Biochemical characterization of homing endonucleases encoded by

fungal mitochondrial genomes

By

Tuhin Kumar Guha

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfilment of the requirements of the degree of:

DOCTOR OF PHILOSOPHY

Department of Microbiology
University of Manitoba
Winnipeg

Abstract

The small ribosomal subunit gene of the *Chaetomium thermophilum* DSM 1495 is invaded by a nested intron at position mS1247, which is composed of a group I intron encoding a LAGLIDADG open reading frame interrupted by an internal group II intron. The first objective was to examine if splicing of the internal intron could reconstitute the coding regions and facilitate the expression of an active homing endonuclease. Using *in vitro* transcription assays, the group II intron was shown to self-splice only under high salt concentration. Both *in vitro* endonuclease and cleavage mapping assays suggested that the nested intron encodes an active homing endonuclease which cleaves near the intron insertion site. This composite arrangement hinted that the group II intron could be regulatory with regards to the expression of the homing endonuclease. Constructs were generated where the codon-optimized open reading frame was interrupted with group IIA1 or IIB introns. The concentration of the magnesium in the media sufficient for splicing was determined by the Reverse Transcriptase-Polymerase Chain Reaction analyses from the bacterial cells grown under various magnesium concentrations. Further, the *in vivo* endonuclease assay showed that magnesium chloride stimulated the expression of a functional protein but the addition of cobalt chloride to the growth media antagonized the expression. This study showed that the homing endonuclease expression in *Escherichia coli* can be regulated by manipulating the splicing efficiency of the group II introns which may have implications in genome engineering as potential 'on/off switch' for temporal regulation of homing endonuclease expression .

Another objective was to characterize native homing endonucleases, cytb.i3ORF and I-OmiI encoded within fungal mitochondrial DNAs, which were difficult to express and purify. For these, an alternative approach was used where two compatible plasmids, HEase.pET28b (+)-

kanamycin and substrate.pUC57-chloramphenicol, based on the antibiotic markers were maintained in *Escherichia coli* BL21 (DE3). The *in vivo* endonuclease assays demonstrated that these homing endonucleases were able to cleave the substrate plasmids when expressed, leading to the loss of the antibiotic markers and thereby providing an indirect approach to screen for potential active homing endonucleases before one invests effort into optimizing protein overexpression and purification strategies.

Acknowledgements

The journey during the last five years has been incredible. I am thankful to many people for their contribution in my research, either directly involved in my research or through friendly discussions.

I am very thankful to my supervisor, Dr. Georg Hausner for giving me the opportunity to work in his laboratory, for believing in me, for the immense guidance and continuous support. The patience, detailed explanation on the research subject whenever needed, friendly approach, your understanding and concern towards the students, have been some of the added benefits working in your lab and I will greatly cherish them forever.

My sincere gratitude to Dr. Peter C Loewen for his mentorship and constant support over the years. This thesis would not have been possible without his support.

To my committee members, Drs. Jörg Stetefeld (Department of Chemistry), Brian Mark, Ivan Oresnik (Department of Microbiology) for their valuable suggestions and comments that helped me to shape the research project better. Moreover, I want to thank the committee members for reading my thesis and for your constructive suggestions.

I am thankful to Dr. Joe O'Neil (Department of Chemistry) for generously allowing access to his laboratory and assisting the circular dichroism spectropolarimetry work.

To Dr. Steven Zimmerly (Biological Sciences, University of Calgary, Canada) for his suggestions on group II introns relating to the *in vitro* splicing assay and also for agreeing to be my external examiner. The necessary corrections/suggestions have been really helpful. I also want to thank Dr. Barry Stoddard (Fred Hutchinson Cancer Research Centre, Seattle, USA) for undertaking the DNA-protein cocrystallographic work.

I am thankful to my past lab mates Dr. Mohamed Hafez, Chen Shen, and Megan Hay for their help, support and friendship and Dr. Mohamed Hafez, again for his suggestions and sharing ideas about mS1247 twintron. Also I am thankful to my present lab mates, Alvan Wai, Iman Bilto, Zubaer Abdullah, Talal Abboud for their suggestions and above all, friendship. My sincere gratitude to Alvan, who has been always beside me, helping me in my research in every possible way and for his constant encouragement.

I am thankful to my departmental colleagues who have helped me, and taught me in handling necessary instruments. My sincere thanks to my fellow colleagues in the department, who have allowed me to use their instruments and chemicals whenever needed. I greatly appreciate yours' generosity.

I would like to extend my gratitude to all the staff members of the Microbiology department.

Moreover, I acknowledge the financial support from the Faculty of Graduate Studies, the Faculty of Science and the Graduate Students' Association at the University of Manitoba. The travel grants did help me to attend and present my research findings in various scientific conferences. The funding from the Faculty of Graduate Studies GETS (Graduate Enhancement of Tri-Council Stipends) program is also gratefully acknowledged.

Last but not least, I am very thankful to my family members who have always believed in me. I am ever grateful to them for their immense support, suggestion and encouragement during my entire Ph.D. program. They have definitely helped me to stay focus in my research work and effectively complete this saga.

*This thesis is dedicated to **Ma Manasha Ma** and **Dadu**, whose blessings have been a significant part of my life.*

Table of contents

List of tables

List of figures

List of abbreviations

BLAST	Basic Local Alignment Search Tool
Cp	Chloroplast
Cfu	Colony forming unit
EBS	Exon Binding Site
ENase	Endonuclease
EtBr	Ethidium Bromide
HEase	Homing Endonuclease
HEG	Homing Endonuclease Gene
IBS	Intron Binding Site
IEP	Intron-Encoded Protein
IGS	Internal Guide Sequence
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Luria Bertani
LHEase	LAGLIDADG Homing Endonuclease
LSU	Large Subunit
LSU rRNA	Large subunit ribosomal RNA
mRNA	Messenger RNA
mt	Mitochondrial
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
NHEJ	Non-Homologous End Joining
nt	Nucleotide

ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
REase	Restriction Endonuclease
rnl	Mitochondrial large subunit ribosomal RNA gene
RNP	Ribonucleoprotein
rns	Mitochondrial small subunit ribosomal RNA gene
rRNA	Ribosomal RNA
RT	Reverse Transcriptase
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SDS	Sodium Dodecyl Sulphate
SSU	Small Subunit
SSU rRNA	Small Subunit ribosomal RNA
TBE	Tris-Borate EDTA
TE buffer	Tris-EDTA buffer

General introduction

New discoveries are sometimes stimulated by initial anomalous results. However, a scientific perspective is required to foresee what these serendipity results have to offer and beyond. The initial discovery of a mobile intron by Bernard Dujon (1980) and harnessing the potential of intron encoded proteins (IEPs) for biotechnological applications in successive studies by other researchers are such examples.

Fungal mitochondrial genomes are highly variable in terms of the overall size ranging from approximately 19 kb to 235 kb (Clark-Walker, 1992; reviewed in Hausner, 2012; Losada *et al*., 2014) due to the presence of intergenic spacer regions mostly consisting of self-splicing group I and group II introns as well as IEPs (Michel and Ferat, 1995; reviewed in Hausner, 2012). Although these self-splicing introns are widespread in the organellar genomes of plant, fungi, algae as well as bacterial genomes (reviewed in Lambowitz and Zimmerly, 2004, 2011; Hausner, 2003, 2012, Hausner *et al*., 2014), these two classes of introns have been characterized based on their sequences, structures and splicing mechanisms (Michel and Westhof, 1990; reviewed in Hausner, 2003). While much of the seminal early research on these elements dealt with the basic mechanism of intron mobility and 'homing' facilitated by the IEPs (reviewed in Dujon, 1989; Belfort *et al*., 2002), extensive research has also been conducted on the utility of IEPs as DNA cutting enzymes for biotechnological purposes (Arnould *et al*., 2006; Takeuchi *et al*., 2011; reviewed in Hafez and Hausner, 2012). These IEPs are commonly known as homing endonucleases (HEases; Thierry and Dujon, 1992).

Homing endonucleases are highly site-specific DNA endonucleases, usually intron- or intein-encoded, which facilitate transfer of intervening sequences (IVS) within target sequences of cognate alleles by mostly catalyzing single- or double-strand breaks (Belfort, 2002; reviewed

in Stoddard, 2006). Based on the conserved nuclease active-site core motifs and catalytic mechanisms, group I encoded HEases are categorized into four major families: LAGLIDADG, GIY-YIG, H-N-H and His-Cys box (reviewed in Stoddard, 2006; Hafez and Hausner, 2012). These enzymes are considered to be the most specific naturally occurring DNA cutting enzymes as they recognize large target sites ranging from 14 to 44 bp within the double-stranded DNA sequences (reviewed in Stoddard, 2006).

Among all the family of HEases, the LAGLIDADG family (LHEases) has become a valuable tool for genome engineering since these molecular “scissors" can be used to replace, modify or eliminate desired sequences with high target specificity; thereby allowing for the modification of various genes in bacteria, plants or animals (Silva *et al*., 2011; Takeuchi *et al*., 2011; reviewed in Stoddard, 2014). Due to their utility, scientists have spent considerable effort on modifying the amino acid residues responsible for HEase target site recognition or on engineering synthetic LHEases in order to increase the target site repertoire that could be covered by these proteins (reviewed in Belfort and Bonocora, 2014). However, in case of LHEases, the endonuclease domain and the binding domain overlap; therefore, engineering of these enzymes is sometimes cumbersome and may lead to labour-intensive and time-consuming extensive trials (reviewed in Hafez and Hausner, 2012; Stoddard, 2014; Sander and Joung, 2014).

Organellar genomes including fungal mitochondrial DNA has been a rich source of mobile introns and IEPs (Hausner, 2003; Sethuraman *et al*., 2009; Hafez *et al*., 2013). Interestingly, these elements tend to localize in the conserved motifs of essential genes such as ribosomal genes (Sethuraman *et al*., 2009), protein coding genes such as, *cyt-b* and *cox1* genes (Ferandon *et al*., 2010; Yin *et al*., 2012); therefore, detection of potential insertions that include putative HEase ORFs can be done by a PCR based survey (Hafez *et al*., 2013, 2014). Also a

number of fungal mitogenomes are available in public data bases and these can be examined for the presence of group I and group II introns that could encode novel HEases with new target sites. The fungal mitochondrial genomes provide a rich resource for ribozymes and homing endonucleases, elements that have applications in biotechnology. Many fungi can be isolated from nature and cultured in laboratories. This project utilized a combination of strategies to find potential intron encoded DNA endonucleases. First, based on a previous study that mined public databases, a homing endonuclease from *Chaetomium thermophilum* was examined in more detail. This HEase was noted to be encoded within a group I intron that has inserted into the mitochondrial small ribosomal subunit gene (*rns*) (mS1247) and what is novel about this element is that the HEase ORF is interrupted by a group II intron. The second strategy was to screen strains of *Ophiostoma ulmi* and related taxa for possible intron insertions within the *rns* and *cyt-b* genes. In this work one *rns* encoded intron ORF from *Ophiostoma minus* and one intron encoded ORF from *Ophiostoma novo-ulmi* subspecies *americana* was examined in more detail to assess if these introns encode functional homing endonucleases. The long term goal of my study is to build towards establishing a catalog of HEases with novel target sites. Therefore, bioprospecting for native HEases will provide an attractive alternative to the extensive protein engineering currently required and contribute further to the genome engineering field by expanding target site repositories.

Chapter 1

Literature review

1.0. Introduction

Genome sequencing reveals a wealth of important information for any organism. For example, molecular technique unveils the presence of intervening sequences (IVS) that might reside within protein-coding genes (Chow *et al*., 1977; Berget *et al*., 1977), ribosomal RNA (rRNA) genes (Back *et al*., 1984; Ralph *et al*., 1993) and transfer RNA (tRNA) genes (Heinemann *et al*., 2010). However, the RNA splicing event restores a continuous gene product by removing these IVS post-transcriptionally (Cech, 1990; Saldanha *et al*., 1993). These sequences are commonly known as introns (Gilbert, 1978) which can be broadly classified into two categories such as (i) self-splicing: group I, II and III introns (group III introns are degenerated group II introns) (Cech, 1990; Palmer and Logsdon, 1991; Robart and Zimmerly, 2005) and (ii) protein assisted splicing: spliceosomal, tRNA and archaeal introns (Cavalier-Smith, 1991; Biderre *et al*., 1998; Lynch and Richardson, 2002; Calvin and Li, 2008; reviewed in Irimia and Roy, 2014).

Even though group I and group II introns are categorized as self-splicing introns, they are distinctive in terms of their sequences, secondary and tertiary structures and splicing mechanisms. Group I introns are self-splicing ribozymes which have high variation in the primary sequence level, however the core secondary structure mostly consists of nine paired regions (P1-P9) which fold into two essential domains required for splicing (Michel and Westhof, 1990). The splicing of group I intron depends on two sequential transesterification reactions mediated by the intron's RNA tertiary structure, an external guanosine moiety and sometimes intron/nuclear encoded proteins such as maturases (Cech, 1990; Saldanha *et al*., 1993). Group II introns, on the other hand are retroelements and are speculated to be the ancestors of the spliceosomal introns and retrotransposons in eukaryotes (Copertino and Hallick,

1993). Like group I introns, group II introns also exhibit conserved secondary and tertiary structures, however they are visualized as six stem-loop domains (DI-DIV) radiating from the central wheel-like structure (Michel and Ferat, 1995; Pyle and Lambowitz, 2006). These six domains interact in an orderly fashion to form a conserved splicing competent tertiary structure that allows the distant intron/exon boundaries to interact in close proximity within the intron's active site (Michel and Ferat, 1995; Qin and Pyle, 1998). In addition, a branch-point nucleotide residue and divalent metal ions also activate the suitable bonds for catalysis (Lambowitz and Zimmerly, 2011).

Interestingly, some group I and group II introns provide an ideal 'hideout' within the host genome where they provide a means for intron encoded proteins (IEPs) to perpetuate (i.e. a mutualistic relationship between the IEPs and their host introns) without adversely affecting the host gene function. The encoded proteins return the favor by rendering mobility to these genetic elements, hence categorized as mobile genetic elements (Dujon *et al*., 1986; Belfort *et al*., 2002). The protein analog of introns, known as inteins (Perler *et al*., 1994) belong to another class of IVS which are removed post-translationally (Anraku *et al*., 1990; Kane *et al*., 1990). Introns/inteins are often conceptualized as 'selfish' genetic elements (Dawkins, 1976) which have mostly evolved mechanisms to prevent their extinction without providing any selective advantage to the host genome (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; see section 1.9. for exceptions).

1.1. Discovery

The discovery of mobile introns was a serendipity and dates back to the experiments conducted in the early 1970s at the Pasteur Institute in Paris. In *Saccharomyces cerevisiae*, a

genetic marker termed 'omega' (ω^+) was observed to be transferred at near 100% frequency (i.e. 'super Mendelian' inheritance) in the genetic crosses involving homozygous ω^+ and ω^- yeast strains (Dujon *et al*., 1974). Later, this marker was shown to correspond to a 1.1 kb group I intron found in the large ribosomal subunit RNA (LSUrRNA) gene of the mitochondrial genome in ω^+ yeast strain (Bos *et al*., 1978; Faye *et al*., 1979). Sequencing of the ω^+ intron revealed a 708 base pair (bp) open reading frame (ORF) which was able to encode a 235 amino-acid protein (Dujon, 1980). Expression from this ORF yielded a functional endonuclease which was crucial for the intron mobility (Colleaux *et al*., 1986, 1988). The initiation of the mobility was due to a transient double-strand break (DSB) near the intron-insertion site in a cognate allele that lacked an insertion. The duplication of the intron and its encoded endonuclease gene into the target site was the result of cellular double-strand repair mechanism via homologous recombination (HR) using the intron-containing allele as the 'repair' template (Zinn and Butow, 1985; Colleaux *et al*., 1986). This functional endonuclease, later named I-SceI (see sub-section 1.7.2. for nomenclature conventions) was the first known representative of the IEPs collectively known as homing endonucleases (HEases; Jacquier and Dujon, 1985).

1.2. Homing

Homing is a site-specific mobility event where a mobile IVS (group I or group II or intein) is horizontally transferred, usually to a homologous allele of the host gene lacking the IVS. The frequency of this genetic event is high and results in uni-directional duplication (or transfer) of the IVS in the cognate intron-/intein-less allele within a diploid genome thereby providing a fitness (increase in numbers) advantage for the genetic element such as persistence in the genome (Dujon, 1989). Site-specific DNA endonucleases i.e. HEases (see section 1.7. for

details) encoded by the ORF residing within the mobile intron or intein (Gimble, 2000; Stoddard, 2006; Perler *et al*., 1997; Southworth and Perler, 2002), sometimes freestanding (i.e. not present within the introns; Herskowitz *et al*., 1992; Zeng *et al*., 2009) initiate the horizontal movement of intron/intein usually to a new location in the host genome. If the horizontal movement of the mobile element involves an orthologous gene, then it is termed as 'homing' (Dujon, 1980) and ectopic integration occurs if this new location happens to be in a different gene (Roman and Woodson, 1995).

DNA sequences both up- and down-stream of the insertion site constitute the homing site which is usually centered near the intron-insertion site. When the homing site is disrupted by an intron (i.e. intron-containing allele), the target site (i.e. the homing site) is lost, thus providing a mechanism to discriminate intron-less (non-self) alleles from intron-containing (self) alleles (Dujon, 1989; Belfort *et al*., 2002).

1.3. Evolutionary speculations

Koonin *et al.* (2006) proposed that group I and group II introns evolved in the precellular RNA world. According to this speculation, the primordial pool of primitive genetic elements was conceptualized also as the source for the original lineages of viruses and related selfish elements. Moreover, it was also speculated that the mitochondrial endosymbionts that gave rise to the eukaryotic organelles probably carried with them mobile elements such as mobile introns and plasmids (Koonin *et al*., 2006; Martin and Koonin, 2006; Hausner, 2012). The acquisition of ORFs by the ribozyme type introns are described in the next section.

The evolutionary origin of homing, particularly the group I intron mobility has always been fascinating. The first hypothesis 'endonuclease-gene invasion' was based on the

biochemical experiments performed on an IEP in the T4 phage. The researchers noted that in the *sunY* gene, the intron sequences flanking the ORF encoding HEase (I-TevII) were similar to the exonic junction sequences which constitute the I-TevII target sequence site. Moreover, they were able to demonstrate that I-TevII ORF was able to cleave a synthetic construct comprising of both up- and down-stream sequences flanking the I-TevII ORF. This result provided evidence for the 'endonuclease-gene invasion' hypothesis where a freestanding HEase can cut an intron sequence which inadvertently resembles the HEase target site. The double-strand recombinogenic-repair event later completes the overall process by inserting the endonuclease gene sequence into the cleaved intron sequence generating a composite mobile element (Loizos *et al*., 1994; reviewed in Hausner *et al*., 2014).

Recently, another theory on the origin of intron homing has been proposed. In cyanobacterial phages, a novel freestanding HEase, F-CphI resides adjacent to *psbA* gene which is interrupted by a self-splicing group I intron. However, this intron does not encode its own endonuclease. Interesting enough, the recognition and the cleavage sites of F-CphI encompassed sequence that includes the intron-insertion site in the intron-less *psbA* genes. However, this mechanism is dependent on the physical proximity of the pre-adapted freestanding endonuclease and an intron in the adjacent gene. Through collaborative effort and plausible illegitimate recombination during coinfection, a non-mobile *psbA* intron can be mobilized by the adjacently encoded F-CphI into the intron-less allele of *psbA* gene. This speculation pointed towards the possibility of collaborative or *trans* homing (Zeng *et al*., 2009; Bonocora and Shub, 2009).

1.4. Distribution of homing

The process of intron homing appears to be widespread. Based on reviews (Lambowitz

and Belfort, 1993; Lambowitz *et al*., 1998), 30% of group I introns are estimated to contain internal ORFs and a significant number of them are assumed to be mobile. Group I intron homing, so far, is the most wide-spread reported event compared to the homing exhibited by the group II intron. It is found in mitochondrial DNA of fungi, mitochondrial and chloroplast genomes of plants, algae and some protozoans as well as nuclear genomes of slime molds, ciliates, algae, fungi, soft corals and sponges (Gimble, 2000; Hafez and Hausner, 2012). In contrast, group I introns are rarely encountered among bacteria (reviewed in Hausner *et al*., 2014). If present, they are predominately inserted within structural RNA genes such as tRNA (Paquin *et al*., 1997; Rudi *et al*., 2002) and rRNA genes (Haugen *et al*., 2007; Salman *et al*., 2012), protein coding genes such as *nrdE* genes in some cyanobacteria (Meng *et al*., 1997; Fujisawa *et al*., 2010), *nrdE* and *recA* genes in various *Bacillus* species (Tourasse *et al*., 2006; Ko *et al*., 2002), flagellin gene in a thermophilic *Bacillus* species (Hayakawa and Ishizuka, 2009, 2012). Even though group I introns are ancient, they are absent in Archaea (Tocchini-Valentini *et al*., 2011). Currently, there are three speculations for such scarcity of group I introns among the prokaryotes. Homing is facilitated by the presence of multiple targets offered by repetitive DNAs (rDNAs) or multi copy genomes such as chloroplast and mitochondrial DNAs in eukaryotes with lower mutation rates (Hausner, 2012). The absence of such multicopy targets in bacteria may be the first factor that explains why mobile introns such as group I introns are not so common amongst bacteria. Second, the extremely prevalent presence of primitive defense mechanisms, possibly, based on the RNA interference principle and the newly discovered CRISPR/Cas defense system in the bacterial genome might limit the spread of foreign DNA elements like mobile group I introns (Barrangou, 2013; Hausner *et al*., 2014; Silas *et al*., 2016). Unlike the eukaryotic transcription and translation machineries which are compartmentalized, insertion of

group I introns into the protein-coding genes in bacteria, which exhibits coupled transcription and translation events may not be welcoming, the later is supposed to interfere by providing lesser time in proper folding of the group I introns to facilitate ribozyme formation and thus efficient splicing. This could be the third factor that would ultimately lead to the elimination of such mobile introns from the bacterial genomes (Öhman-Hedén *et al*., 1993; Edgell *et al*., 2000; Hausner *et al*., 2014).

In addition, many group II introns, archaeal introns and inteins also exhibit homing. The genomes of fungal and plant mitochondria, chloroplast genomes of eubacteria, algae and plant encounter group II intron homing (Belfort *et al*., 1995; Lambowitz *et al*., 1998). Group II introns have been recorded in early branching metazoans (reviewed in Hausner *et al*., 2014; Huchon *et al*., 2015), however they are rare in archaea (Rest and Mindell, 2003) and have not been found in the nuclear genomes of eukaryotes. It has been suggested that group II introns gave rise to spliceosomal introns and various types of retroelements which are highly abundant in eukaryotes (Xiong and Eickbush, 1990; Lambowitz and Belfort, 2015). It is interesting to note that spliceosomal introns are not known to be mobile (Lambowitz and Zimmerly, 2011). The structural similarities between group II introns and spliceosomal messenger RNA (mRNA) introns in eukaryotic genome suggest that they might be derived from once-mobile group II introns (Weiner, 1993; Sharp, 1994; Koonin *et al*., 2006). Engineered group II introns invading ectopic sites in the eukaryotic chromosome further support this theory (Guo *et al*., 2000; reviewed in Molina-Sánchez *et al*., 2015).

Archaeal introns are present within tRNA and rRNA genes, although they have rare occurrence (Lykke-Andersen *et al*., 1997). Inteins however are found in archaea, bacteria, nuclear and organellar genomes of few eukaryotes such as yeast (Perler *et al*., 1997).

1.5. Homing mechanisms

Even though group I introns and group II introns are widespread, which is attributed to the endonuclease reaction catalyzed by the IEP or in part by the group II intron lariat, the overall mechanism varies dramatically for these two types of introns (Figure 1). The generalized mechanism for each will be discussed in the following sub-sections.

1.5.1. Group I intron homing

The mobile group I intron depends on the translation of the IEP which is highly specific in recognizing and binding to a homing site in the intron-less cognate allele. Once bound to the homing site, a DSB is created by the endonuclease. The cellular repair machinery mends this breakage by means of HR using the intron-containing allele as the corrective template (reviewed in Lambowitz and Belfort, 1993). In the process, sometimes, the flanking regions of the homing site are also transferred (co-conversion) into the cognate intron-less allele (Bell-Pedersen *et al*., 1989). Initial studies of group I intron homing with the *td* intron of phage T4 indicated the requirements of various exonucleases and *E.coli* recombinase RecA for homologous strand invasion of an intron containing allele thereby facilitating repair of the DSB and precise intron-insertion (Bell-Pedersen *et al*., 1989; Clyman and Belfort, 1992).

Transfer of the mobile elements with reduced efficiency in T4 phage infection system using hosts deficient in enzymes responsible for crossover resolution indicated that sometimes, mobile group I intron after the DSB event uses other gene conversion pathways like recombination-independent pathway. The mechanisms might include synthesis-dependent strand annealing (SDSA) or a topoisomerase-mediated (TM) pathway (Mueller *et al*., 1996). These mechanisms overall ensure for a faithful duplication i.e. horizontal transfer of the IVS to the

cleaved allele.

Not much is known about the homing of archaeal introns and inteins, however, it is thought to proceed using similar DNA based mechanism employed as seen in other group I introns for their mobility (Lykke-Andersen *et al*., 1997).

1.5.2. Group II intron 'retrohoming'

The homing mechanism exhibited by group II intron is termed 'retrohoming' which is significantly different as well as more complicated from the homing mechanism used by group I intron, archaeal introns and inteins (Curcio and Belfort, 1996). Detailed biochemical and genetic studies of the mobile *aI*2 intron in the mitochondrial yeast *cox1* gene suggested a mechanism which involved the presence of a ribonucleoprotein (RNP) complex consisting of the encoded protein and the spliced intron RNA (Yang *et al*., 1998). Interestingly, this single encoded protein has four distinct domains namely DNA-binding domain, endonuclease, reverse-transcriptase and maturase which work synergistically. The excised RNA intron participates in this homing process by associating with the IEP to form a stable RNP and recognizes the intron-less allele via base-pairing between the protein bound RNA and the target DNA sequence. Subsequently, the 3′ hydroxyl group of the intron serves as a nucleophile and cleaves just one strand of the DNA homing site. The RNA lariat is reverse spliced into the target site and the endonuclease domain of the assisted protein partner cleaves the complementary DNA strand (Zimmerly *et al*., 1995a, b). The reverse-transcriptase domain synthesizes DNA using the invading RNA template. Cellular machinery completes the homing process by replacing the RNA with DNA (Lambowitz and Zimmerly, 2011).

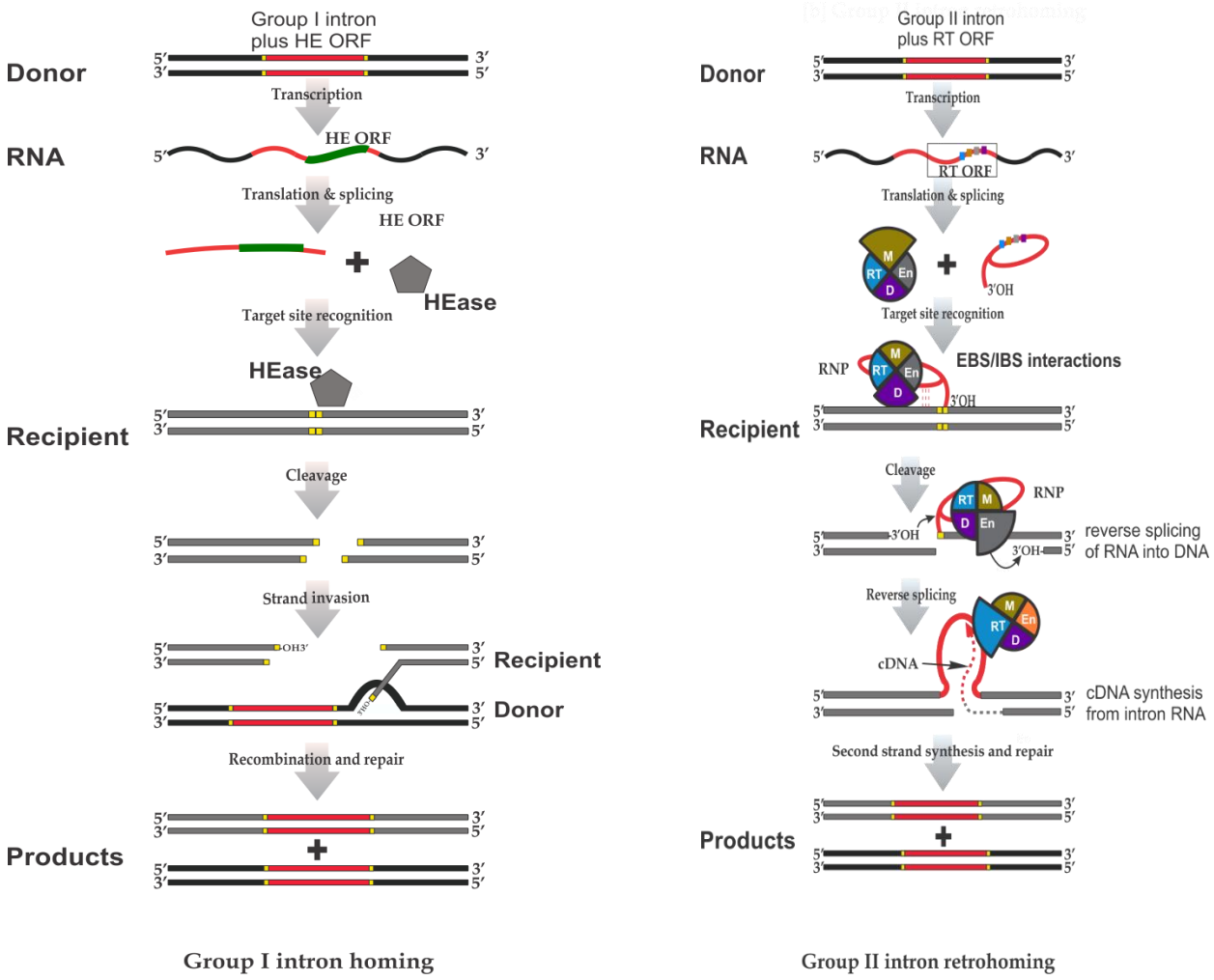


Figure 1. Generalized homing mechanisms for mobile group I introns and group II introns. While the group I intron homing rely on cell-based repair of DSBs induced by the endonuclease, the group II intron uses a more complex mechanism of reverse splicing and subsequent reverse transcription of the intron, as described in the text.
Picture courtesy: Mohamed Hafez, 2016.

1.6. The homing cycle

The availability of intron-/intein-less alleles for endonuclease-mediated homing, the phenotypic cost associated with the insertion of a mobile element, the presence of efficient homology-based DSB repair systems are the few important factors on which the evolutionary dynamics of mobile introns/inteins depend (reviewed in Hausner *et al*., 2014).

In order to test the idea that horizontal transmission was necessary for the long-term persistence of selfish genes, 20 species of yeasts were surveyed for the group I intron and its embedded 'ω' ORF. The survey revealed three evolutionary stages for the 'ω' ORF i.e. functional, nonfunctional, or absent. Moreover, the phylogeny of the ORF differed significantly from that of the host strain which indicated a strong evidence of horizontal transmission. The results from this observation was rewarding. The life cycle for the homing endonuclease genes (HEGs) commonly known as the 'homing cycle' was proposed (Goddard and Burt, 1999). According to this event, an empty site within a genome is invaded from another organelle or organism by a group I intron- or intein-associated HEG via horizontal transmission. Subsequently, the homing mechanism stably replicates the group I intron or intein gene and its associated ORF to identical loci in a recipient intron-less or intein-less cognate alleles. As these elements appear to be neutral, there is a lack of selection so inactivation and eventual elimination of the intron or the intein gene arises due to point mutations within the HEG leading eventually to the loss of the HEG and intron. Thus an empty site is regenerated and this step prepares the stage for the second invasion which continues the cycle of invasion and loss.

1.7. Homing endonucleases

Homing endonucleases are small (<40 kDa), diverse, usually intron- or intein-encoded, highly site-specific (rare-cutting) DNA endonucleases which facilitate homing of IVS by catalyzing single- or double-strand break within target sequences of cognate alleles. By transmitting their own genes horizontally within a host population, they can increase their allele frequency greater than that acquired through Mendelian inheritance (Chevalier and Stoddard, 2001; Stoddard, 2006). These proteins are small (e.g. I-CreI is a homodimer of 163-residue monomers) due to the length limitations of the mobile elements in which they reside (Kowalski and Derbyshire, 2002; Chevalier and Stoddard, 2001). To avoid deleterious effects on intron splicing, HEGs are usually inserted in loops of the intron tertiary structure (e.g. P9 stem-loop for group I introns or domain IV for group II introns) that presumably do not interfere with the splicing-competent folding of the introns (Michel and Westhof, 1990). However, an alternative scenario, 'core creep' which is essentially an extension of the IEP coding region has been described. Here the HEG which is inserted into the peripheral loop of the intron overlaps with the core intron sequences. This occurs due to mutations of sequences upstream of the original HEase start codon being converted into coding sequences that can extend the ORF further upstream and eventually fuse the HEase ORF with the upstream exon sequence (Edgell *et al*., 2011).

Reading until now, it is tempting to align the homing event with genetic transposition event, as both involve transfer of the mobile element which is initiated by the protein encoded within that sequence. As a matter of fact, HEases have several distinguishing features that differentiate them from transposable elements and restriction enzymes (REases).

1.7.1. Distinguishing characteristics of HEases

In the transposition event, an enzyme transposase recognizes and interacts with the ends of the transposon analogous to the mechanism used by an integrase or a recombinase protein. HEase, on the contrary, does not recognize its corresponding mobile DNA but simply cleaves the target site, and usually the cellular DSB repair mechanism completes the homing process (Jurica and Stoddard, 1999).

Most HEases and REases (type II enzymes) share a common characteristic: the ability to cleave at a site-specific double-stranded DNA. Apart from this only similarity, these enzymes have few differentiating characteristics. Both of them have evolved independently in different genomic locations (Wilson, 1988), differ in substrate recognition properties and the requirement of accessory factors for endonuclease activity (Belfort and Roberts, 1997). HEases have long, generally asymmetric recognition sequences which span 14 bp (I-DmoI) up to 40 bp (I-TevI) and, therefore they are also known as meganucleases (Thierry and Dujon, 1992). They are more flexible towards several base pair changes i.e., can tolerate limited sequence polymorphisms within their recognition sites (reviewed in Stoddard, 2006; see sub-section 1.8.1.2 for details). Restriction enzymes on the other hand, recognize dyad symmetrical target sequences of much shorter length in the range of 3-8 bp and are highly sensitive to single-site mutations in their recognition sites. For example, EcoRV makes 11 base-specific contacts with its 6 bp substrate, therefore it is exquisitely sensitive towards single base changes (Chevalier and Stoddard, 2001).

The HEGs are usually found in mobile introns, inteins and sometimes freestanding which may be present in various genomic microenvironments (nuclei, mitochondria, and chloroplast) of eukaryotic cells apart from being found in archaea and bacteria (Lambowitz and Belfort, 1993). They are basically ‘selfish’ elements and help in the transfer of introns or inteins in which they

reside (reviewed in Stoddard, 2006). In contrast, the genes encoding REases have always been found in archaea, bacteria and few eukaryotic viruses (Roberts and Macelis, 1997). They are always freestanding and mostly found in close association with the genes coding cognate DNA modifying enzymes (Wilson, 1988) and serve a specific "purpose" i.e., defense against foreign DNA (e.g. phage DNA) (Roberts, 1976).

1.7.2. Nomenclature conventions

Initial discovery of HEase followed by its functional demonstration qualifies for a nomenclature and this springs from the convention used for REases (Smith and Nathans, 1973). The naming consists of three letters: first letter belongs to the genus, second and third letter belongs to the first two letters of the species from where the enzyme is discovered followed by a Roman numeral to distinguish multiple enzymes from the same organism (REBASE; Roberts and Macelis, 1997). Italicization is not necessary for the first three-letter acronym of the HEase as most journals avoid retaining the italic convention because it is sometimes hard to translate in machine language during publication (Roberts *et al*., 2003). However, due to historical reasons, the nomenclature convention was not implemented for the Homothallic (HO) endonuclease from *S. cerevisiae* and the 'Similar to Endonucleases of Group I introns' (SegA) endonuclease from T-even phage T4 (Belfort and Roberts, 1997). Several prefixes are used, such as (i) I = intron-encoded (e.g. IAniI), (ii) PI = protein insert/intein-encoded (e.g. PI-SceI), (iii) F = freestanding (e.g. F-TevII), (iv) H = hybrid (H-DreI) (Waring *et al*., 1982; Kane *et al*., 1990; Kadyrov *et al*., 1994; Chevalier *et al*., 2002). Thus, the first discovered, biochemically characterized intron-encoded HEase from *S. cerevisiae* was named I-SceI.

It was also decided that the above designation will only be strictly applied for the HEase

demonstrating endonuclease activity. Therefore, the nomenclature convention of the HEase should not be based just on the mere similarity to one of the several conserved sequence motifs which group the vast majority of HEases into various HEase families.

1.8. Homing endonuclease families

Based on the conserved nuclease active-site core motifs and catalytic mechanisms, group I encoded HEases are categorized into four main families: LAGLIDADG, GIY-YIG, H-N-H, His-Cys box (Stoddard, 2006). While another motif PD-(D/E)-XK found in cyanobacterial tRNA group I introns (Zhao *et al*., 2007) has secured an additional family status, suggestions based on superposition of the active sites have been put forward to combine H-N-H and His-Cys box into the ββα-Me superfamily (Friedhoff *et al*. 1999; Kühlmann *et al*., 1999). Several HEase-like protein families such as the Vsr (very-short patch repair) endonucleases identified from the cyanobacterial tRNA group I introns (Dassa *et al*., 2009), the Holliday junction resolvase-like HEases in phages (Zeng *et al*., 2009) and EDxHD family found within the *recA* gene of *Bacillus thuringenesis* 0305φ8-36 bacteriophage (Taylor *et al*., 2011) have been recognized as additional minority families.

HUH endonuclease family (in which U represents a hydrophobic residue) should not be confused with an H-N-H family. This family of enzymes helps in the mobility for genetic elements by catalysing cleavage and rejoining of single-stranded DNA using an active-site tyrosine residue to make a transient 5'-phosphotyrosine bond with the DNA substrate. They also have a key role in rolling-circle replication of plasmids and bacteriophages and in various types of transposition events (Kornberg and Baker, 1992; reviewed in Chandler *et al*., 2013).

Even though a few LAGLIDADG ORFs typical of group I introns have been discovered

to be encoded by group II introns (Toor and Zimmerly, 2002; Mullineux *et al*., 2010), the group II intronic ORFs typically encode reverse-transcriptases (RTs) and these enzymes are not the focus of my thesis. With regards to group II introns and RTs, the readers are referred to the reviews by Lambowitz and Belfort, 2015; Zimmerly and Wu, 2015; Zimmerly and Semper, 2015; McNeil *et al*., 2016. The following sub-sections provide a brief overview of the four major HEase families.

1.8.1. LAGLIDADG family

The discovery of the omega intron encoding the I-SceI by Bernard Dujon (1980) was beneficial in two ways. First, a novel HEase made its fortuitous appearance. Second, this enzyme along with other HEases bearing the most recognizable conserved 10-sequence motif were assigned to the largest family of HEases termed the LAGLIDADG family (Hensgens *et al*., 1983; Stoddard, 2006; Takeuchi *et al*., 2011). The members of this family includes single motif LAGLIDADG proteins such as I-CreI (Thompson *et al*., 1992), I-CeuI (Turmel *et al*., 1997) and their molecular dimeric cousins (double motif LAGLIDADG motifs separated by 80-150 residues) such as PI-SceI (Duan *et al*., 1997), and I-DmoI (Silva *et al*., 1999). The single motif and double motifs behave as homodimers and monomers respectively. This segregation within the family may be attributed to the possible gene duplication event (Haugen and Bhattacharya, 2005). It is interesting to note that even so they are part of the same family, they have unique characteristics. The homodimers recognize palindromic or near-palindromic DNA target sites while the monomeric enzymes are not constrained to symmetric DNA target sequences. Even though the active sites are found adjacent to each other in the tertiary/quaternary structure, both forms of these enzymes display two unique active sites responsible for the catalytic cleavage of

the DNA strand (reviewed in Stoddard, 2006).

The members have a broad biological host range ranging from archaea, bacteria, bacteriophages, mitochondrial genomes of fungi and protozoa, chloroplasts of algae and plants and some of the early branching metazoans (Hafez and Hausner, 2012). They are usually encoded from group I introns, inteins, archaeal RNAs but are sometimes freestanding (Dalgaard *et al*., 1997).

It is interesting to note that group I intron-encoded LAGLIDADG HEases (LHEases) can act as maturases which aid in efficient splicing of introns by folding the RNA into a conformation that favours splicing, while at other times, LAGLIDADG maturases behave as functional endonucleases such as I-AniI from *Aspergillus nidulans* (Ho *et al*., 1997; Monteilhet *et al*., 2000; Belfort, 2003; Longo *et al*., 2005).

The crystallographic structure of several LHEases bound either to their target DNA such as I-CreI (Chevalier *et al*., 2001), I-MsoI (Chevalier *et al*., 2003), I-AniI (Bolduc *et al*., 2003), I-SceI (Moure *et al*., 2003), I-OnuI (Takeuchi *et al*., 2011), I-SmaMI (Shen *et al*., 2016) or in the absence of DNA such as I-DmoI (Silva *et al*., 1999), PI-PfuI (Ichiyanagi *et al*., 2000) have been resolved with great precision. These molecular structures were instrumental in dissecting the structural and functional significance of the LAGLIDADG motif, the mechanism of DNA recognition and endonucleolytic activity.

1.8.1.1. Structure

Even though the LAGLIDADG members have very little homology with regards to their primary sequence, the endonuclease domain with an average dimension of approximately 25 Å x 25 Å x 35 Å consisting of a core fold with mixed α/β topology (α-β-β-α-β-β-α) is significantly

conserved across the family. The conserved motif found in each folded domain (for monomeric versions) is vital to both structural and functional integrity of the enzyme. While the last three residues of the motif facilitate a tight turn from the amino-terminal α-helix into the first β-strand of each DNA-binding surface, the individual side-chains from the amino-terminal helices created by the first seven amino-acids, on the other hand, engage either in core packaging or establish contacts across the interdomain surface. Small residues such as glycine and alanine that allow van der Waals contacts between backbone atoms mediate the helix packing at this interface (Heath *et al*., 1997). The overall structure can be visualized as half-cylindrical or 'saddle-like' with a groove formed by the underside of the saddle. The antiparallel four-stranded continuous β-sheets which cross the groove axis at an angle of 45° are composed of large number of exposed basic and polar moieties required for DNA contacts and binding (reviewed in Stoddard, 2006; Figure 2A and 2B). Structural alignment of several endonuclease domains revealed the conserved nature of the central core of the β-sheets which correspond to the residues that make contacts to base pairs in each DNA half-site (Bolduc *et al*., 2003).

1.8.1.2. DNA recognition

High resolution complex crystallographic structures of the protein-DNA interface of I-CreI and I-PpoI HEases provided insights into the flexible recognition strategy for the DNA substrate site where some individual polymorphisms are allowed without significantly compromising the binding efficiency (Moure *et al*., 2002, 2003; Chevalier *et al*., 2003). This moderately relaxed specificity due to undersaturation of the DNA-protein contact may be useful for intron mobility where differences between homologous alleles of the host gene are

interrogated by the HEases to promote lateral transfer between closely related species (Belfort and Bonocora, 2014).

In each recognition interface, a set of four antiparallel flexible β-strands from each LAGLIDADG domain provide both direct and water-mediated contacts between amino-acid side chains and nucleotide atoms of the major groove in each DNA half-site. Typically, strands β1 and β2 extend from base pairs 3 to base pairs 11 on either side of the scissile phosphate groups thereby making contacts with the entire length of the interface. While the central four base pairs (+2 to -2) are devoid of any direct contact with the residue side chains, additional contacts to base pairs 3, 4 and 5 in each half site are provided by strands β3 and β4. For illustration, the direct and water-mediated contacts between the I-CreI enzyme and the bases of its DNA target site have been depicted in Figure 3A. It has been estimated that approximately three-fourth of the possible hydrogen-bond donors and acceptors of the base pairs in the major groove are contacted by the β strands, making few or no additional contacts in the minor groove. Within the target site between bases -3 to +3, DNA around the endonuclease binding surface is locally overwound and twisted (~50°) leading to narrowing of the minor groove. This configuration not only positions the scissile phosphates in close proximity with each other (approximately 5-8 Å apart) but also places these phosphates in near vicinity of the bound metal ions in the LAGLIDADG domain for efficient DNA-cleavage.

1.8.1.3. DNA-cleavage

The general mechanism of a nucleolytic cleavage involves positioning of an activated nucleophile (usually a deprotonated water molecule) for an in-line attack on the electrophilic 5′ phosphate (Chevalier and Stoddard, 2001). The well-studied I-CreI homodimeric enzyme

provided important insights to the DNA-cleavage mechanism (Chevalier *et al*., 2004); the results can be generalized to the other LHEases. The members of this family cleave target DNA across its minor groove yielding 4-nucleotide (nt), 3′ hydroxyl (OH) overhangs. Even though the enzyme positions itself on the major groove of the DNA, it must overcome the steric hindrance in order to load its catalytic machinery across the narrow minor groove (9 Å). The enzyme counteracts this problem by adopting a tight packaging of the LAGLIDADG helices at the dimer interface, positioning the catalytic side chains of Aspartate 20 (Asp 20) and Asp 20′ into the DNA-binding groove.

Similar to most known endonucleases, this enzyme family is dependent on divalent cations, particularly magnesium for activity. The co-crystal structures of PI-SceI (Moure *et al*., 2002), I-AniI (Bolduc *et al*., 2003) and I-CreI (Chevalier *et al*., 2004) have shown that two divalent metal ions, each bound to one Asp residue independently coordinates with a single oxygen from the scissile phosphate (i.e., a phosphate bond that can be broken by an endonuclease). The presence of a third divalent metal ion coordinated by a pair of overlapping active sites, also participates in cleavage reactions of both the DNA strands by virtue of its interaction with the scissile phosphates and 3′ hydroxyl leaving groups (Figure 3B). The precise use of bound divalent metal ions for cleavage is still unknown, however, the role of divalent cation is usually thought to lower the dissociation constant (pKa) of a bound water molecule for easy deprotonation, thereby activating the water molecule during the endonuclease reaction. Moreover, high-resolution structures of I-CreI show a metal-bound water molecule in each active site that is appropriately positioned for an in-line hydrolytic attack on a scissile phosphate group (Figure 3B). In addition, within a large pocket surrounding the DNA scissile phosphate group which extend from the metal-bound nucleophile to the leaving 3′ oxygen group, a well-ordered

network of solvent is positioned and coordinated by several basic residues. The LAGLIDADG family of enzymes is unique compared to the other hydrolytic endonucleases. For these enzymes, there has been no essential acidic residue that has been unambiguously identified as a general base for activation of a nucleophilic water molecule. In addition, the basic residues in their active sites are not generally found in contact distance with the metal-bound waters; however, the nucleophilic water is in direct contact with the surrounding ordered water molecules (Chevalier *et al*., 2004; Figure 3B). Therefore, the capacity to either donate or accept one or more hydrogen bonds is the only obvious common chemical feature of many of these residues. Thus, it has been speculated that each branch of closely related LAGLIDADG enzymes may have adopted a unique active-site solvent-packing arrangement that is highly specialized to fulfill DNA hydrolysis (reviewed in Stoddard, 2006).

The cleavage and the intron-insertion site for the members of the LAGLIDADG family are coincident or sometimes separated by few nucleotides (Belforts and Roberts, 1997; Guha and Hausner, 2014). However, PI-Mle*I* from *Mycobacterium leprae* RecA intein is different from other members of the family for its modular structure with functionally separable domains for DNA-binding and cleavage (Singh *et al*., 2009, 2010).

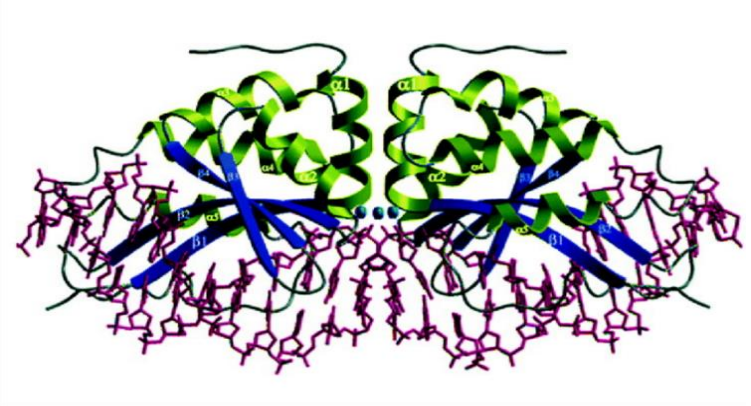


Figure 2A. Cartoon representation of the structure of 22 bp DNA bound complex of the I-CreI homodimer made using PYMOL (DeLano, 2002). The alpha helices (marked as α1 through α4) in each domain are shown in green colour while the anti-parallel beta sheets (marked as β1 through β4) are shown in dark blue colour which are positioned on the major groove of the DNA in each recognition half-site (DNA backbone shown in purple colour). Three metal ions shown in dark green are visible at the base of the LAGLIDADG interdomain are required for DNA endonuclease activity.
Stoddard BL. 2006. Homing endonuclease structure and function. Q. Rev. Biophys. **38**(1): 49-95. (Springer Publications. Image reproduced with permission. License number: 3831560857433).

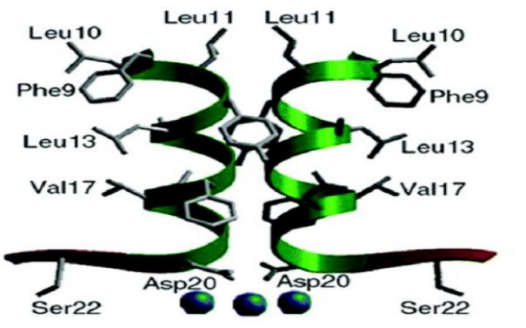


Figure 2B. The LAGLIDADG motifs form the helices at the domain interface of the I-CreI structure. The hydrophobic residues present in the domain interface such as leucine (Leu), phenylalanine (Phe), valine (Val), serine (Ser) are indicated. The Aspartate residues in both the domains bind metal ions. A shared metal ion is also visible in the middle. Small residues such as glycine and alanine that allow van der Waals contacts between backbone atoms mediate the helix packing at this interface.
Stoddard BL. 2006. Homing endonuclease structure and function. Q. Rev. Biophys. **38**(1): 49-95. (Springer Publications. Image reproduced with permission. License number: 3831560857433).

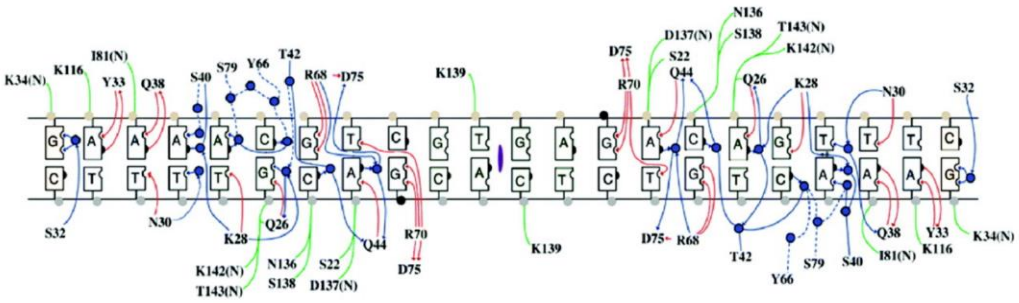


Figure 3A. Summary of direct contacts and water-mediated contacts between the I-CreI enzyme and the bases of its DNA target site. Solid arrows (red to bases, green to backbone atoms) depict the direct bonds. Blue dashed and solid lines represent the water-mediated contacts. The scissile phosphates are indicated with black closed circles.
Stoddard BL. 2006. Homing endonuclease structure and function. Q. Rev. Biophys. **38**(1): 49-95. (Springer Publications. Image reproduced with permission. License number: 3831560857433).

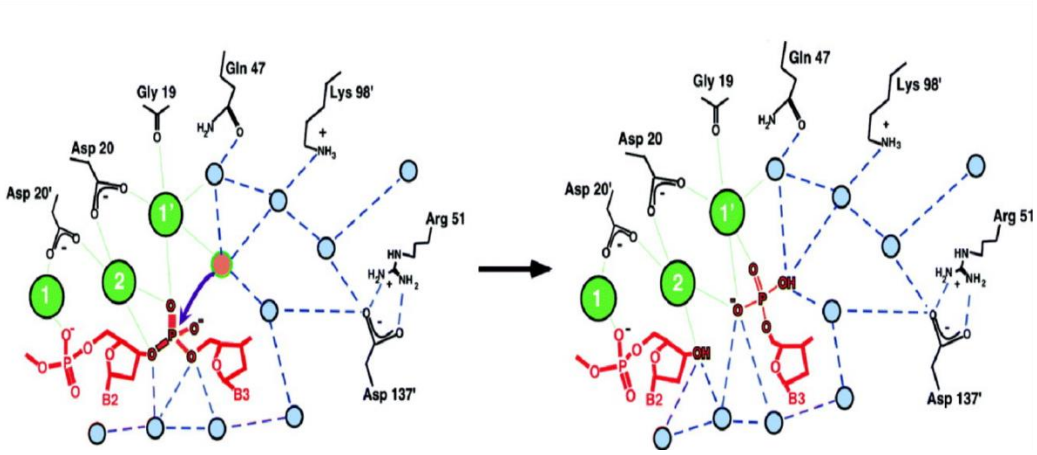


Figure 3B. Proposed catalytic mechanism for I-CreI, as described in the DNA cleavage (subsection 1.8.1.3). The nucleophilic water is orange; surrounding ordered water molecules are blue. The metal ions (1 and 1′) are represented with green circles which are attached to each of the aspartate residues (Asp 20 and Asp 20′). The metal ion (2) is shared between the overlapping active sites. The arrow shows the inline hydrophilic attack on the scissile phosphate.
Stoddard BL. 2006. Homing endonuclease structure and function. Q. Rev. Biophys. **38**(1): 49-95. (Springer Publications. Image reproduced with permission. License number: 3831560857433).

1.8.2. GIY-YIG family

The GIY-YIG is a short module of 70-100 amino-acids characterized by the two conserved sequence motifs flanking a 10-11 amino-acid segment. The signature motifs for this family start with `GIY' and end with `YIG' tripeptides i.e. GIY-(X_{10-11})-YIG. Members of this smaller family of endonucleases have been identified as both freestanding (F-TevI, F-TevII; Sharma *et al*., 1992) and within mobile group I introns (I-TevI, I-TevII; Bell-Pedersen *et al*., 1990); however, one example has been noted within a group II intron of the protozoan *Amoebidium parasiticum* (Li *et al*., 2011). Moreover, they have been identified from introns within fungal and algal mitochondrial genomes (Saguez *et al*., 2000; Kroymann and Zetsche, 1997), chloroplast genomes (Paquin *et al*., 1995), and bacteriophage genomes (Sharma *et al*., 1992).

Besides mediating intron homing, the GIY-YIG endonuclease domain serves additional functions. This domain is found within protein scaffolds that participate in diverse cellular pathways such as assisting other DNA-binding proteins: the UvrC subunit of the archaeal and bacterial UvrABC DNA excision repair complex with the UvrC providing the required endonuclease activity (Dunin-Horkawicz *et al*., 2006).

Biochemical approaches such as DNA footprinting and limited proteolysis analysis performed on monomeric 28 kDa I-TevI protein and other members of this family have shown that unlike LAGLIDADG family, these enzymes have a modular structure (bipartite structure) where the carboxyl-terminal DNA-binding domain and the amino-terminal catalytic domain are separated by a long flexible linker (Bryk *et al*., 1993; Derbyshire *et al*., 1997). Although homotetrameric version (Cfr42I) from this family has been reported (Gasiunas *et al*., 2008), the members typically act as monomers and contain a single active site that hydrolyzes DNA by a

one-metal ion mechanism. For example, I-BmoI functions as a monomer at all steps of the reaction pathway and does not transiently dimerize or use sequential transesterification reactions to cleave its substrate. Instead, the DNA-binding domain acts as a molecular anchor to tether the GIY-YIG domain to the substrate, permitting rotation of the nuclease domain to sequentially nick each DNA strand (Kleinstiver *et al*., 2013). This group of enzymes after binding primarily across the minor groove and phosphate backbone significantly distort their DNA homing site and cleave the substrate yielding 2-nt, 3′ overhangs (Loizos *et al*., 1996). The members have been noted to recognize a long homing site (~ 31-40 bp) and cleave their respective DNA target sites many bases away from the intron-insertion site. For example, the cleavage site bound by the cleavage domain of I-TevI is ~25 bp away (Mueller *et al*., 1995; Bryk *et al*., 1995). One exception being the *Bacillus thuringiensis* ssp. *pakistani nrdF* intron-encoded HEase, I-BthII with an unconventional GIY-$(X)_8$-YIG motif that cleaves an intronless *nrdF* gene only 7-nt upstream of the intron-insertion site, typical for LHEases (Nord and Sjöberg, 2008; Guha and Hausner, 2014).

The cleavage domain has a mixed α/β topology and the GIY-YIG residues are located in the three-stranded β-sheet as evident from nuclear magnetic resonance studies (Kowalski *et al*., 1999). Apart from these motifs, two highly conserved residues residing in the α-helices are important for endonucleolytic activity (Kowalski *et al*., 1999). Like other endonucleases, divalent cations are essential for DSB formation (Bryck *et al*., 1993). However, nicks can be created without the metal ions *in vitro* (Bryck *et al*., 1993).

1.8.3. H-N-H family

The H-N-H family includes protein members which contain a consensus sequence spanning 30-33 amino-acids with two pairs of conserved histidine flanking a conserved asparagine (Shub *et al*., 1994; Gorbalenya, 1994). Based on the exact sequence of the core H-N-H signature and on the presence or absence of uniquely conserved flanking residues, this family can be further divided into at least eight subfamilies (Mehta *et al*., 2004). Recently, a thermostable DNase, EheA isolated from *Exiguobacterium* sp. yc3 is the first experimentally determined bacterial source endonuclease belonging to the second H-N-H subfamily (H-N-N; Zhou *et al*., 2015).

Although very few structural and biochemical studies have been performed, this conserved endonuclease catalytic domain has been found in several well-known and some obscure genomic locations such as group I intron-encoded HEases (e.g. I-HmuI and I-HmuII; Goodrich-Blair *et al*., 1990; Goodrich-Blair and Shub, 1994), proteins encoded by mobile group II introns (e.g. I-SceV, I-LlaI; Zimmerly *et al*., 1995a; Shearman *et al*., 1996), non-specific antibacterial endonucleases (colicins E7, E9; Ko *et al*., 1999; Kleanthous *et al*., 1999), DNA repair enzyme (MutS; Malik and Henikoff, 2000), DNA rewinding motor enzyme (annealing helicase 2; Yusufzai and Kadonaga, 2010), cas9 nuclease (Jinek *et al*., 2014), the *nrdB* intron of RB3 bacteriophage (I-TevIII; Eddy and Gold, 1991), *Phormidium foveolarum* phage Pf-WMP3 (I-PfoP3I; Kong *et al*., 2012), and the *psbA* gene of *Chlamydomonas moewusii* chloroplast (I-CmoeI; Drouin *et al*., 2000). Recently, this motif has also been reported to be a widespread component of phage DNA packaging machines (Kala *et al*., 2014).

The structure of the best studied member, I-HmuI has been determined in complex with its bound cognate DNA target site which spans 25 bp (Shen *et al*., 2004). The structure of this

protein is extraordinarily elongated with a series of distinct sequential structural domains and motifs distributed along the DNA target. The carboxyl-terminal helix-turn-helix (HTH) motif binds the target site within its major groove while an amino-terminal antiparallel β-sheet binds the opposite end within its major groove, along with two α-helices bind along its middle seven base pairs within the minor groove. A nuclease active site consisting of a ββα associated with one catalytic metal ion (ββα-Me) binds across the minor groove and interacts nonspecifically with the DNA backbone near the site of cleavage (Shen *et al*., 2004). The crystal structure of Gmet_0936 protein, a putative H-N-H endonuclease from *Geobacter metallireducens* GS-15 reveals a zinc ion coordinated by various cysteine residues bound in each monomer which likely plays an important structural role in stabilizing the overall conformation (Xu *et al*., 2013). The DNA displays a distinct bend of 40° approximately 4-5 bp downstream to the site of cleavage, corresponding to a significant widening of the minor groove in that region (Xu *et al*., 2013).

The H-N-H family members display significantly diverged biochemical properties. For example, I-HmuI cleaves only one strand of its DNA substrate (Landthaler *et al*., 2004) whereas I-TevIII and I-CmoeI generate DSBs with 5′ and 3′ overhangs respectively. Yet extraordinary are the group II intron-encoded I-SceV and I-SceVI which form RNP complexes with their respective intron RNAs where both the components participate in cleavage. It has been demonstrated that the protein-mediated cleavage of one strand was hampered when the H-N-H component was mutated (Zimmerly *et al*., 1995b). However, normal nicking was observed on the other strand which was mediated by the intron RNA component (Zimmerly *et al*., 1995b; Matsuura *et al*., 1997).

1.8.4. His-Cys box family

A smaller group of HEases are characterized by two conserved histidines (His) and three conserved cysteines (Cys) within a 30 amino-acid stretch which are likely to form a metal coordination site constitute the His-Cys box family (Muscarella *et al*., 1990). Although it is expected that a sequence rich in His-Cys residues has the potential to bind zinc atoms in the active site, it does not align well with the consensus sequence of any previously identified zinc binding domain (Flick *et al*., 1998).

The distribution of this family is very limited, being only encoded within the known mobile group I introns present in eukaryotic nuclear genomes such as highly conserved regions of small and large subunit of ribosomal DNA of slime molds, fungi and amoebae (Johansen *et al*., 1993).

Even though experiments for intron mobility have not been performed for closely related group of His-Cys box proteins encoded from *Naegleria* (I-NanI, I-NjaI, I-NitI; Elde *et al*., 1999, 2000), the I-PpoI from *Physarum polycephalum* (Muscarella *et al*., 1990) and I-DirI from *Didymium iridis* (Johansen *et al*., 1997) have been experimentally shown to promote mobility of their respective introns in their natural hosts. Biochemical and structural studies of I-PpoI revealed the significance of the His-Cys box along with the detailed cleavage mechanism. One interesting observation is that both the His-Cys box family and LAGLIDADG family are similar in some aspects and dissimilar in other aspects (Chevalier and Stoddard, 2001). Although I-PpoI monomer (18 kDa) consists of two α-helices and ten β-helices folded into three separate β-sheets, these two families share a mixed α/β topology. Similar to monomeric LAGLIDADG enzymes, I-PpoI recognizes a 14 bp pseudo-palindromic site on major groove of DNA homing site and cleaves minor groove to generate 4-nt, 3′-OH overhangs. Unlike I-CreI (a well-studied

LAGLIDADG enzyme) which display extended protein folds surrounding long β-ribbon platforms, I-PpoI lacks a tight hydrophobic core packaging and severely bends and distorts its DNA substrate to widen the minor groove enough (~20Å) to place the scissile phosphates into the active sites of the enzyme (Ellison *et al*., 1993).

Even though all of these above families appear to evolve independently, they are usually tolerant towards minor variations in DNA homing target sequences and optimized in their ability to recognize long DNA targets for efficient homing process. However, it has been demonstrated that chromatin compaction (Daboussi *et al*., 2012) and presence of 5′-cytosine-phosphoguanosine (CpG) methylation affect the activity of both natural and engineered meganucleases *in vivo* (Valton *et al*., 2012). Moreover, sequestration/localization of most HEGs within mobile elements or within specific compartments of the cell along with low protein expression profiles prevent the host genome from non-specific nuclease activity.

1.9. Beneficial functions

The impression of mobile intron and its encoded HEase is that they mostly bear a negative consequence in the host genome, i.e., HEases and their carriers, introns/inteins are 'selfish elements' or 'free-loaders' in the genome. There are rare instances where HEases stand out for providing a benefit to the host, and thus these HEases can escape the invasion/elimination cycle (Stoddard, 2006). An intron-encoded bifunctional LHEase, I-AniI acts as a 'maturase' which corresponds to the direct interaction of the IEP with the surrounding intron in a high-affinity binding which is required to act as a 'chaperone' to the RNA for effective splicing. This activity might have provided enough selective pressure to maintain the intron encoded ORF (Ho and Waring, 1999).

DSBs have also been shown to initiate a mitotic recombination reaction, one of which is known as the yeast mating type switching (Herskowitz *et al*., 1992). Haploid yeast cells come in two mating types, designated *a* and *α* and these mating types are determined by a master regulatory locus on chromosome III called *MAT* (mating type) which contains four identical sequence blocks: W, X, Z1 and Z2; however, each mating type contains a unique *MAT* Y sequence: Y*a* in *a* cells and Y*α* in *α* cells. In this aspect, an intein-encoded LHEase, HO (closely related to the endonuclease domain of PI-SceI) has been shown to be responsible for the 'mating type switch' in *S. cerevisiae* by making a DSB at the *MAT* YZ border. In order to repair the gap, the new Y*a* or Y*α* information transposes into *MAT* from two unlinked, silent cassettes known as *HMLα* (Hidden Mat Left) and *HMRa* (Hidden Mat Right) thereby switching the cell's mating type. This process allows haploid yeast cells of one mating type to produce haploid cells of the other type, thereby allowing sister cells to mate and become diploid. Diploid cells, in turn, can sporulate to produce haploid cells and offer yeast the advantages coincident with a sexual life cycle (Jin *et al*., 1997).

The PI-SceI HEase present within some species of *Saccharomyces* is another instance for the beneficial role of a HEG. This protein has shown to increase the sporulation speed of homozygotes by around 40% compared to the ones lacking the protein (Giraldo-Perez and Goddard, 2013). This increased promiscuity allowed PI-SceI HEase to invade yeast populations 20 times more rapidly compared with when there is no effect on promiscuity. This increased sporulation rate proves advantageous in the carrier cell's fitness because yeasts generally sporulate to more resistant spore structures within ascus (i.e. ascospores) when unfavourable conditions are encountered compared to the vegetative budding stage when conditions are favourable (Giraldo-Perez and Goddard, 2013). Moreover, the spores survive better than

vegetative cells in the gut of a known vector which is *Drosophila* (Reuter *et al*., 2007). Therefore, this suggests a *VDE* carrier would enjoy increased survival under harsh conditions.

The T4 phage-encoded I-TevI endonuclease not only displays DNA-cleavage activity, but also acts as a transcriptional autorepressor of its own expression by binding a DNA sequence that overlaps a late promoter within the 5′ region of its own ORF. Thus it represents another system in which a HEase displays a secondary function with potential benefit to the host (Edgell *et al*., 2004).

HEases have been useful for survival in some species. For example, during mixed infections, members of the H-N-H family, I-HmuI and I-HmuII from closely related *Bacillus subtilis* phage prove advantageous to their host by specifically cleaving the DNA of the heterologous phage (Goodrich-Blair and Shub, 1996).

Being mostly selfish or sometimes beneficiary to the host may explain the overall biology of the HEGs and their related homing life cycle. However**,** these HEases by virtue of their rare-cutting activity are harnessed as targetable nucleases for biotechnological applications such as genome editing (Silva *et al*., 2011; Takeuchi *et al*., 2011). However, so far only a very limited number of well characterized native LHEases (I-SceI, I-CreI, I-DmoI, I-AniI, and I-OnuI; Jacoby *et al*., 2012; Prieto *et al*., 2012) are used in genome engineering. Therefore, one should either bioprospect for more native HEases (Sethuraman *et al*., 2009; Hafez and Hausner, 2012) or reprogram existing scaffolds of the above HEases along with engineering synthetic versions so that a wider choice of genomic sites can be targeted.

1.10. HEases reprogramming

The crystal structure of the various DNA-bound HEases allowed identification and determination of essential amino-acid residues that are in close contact with the base pairs in their respective DNA targets (Jurica *et al*., 1998; Moure *et al*., 2003; Chevalier *et al*., 2003). This led to the conclusion that not all residues present in the active site are equally important in DNA-binding. Therefore, altered specificity variants of LHEases could be generated (Arnould *et al*., 2006) along with development of various selection systems designed to isolate and characterize these mutant enzymes (Seligman *et al*., 2002; Gruen *et al*., 2002; Jarjour *et al*., 2009).

1.10.1. Alteration at individual base pairs

The rational design based on the crystal structure of the various DNA-bound HEases (as described above) coupled with randomization and screening in a high-throughput format resulting in altered specificity took shape after solving the various crystal structures of both monomeric and dimeric LHEases (Chevalier *et al*., 2001; Bolduc *et al*., 2003; Arnould *et al*., 2006). Earlier studies have shown that mutation of individual DNA-binding residues of HEases resulted in a change in the target site specificity. The identification of such mutated constructs (new variants) that displayed altered recognition specificity were assessed using the bacterial two-hybrid screening strategy (Gimble *et al*., 2003) or methods that coupled endonuclease activity to the elimination of a toxic reporter gene (Seligman *et al*., 2002; Gruen *et al*., 2002). In parallel to this approach, redesigning of LHEase DNA-binding and -cleavage specificities were also achieved through sophisticated computational programming (Ashworth *et al*., 2006, 2010). Notably, in case of the I-CreI, thousands of variants of the enzyme targeting 16 different base pair positions in the 22 bp I-CreI target site have been generated by RosettaDesign (RD; Ulge *et*

al., 2011) of which, over two-thirds have the intended new site specificity. Assisting software FoldX further determines the energetic effect of point mutations as well as the interaction energy (i.e., the total energy caused by an interaction between the objects being considered) of protein complexes including protein-DNA (Schymkowitz *et al*., 2005). However, as an alternative for reengineering, one can still search for naturally occurring enzymes to increase the number of LHEases with different target sites (Sethuraman *et al*., 2009; Hafez and Hausner, 2012).

1.10.2. Hybrid endonucleases

Expanding HEase specificity can also be attained by creating hybrid endonucleases such that these chimeric enzymes recognize corresponding hybrid target sites. Earlier studies demonstrated that creation of novel chimeric HEases capable of recognizing corresponding chimeric DNA target sites could be made by mixing/matching and fusing the entire domains or subunits from unrelated LAGLIDADG enzymes (Chevalier *et al*., 2002; Steuer *et al*., 2004). For example, a highly active artificial enzyme was created by fusing domains from two different LHEases, I-DmoI and I-CreI. This hybrid endonuclease was named H-DreI which also recognized a hybrid substrate derived from the original parent donors (Grizot *et al*., 2010). However, recent approaches are focused mainly on systematic exchange of domains between HEases with ~40 to 70% sequence identity between the individual proteins, selected from relatively closely related clades e.g., from mesophilic fungal mitochondrial genomes (Baxter *et al*., 2012). Another example of hybridization strategy is to generate monomerized versions of homodimeric HEases by combining two single motif LHEases using short (∼33 residues) linker peptides. For example, I-CreI or I-MsoI peptides have been fused with respective canonical HEase amino sequence generating monomerized versions (Li *et al*., 2009). Catalytically inactive

I-SceI fused to the restriction enzyme PvuII as the cleavage module represents another type of hybrid endonuclease (Fonfara *et al*., 2012) that cleaves double-stranded DNA at a defined site outside the original I-SceI target site (Lippow *et al*., 2009).

Although LHEases are architecturally well adapted for therapeutic applications, tightly coupled cleavage and binding domains of these enzymes have limited the engineerability of the DNA-binding interface resulting in a compromised functionality and targeting novel DNA sites (Takeuchi *et al*., 2011). To address this limitation, a novel designer rare-cleaving endonuclease was designed by fusing DNA-binding transcription activator-like effectors (TALEs) to the amino-terminal nuclease domain of a meganuclease I-AniI. This modular architecture named 'megaTAL' has shown to rescue the activity and boost repair activity well beyond that of their high affinity HEase counterparts (Boissel *et al*., 2013).

Even though the non-homologous end-joining (NHEJ) pathway is usually responsible for introducing mutations at the DSBs within the genome (Shrivastav *et al*., 2008; Lieber *et al*., 2011), sometimes, DSBs are non-productive for genome engineering because they are repaired without any intended mutation (Wolfs *et al*., 2014). This may be due to the formation of compatible cohesive ends generated by the nucleases that are re-ligated through the typical NHEJ pathway which utilizes short homologous DNA sequences called microhomologies to guide repair (McVey and Lee, 2008; Wolfs *et al*., 2014). These microhomologies are often present in single-stranded overhangs on the ends of DSBs and when these overhangs are perfectly compatible, NHEJ usually repairs the break accurately which is not desirable for gene disruption (Wilson and Lieber, 1999; Budman and Chu, 2005). One strategy to bias repair events toward gene disruption is to co-express a DNA end-processing enzyme with the genome-editing nuclease. Coexpression of designer endonucleases with DNA end-processing enzyme (e.g.

TREX2, which is a 3′-5′ exonuclease) *in trans* has been employed to bias repair events towards gene disruption (Certo *et al*., 2012). However, transfection of the *Trex2* coding region with meganuclease constructs seemed problematic in size-constrained vectors and overexpression of this exonuclease may cause mutagenic repair at unwanted off-target sites (Delacôte *et al*., 2013). Therefore, coupling two different nuclease active sites into a single polypeptide could be a useful strategy for gene disruption. Recently, the 'MegaTev' architecture has been created. It is the fusion of a meganuclease (Mega) with the nuclease domain derived from the GIY-YIG HEase, I-TevI (Tev) designed to position ~30 bp apart on DNA substrate and generate two DSBs with non-compatible cohesive ends. Without overexpressing the DNA-end processing enzymes, high gene disruption was observed in a HEK 293 cell line when this dual active MegaTev was expressed (Wolfs *et al*., 2014).

1.10.3. Nicking endonucleases (Nickases)

Site-specific nicking HEase or 'nickases' can be viewed as promising candidates since single-strand DNA nicks are less prone to recombination events than double-strand breaks. The single-strand DNA nicks are repaired by the error prone NHEJ reactions thereby reducing cellular toxicity (Belfort and Bonocora, 2014). Various studies have demonstrated that the GIY-YIG endonucleases can perform sequential cleavage of their DNA target sites while the H-N-H enzymes perform site-specific nicking of DNA single-strand. More importantly, these enzymes are effective in gene conversion events (Mueller *et al*., 1995; Carter *et al*., 2007; Landthaler and Shub, 2003) that do not involve NHEJ. Therefore, an effort has been carried out in order to transform double strand cutters LHEases into nickases. In a nicking variant of I-AniI (I-AniI K227M), an engineered mutation of a basic residue essential for proton transfer and solvent

activation in one active site has been generated. This artificially constructed nickase pair has shown to cleave its DNA target site ~8-fold faster than wild-type I-AniI to generate a DSB (McConnell Smith *et al*., 2009).

1.10.4. Inserting ribozyme based switch

Protein-responsive ribozyme switches (Kennedy *et al*., 2014) and ribozyme based aminoglycoside switches (Klauser *et al*., 2015) are among the growing class of genetic controllers applied in synthetic biology to engineer cellular functions in eukaryotic and mammalian cells (reviewed in Bradley *et al*., 2016).

In biotechnological applications, it is sometimes desirable to control the activity of a HEase in order to reduce genotoxicity. Previously, a molecular switch was developed that controlled the endonuclease activity of PI-SceI *in vitro.* In that study, two cysteine amino-acid residue pairs were separately inserted into the HEase DNA-binding loops to allow for disulfide bond formation that locks the endonuclease into a nonproductive conformation. However this approach was not practical for *in vivo* applications as the targeted cell would probably suffer damage if oxidizing conditions were applied (Posey and Gimble, 2002).

1.11. Applications

Early approaches to genome engineering relied heavily on HR. For example, the modification of the genomic DNA of mouse involved the modification of embryonic stem cells in culture followed by injecting these engineered cells into mouse embryos (Capecchi, 1980; Thomas *et al*., 1986; Thomas and Capecchi, 1987). However, modifying genomes depending on this singular factor (i.e., the cell's intrinsic HR mechanism) remained a challenging

process. The introduction of a site-specific I-SceI endonuclease into a mammalian genome in somatic cell types was the first experiment to show that increased efficiency of a site-specific sequence conversion event is possible through induced rate of HR mechanism (Choulika *et al*., 1995). If genome engineering could be triggered by introducing DSB at the site of interest, then REases could have done this task efficiently. The disadvantage being, they are likely to cut the DNA molecule several times due to their very short recognition site (see sub-section 1.7.1). Therefore, 'genome surgery' needs more precise 'molecular scissors' which can offer a higher degree of specificity. To their advantage, HEases recognize longer target sites (14-44 bp), so the probability of finding the target site elsewhere will be a rare occurrence in the genome. Moreover, they are highly specific and also tolerant to some target site degeneracy with reduced fidelity which often corresponds to the wobble position within the reading frames of the protein-coding host genes at the HEase recognition site (Edgell *et al*., 2004; Scalley-Kim *et al*., 2007; Barzel *et al*., 2011). These characteristics facilitate using HEases as potential tools in different genomic applications.

1.11.1. HEases as therapeutic agents

Soon after the discovery of I-SceI-induced DSB repair, LHEases were used to stimulate recombination in mammalian cells (Jasin, 1996). Although genome engineering using HEase is not absolutely accurate, this approach is far more achievable compared to random transgenics using viral vectors that can integrate indiscriminately causing off-target issues (Williams and Baum, 2003; Davé *et al*., 2009). However, tail vein injection of adenovirus that expressed the I-SceI meganuclease was successful in being delivered in the liver of mice and the HEase induced targeted genomic recombination (Gouble *et al*., 2006). Recently, I-SceI was also used for the

correction of Artemis deficiency Art (-/-) in murine hematopoietic stem cells (mHSCs). In this approach, the I-SceI enzyme and the Artemis correction template were each delivered by a self-inactivating (SIN)-integrase-defective lentiviral vectors. Reversion of the mutant phenotype into wild type with $1/5^{th}$ success rate was observed when the Art (-/-) mHSCs were transduced with these two vectors. Even with low success rate and potential for genotoxicity yet to be evaluated, this new approach to gene editing appears to be promising (Rivière *et al*., 2014). In other instances, engineered I-CreI endonucleases were designed to correct the *XPC1* gene, the *RAG1* gene, dystrophin gene in cells from patients suffering from Xeroderma pigmentosum (XP; Arnould *et al*., 2007), severe combined immunodeficiency disease (SCID; Grizot *et al*., 2009), and Duchenne muscular dystrophy (DMD; Chapdelaine *et al*., 2010) respectively. In another example, a sevenfold increase in gene disruption of the endogenous HIV coreceptor *CCR5* was observed with an engineered I-CreI coupled to a DNA end-processing enzyme (Trex2) over the untailored nuclease (Certo *et al*., 2012). Recently, pre-treatment of human corneas *ex vivo* with a specific meganuclease HSV-1m2 before transplantation into herpetic keratitis patients avoided the recurrence of the infection and allograft rejection in corneal treatments (Elbadawy *et al*., 2014).

1.11.2. HEases in facilitating transgenesis

Transgenesis is the process of deliberate modification of a genome by introducing an exogenous gene (transgene) into a living organism so that the organism exhibits a new trait and its offspring inherits this characteristic (Gordon and Ruddle, 1981). In the past, transgenic zebrafish were generated by microinjecting plasmid DNA into early embryos. The mosaic distribution of the injected transgene and late integration within genome were the two main

hurdles which made generation of transgenic lines difficult (Soroldoni *et al*., 2009). However, coinjection of I-SceI meganuclease with a reporter construct flanked by I-SceI sites overcomes these problems by earlier transgene integration into the host genome. This same enzyme has been effectively used to facilitate transgenesis in *Xenopus tropicalis* (Ogino *et al*., 2006), sea urchin embryos (Ochiai *et al*., 2008), zebrafish eggs (Soroldoni *et al*., 2009) for nearly a decade. However, the native I-SceI is not capable of facilitating transgenesis in mammalian embryos via cytoplasmic microinjection as it did for the above examples. Recently, I-SceI stitched to a nuclear localization signal (i.e., NLS-I-SceI) has been able to transfer DNA fragments from cytoplasm into nuclear compartment thereby mediating germline transgenesis in both mice and porcine embryos (Wang *et al*., 2014).

1.11.3. HEases as mutagenic agents

HEases have been used to disrupt the genetic material of an organism at specific sites by introducing DSB in which, mainly the NHEJ based repair process creates indels surrounding the cleavage site. This allows for site-directed mutagenesis in a variety of organisms such as bacteria, fungi, plants, insects, and metazoans (Wong, 2004; Flannagan *et al*., 2008; Siegl *et al*., 2010; Wang *et al*., 2012; Lyznik *et al*., 2012). Engineered variant of I-AniI (Y2 I-AniI) HEases have been used to develop methods to selectively disrupt integrated HIV proviruses within latently infected immortalized human cell lines (Aubert *et al*., 2011).

Introducing a DSB has another advantage. It increases the frequency of mutations above the natural background level. For example, a system referred to as *delitto perfetto* is a genetic method for *in vivo* site-directed mutagenesis in *S. cerevisiae* where the I-SceI was used to generate a DSB at the appropriate target site. As a result, the frequency of targeted HR increased

by 4000-fold with the engineered DNA compared to the experiments where DSBs were not generated (Storici and Resnick, 2003; Storici *et al*., 2003).

1.11.4. HEases in curbing pest population

Insects play a major role as vectors of human diseases as well as causing significant agricultural losses. Harnessing the activity of customized HEases has been proposed as a method to curtail the spread of dreadful diseases like malaria by promoting the spread of 'engineered' insects through populations, subsequently reducing the number of disease causing malarial parasites (e.g., *Plasmodium falciparum*). The I-PpoI HEG in male mosquitoes was capable of introducing high levels of infertility in target populations in cage trials, and therefore has potential in malaria control. This HEG under the control of a male spermatogenesis specific promoter (β-tubulin promoter) has been inserted into the Y chromosome in *Anopheles gambiae*, which when expressed during meiosis could cut a sequence inserted within the repetitive DNA encoding a gene coding for 28S ribosomal RNA located on the X chromosome. Even though this process allows for the normal development of the zygotes, shredding of the paternal X chromosome favours a severe male-biased sex ratio (Windbichler *et al*., 2008; Deredec *et al*., 2011; Klein *et al*., 2012). In a separate study using the I-SceI endonuclease, it was shown that higher rates of homing can be achieved within spermatogonia and in the female germline and more importantly, the 'homed constructs' continue to exhibit HEase activity in the subsequent generations essential for the successful deployment of a HEG-based gene drive strategy (Chan *et al*., 2011, 2013).

The strategy involving an engineered genetic construct that is heritable and biases offspring sex ratio towards males has expanded beyond controlling the mosquito population. For

the biological control of carp and more recently cane toads, an organization abbreviated as CSIRO operated by the Australian government adapted this 'daughterless technology'. With a distorted sex ratio containing fewer females and mostly males, it is predicted that the carp population will significantly decrease with repeated releases over 25 years through to pseudo-extinction within 75-90 years in the Murray-Darling Basin of Australia (Saunders *et al*., 2010).

1.11.5. HEases in agronomy

Site-directed sequence modifications using HEases have been used in studies of both model and crop plant species. For example, a proof-of-concept (POC) study was undertaken in *Arabidopsis thaliana* where I-SceI-induced DSB lead to a much more efficient gene targeting (GT) system that is independent of transformation (Fauser *et al*., 2012). In another example, engineered I-CreI endonuclease was used for targeted mutagenesis in maize that was designed to produce DSBs at the *liguleless1* chromosomal locus (Guo *et al*., 2010). Recently, GT based on DSB induction by the I-SceI and a transgenic artificial target locus was employed in barley and this approach could be developed as a routine genome editing tool for other important cash crops (Watanabe *et al*., 2016).

One of the most important issues in plant genome editing is delivery and expression of the engineered nucleases in plant cells, since not all useful plant species are amenable to regeneration and transgenic methods. Re-engineered HEases were also used for developing transformation vector systems for plant genome editing and targeted mutagenesis (Yang *et al*., 2009; Vainstein *et al*., 2011; Lyznik *et al*., 2012). Recent studies using viral vectors with HEases such as recombinant adeno-associated virus (Ellis *et al*., 2013) and geminivirus-based replicons

(Baltes *et al*., 2014) to deliver the DNA containing genome editing tools have enabled efficient genome engineering in various plant species.

1.12. Other genome engineering platforms

Apart from HEases or meganucleases and their derivatives, four major classes of nucleases such as zinc finger nucleases (ZFNs; Kim *et al*., 1996; Pernstich and Halford, 2012), transcription activator-like effector nucleases (TALENs; Boch *et al*., 2009; Li *et al*., 2011), targetrons (Lambowitz and Zimmerly, 2011) and clustered regularly-interspaced short palindromic repeats (CRISPR) associated Cas9 nucleases (Bhaya *et al*., 2011; Jinek *et al*., 2012) exist that enable site-specific genome engineering. Based on the various modes of DNA recognition, these nuclease platforms can be broadly divided into two groups. First group consists of meganucleases, ZFNs and TALENs which achieve sequence-specific DNA-binding via protein-DNA interactions. The second group comprises of targetrons and CRISPR-Cas9 where modified group II intron lariats and short sequence-specific guide RNA molecules aid in base-pairing directly with the target DNA sequence respectively (Enyeart *et al*., 2014; Cox *et al*., 2015). However, one stand-alone group exists which depends on the sequence-specific DNA ligands instead of the protein-DNA based or RNA-DNA based binding. These are known as the triplex-forming oligonucleotide (TFO) nucleases where a type II REase (e.g. PvuII) domain is coupled to TFO which can be engineered according to the target site specificity (Eisenschmidt *et al*., 2005; Schleifman *et al*., 2008).

Several excellent reviews exist for learning the mechanism, engineering, application, current progress as well as future prospects and challenges of each of the above genome editing tools (Gaj *et al*., 2013; Sander and Joung, 2014; Kim and Kim, 2014; Stoddard, 2014; Cox *et al*.,

2015). Therefore, this section will refrain from being a redundant one, instead it will briefly describe the above key 'players' and conclude by providing the present scenario and importance of LHEases with respect to the other programmable nuclease platforms, especially those which are relevant in the recent era of genome engineering.

ZFNs are artificial endonucleases that have been generated by combining a small zinc finger (ZF; ~30 amino-acid) DNA-binding/recognition domain (Cys_2His_2) to the type IIS nonspecific DNA-cleavage domain from FokI. One ZF module recognizes a 3 bp sequence and cleavage activity of the FokI endonuclease demands dimerization (Cathomen and Joung, 2008; Urnov *et al*., 2010). Even though ZFNs showed impressive results in modifying the HIV CCR5 co-receptor surface protein in the autologous CD4 T lymphocytes of persons infected with HIV (Tebas *et al*., 2014), the risk of modular synchronization between the ZF and the endonuclease domain still lurks leading to cleavage at ectopic sites (Urnov *et al*., 2005; Pattanayak *et al*., 2011).

TALENs are artificial endonucleases designed by fusing the DNA-binding domain (multiple nearly identical repeats, each comprised of ~34 amino-acids) obtained from TAL effector (TALE) protein to the nonspecific DNA type IIS cleavage domain from the FokI endonuclease (Christian *et al*., 2010). Each TALE repeat independently recognizes its corresponding nucleotide base and despite a tolerance of mismatches of longer TALENs *in vitro*, they seem to have higher genome editing activity and less toxicity than ZFNs (Chen *et al*., 2013). TALEs can be redesigned to bind to other user-defined sequences by simply joining appropriate repeat units. Like ZFNs, they require dimerization partner for effective cleavage (Boch *et al*., 2009).

The targetron is a RNP that consists of an engineered group II intron RNA lariat molecule and a multidomain IEP which has been used for mutagenesis in bacterial genes (Lambowitz and Zimmerly, 2011). The strategy is based on group II retrohoming (see sub-section 1.5.2.) where the intron lariat recognizes the native DNA target by an exon binding sequence (EBS) base-pairing with the DNA over ~14 bp (Mohr *et al*., 2000). Although compromised activity is observed in eukaryotes and mammalian system due to the suboptimal codon usage, translational repression of the RT, nonsense-mediated decay (NMD) of group II intron-containing RNAs and suboptimal magnesium ion (Mg^{+2}) concentrations (Truong *et al*., 2015), this RNA-guided endonuclease (RGEN) has shown potential for high site-specific retargeting in prokaryotes by reprogramming the intron EBS (Karberg *et al*., 2001; Lambowitz and Zimmerly, 2011). Even though group II introns have the potential to integrate into DNA by creating site-specific DSBs, the entry of the targeting RNA in the form of a RNP into the nucleus along still remains the major obstacle for applications for targetrons among eukaryotes (Cui and Davis, 2007; Enyeart *et al*., 2014)

The components derived from the bacterial immunity system: duet of CRISPR locus and Cas9 nonspecific endonuclease (CRISPR/Cas9) is a novel RGEN for precise and efficient gene targeting (Jinek *et al*., 2012; Mali *et al*., 2013; Calos, 2016). The uniqueness of this platform is based simply on designing guide RNAs (gRNAs) since the Cas9 nuclease does not require any engineering for retargeting. First, the gRNA attaches the Cas9, complementary base pairing allows the gRNA sequence (~18-20 nt) to hybridize with the targeted DNA sequence as a result docking the Cas9 nuclease at that location. The H-N-H and the RuvC nuclease domain of the Cas9 cleave both DNA strands to create DSBs provided target site that must lie immediately 5′ of a PAM (protospacer adjacent motif) sequence. Eventually, cellular DSB repair mechanism

either creates or rectifies a mutation based on the presence or absence of the user-provided corrective template. In addition, this system has been successfully modified to accommodate mammalian transcription and translation requirements (Qi *et al*., 2013; Maeder *et al*., 2013). Although multiplexed editing is very much possible through this system, the requirement for PAM sequence and genomic DNA accessibility due to chromatin and methylation states are major potential constraints. The risk of off-target issues still exists, henceforth, gRNA modifications like paired Cas9 nickases (Ran *et al*., 2013) and truncated gRNA (Fu *et al*., 2014) have been constructed which have shown promising results.

The current genome editing technologies have shown immense success, and thereby revolutionized the field of life science. It is tempting to comment at this juncture that the practice of stimulating HR, mediated by various ZFNs, TALENs or CRISPR/Cas9 systems has come to fruition only after the first reports of using the I-SceI HEase in mammalian cells were published. Utilizing the above state of the art tools, especially CRISPR/Cas9, researchers have been able to disrupt specific genes, introduce single-nucleotide substitutions, add exogenous DNA into intended genomic sites and perform many other applications like construction of transgenic model organisms and corresponding cell lines. In all these cases, targeting specificity may be compromised to some extent. It is true that the engineering of LHEases has been challenging because the DNA recognition and cleavage functions of these enzymes are intertwined in a single domain. However, therapeutic applications do demand the highest precision in gene modification activity through the highest level of target specific ‘molecular scissors’. For such applications, continued development of compact, highly specific nuclease domains that do not rely upon additional DNA or RNA targeting moieties will be of value. LHEases may always be in demand as components of vector/cloning systems that require rare-cutting enzymes.

1.13. Research objectives

Previous studies by Hafez and coworkers discovered a novel twintron (nested intron) inserted at position mS1247 within the small ribosomal subunit (*rns*) gene of a thermophilic fungus, *Chaetomium thermophilum* DSM 1495 (Hafez *et al*., 2013). This twintron (nested intron) is composed of a group IC2 intron encoding a double motif LAGLIDADG HEase interrupted by an ORF-less group IIA1 intron (Hafez *et al*., 2013). It is hypothesized that splicing of internal group II intron would reconstitute the LAGLIDADG ORF, thereby produce an active HEase. This twintron (nested intron) arrangement offers the opportunity to examine if the nested group II intron could be utilized as a regulatory element for the expression of the HEase.

Previous studies also indicated that the fungal mitochondrial encoded cytochrome oxidase b (*cyt-b*) gene (unpublished) and the small-subunit ribosomal RNA (*rns*) gene appears to be a reservoir for a number of group I and II introns along with IEPs such as HEases and reverse transcriptases. Therefore, it would be an excellent opportunity to bioprospect for native and active HEases which will provide an attractive alternative to the labour intensive protein engineering required to increase the target site repertoire for genome engineering purposes. Therefore, the specific research objectives of this thesis are as follows:

1.13.1. Biochemical characterization of the twintron (nested intron) encoded LHEase from *Chaetomium thermophilum* DSM 1495.

1.13.2. Applying group II introns in order to the attenuate *in vitro* and *in vivo* expression of a functional LHEase characterized in objective **1.13.1.**

1.13.3. Bioprospecting for native LHEases from i) the c490 intron of the *cyt-b* gene in *Ophiostoma novo-ulmi* subspecies *americana* and ii) the mS569 intron of the *rns* gene in *Ophiostoma ulmi*.

Chapter 2

General Materials and Methods

2.1. Chemicals and common reagents

Most of the chemicals and common reagents, unless otherwise stated were purchased from ThermoFisher Scientific (Mississauga, Ontario, Canada). The bacterial growth media were purchased from Difco™ BD chemicals (Mississauga, Ontario, Canada). MilliQ® water was prepared with Millipore filtration system (Billerica, Massachusetts, USA) which was used to prepare solutions and growth media. For preparing antibiotic solutions and performing biochemical experiments, DNase-RNase-free water (catalog # 821739, referred in the text as 'nuclease-free water'), purchased from MP Biomedicals (Solon, Ohio, USA) was used. Restriction enzymes, Taq polymerases, modifying enzymes, and *in vitro* translation kits were purchased from New England Biolabs (NEB, Pickering, Ontario, Canada).

2.2. Bacterial strains

Chemically competent *E.coli* NEB 5-alpha cells (catalog # C2987H) which is a derivative of *E.coli* DH5α were purchased from NEB. Because of its high transformation efficiency of unmethylated DNA (*hsdR*), reduced recombination of cloned DNA (*recA1*) and nonspecific endonuclease I deficient (*endA1*) for highest quality plasmid preparations, this strain was selected for cloning and routine maintenance of plasmids.

Chemically competent *E.coli* BL21 (DE3) (catalog # C2527H) was also purchased from NEB. This strain has genome-encoded T7 RNA polymerase which is under the control of the *lacUV5* promoter. Isopropylthio-β-galactopyranoside (IPTG) which acts as a substrate analogue of lactose, was used to activate the *lacUV5* promoter to stimulate the production of T7 RNA polymerase, this in turn allows for the overexpression of the mRNA for the recombinant ORF present on an expression plasmid (which has the T7 promoter) thereby resulting in the desired

protein to be overexpressed. The strain is also *endA* deficient, therefore lacks the endonuclease responsible for the degradation of the plasmid DNA in miniprep methods. The chemical competent cells were stored in -80 °C freezer (ThermoFisher Scientific) until needed.

2.3. Resuspension of PCR primers and lyophilized plasmids

The primers used in this study (see Table 2.1) were ordered from Alpha DNA (Montreal, Quebec). Desalted, lyophilized oligonucleotides usually in the concentration range of 30-35 μg/OD were shipped from the manufacturer. The primers were adjusted to 400 picomoles by adding respective amounts of nuclease-free water, vortexed till mixed thoroughly, quick spun and stored at -20 °C (Frigidaire, Mississauga, Ontario, Canada) until required.

Plasmids containing the substrate sequences and plasmids with the suitable codon optimized ORFs encoding HEase used in this study were synthesized and cloned in plasmids by GenScript (Piscataway, New Jersey, USA). The lyophilized plasmids (4 μg) were suspended in 20 μL nuclease-free water, vortexed till mixed thoroughly, quick spun and stored at -20 °C until required.

2.4. DNA amplification

A TC-512 DNA thermal cycler from Techne (Burlington, New Jersey, USA) was used for DNA amplifications. Usually, the reaction mixture contained 1 x One *Taq*® standard reaction buffer (20 mM Tris-HCl pH 8.9, 22 mM NH_4Cl, 22 mM KCl, 1.8 mM $MgCl_2$, 0.06% IGEPAL® CA-630, 0.05% Tween® 20), 0.2 mM dNTPs, 1 μM of both forward and reverse primers (see Table 2.1.), 1-10 ng of plasmid DNA and 1.25 units of One *Taq*® Hot start DNA polymerase (NEB, catalog # M0481S). Nuclease-free water was used to adjust the volume to 50 μL. The

reaction conditions such as annealing temperatures, extension time of the primers etc. required for optimum amplification of different DNA fragments varied. The reaction conditions have been specified in the respective experimental chapters.

2.5. Transformation

The chemical transformation protocol was followed according to the manufacturer's recommendation (NEB). Briefly, 1-5 μL of 0.1 μg of the plasmid DNA (either substrate plasmid or HEase overexpression plasmid) was added to the tube containing chemical competent cells (see section 2.2 for the bacterial strains). The tube was then carefully flicked 4-5 times to mix the cells with the plasmid DNA and incubated for 30 minutes on ice. Heat shock was applied for exactly 30 seconds in a temperature controlled water-bath (Grants Instruments, Cambridge, Barrington, UK) pre-set at 42 °C. The tube was placed back on ice for another 5 minutes. A 950 μL of room temperature (RT) Super Optimal broth with Catabolite repression (SOC) medium was added aseptically to the mixture and placed in a 37 °C incubator room for an hour with vigorous shaking (250 rpm). The cells were thoroughly mixed and either 50 μL or 100 μL of the cells were aseptically spread-plated on pre-warmed LB-antibiotic selection plates. The plates were incubated overnight at 37 °C in an incubator room until the colonies were clearly visible.

2.6. Bacterial growth media

E. coli cultures were grown in the Luria Bertani (LB) liquid medium (10 g of Tryptone, 5 g of yeast extract, 5 g of NaCl per litre of culture). Likewise, solid LB plates were prepared that contained the same ingredients supplemented with 20 g/L of agar.

Ampicillin (catalog # BP1760-25) was added to the liquid and solid media to a

concentration of 60 μg/mL for selecting the bacterial cells containing the plasmids encoding an ampicillin resistance gene. For plasmids carrying the kanamycin resistance gene, kanamycin (catalog # BP906-5) was added to solid and liquid media to a concentration of 100 μg/mL. Chloramphenicol (catalog # BP904-100) was used at a concentration of 60 μg/mL for the plasmids encoding the chloramphenicol resistance gene. During cotransformation experiments, both kanamycin and chloramphenicol were used in the growth media at the aforementioned concentrations.

2.7. Plasmid miniprep

The bacterial cells harbouring the plasmid of interest were inoculated in 5 mL LB medium in the presence of an appropriate concentration of antibiotic and incubated overnight on a TC-7 rotatory incubator (New Brunswick Scientific Co., Inc., Connecticut, USA) located within a 37 °C incubator room. The plasmids were isolated from the *E.coli* cells using Presto™ Mini Plasmid Kit (Geneaid, New Taipei, Taiwan, catalog # PDH100). Briefly, 1 mL of cultured bacterial cells was transferred to a 1.5 mL microcentrifuge tube and centrifuged at 14000 x g for 1 minute at room temperature using a benchtop Spectrafuge 24D ultracentrifuge (Mandel, Guelph, Ontario). The supernatant was completely discarded by aspiration and the above step was repeated until the cell pellet from 4 mL of the cell suspension was harvested in the same 1.5 mL microcentrifuge tube. Two hundred μL of cell resuspension buffer: PD1 Buffer (containing 50 mg/mL RNase A) was added to the 1.5 mL microcentrifuge tube containing the cell pellet which was resuspended completely by pipetting until all traces of the cell pellet have been dissolved. Two hundred μL of cell lysis buffer: PD2 Buffer was added to the resuspended sample and then mixed gently by inverting the tube 10 times. The mixture was incubated at room

temperature for at least 2 minutes followed by addition of 300 µL of neutralizing buffer: PD3 Buffer which was mixed immediately. The sample was centrifuged at 14000 x g for 3 minutes at room temperature. The supernatant was carefully transferred to a PDH column (placed on a 2 mL collection tube) without disrupting the white precipitate and centrifuged at 14000 x g for 30 seconds at room temperature. The flow-through was discarded. Added 400 µL of W1 buffer into the PDH column and centrifuged at 14000 x g for 30 seconds. The flow-through was discarded and then placed the PDH column back in the 2 mL collection tube. Six hundred µL of wash buffer (pre-mixed with absolute ethanol) was added into the PDH column and centrifuged at 14000 x g for 30 seconds at room temperature. The flow-through was discarded. The PDH column back was again placed back in the 2 mL collection tube. The column matrix was dried (to remove any traces of ethanol) by centrifuging at 14000 x g for 3 minutes at room temperature. The dried PDH column was transferred to a new 1.5 mL microcentrifuge tube. Fifty µL of nuclease-free water was added into the center of the column matrix and let stand for at least 2 minutes to allow the nuclease-free water to be completely absorbed. Finally, the purified plasmid DNA was eluted by centrifuging at 14000 x g for 2 minutes at room temperature. The purity (Absorbance 260/280 ratio = 1.8 for purified DNA samples) as well as the concentration (in ng/µL) was determined using a Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). The plasmid was stored at -20 °C until needed.

2.8. Restriction digestion

The transformed colonies were selected based on the correct restriction pattern of the plasmids by performing digestion with various restriction enzymes. The restriction enzymes which were selected to undertake the digestion process were usually the ones previously used for

cloning the fragment into the plasmid backbone. Typically for colony screening, 0.5 µg of the plasmid DNA was digested following the manufacturer's protocol. The plasmid DNA was incubated with 1 unit of restriction enzyme in the presence of 1 x restriction buffer recommended for the optimal activity for that enzyme. Nuclease-free water was added accordingly to make the final volume of 20 µL. Sometimes, 1 hour incubation at 37 °C was not enough, therefore incubation for 2 hours to ensure complete linearization of the plasmid DNA was often carried out.

2.9. Agarose gel electrophoresis

A one percent (w/v) agarose gel was prepared by weighing a calculated amount of agarose powder (UltraPure™ Agarose, catalog # 15510-027) on a MXX-412 balance (Denver Instrument, Bohemia, New York, USA) followed by boiling in 1 x TBE buffer (89 mM Tris-HCl pH 8.3, 89 mM Boric acid, 100 mM EDTA) in a microwave till the agarose was completely dissolved. After cooling for 10 minutes, agarose gels were cast in either Bio-Rad Mini (6.5 cm x 10 cm) or Midi Sub Cell Plexiglass (10 cm x 20.5 cm) horizontal electrophoresis trays depending on the number of the samples to be run on the gel. As the reference for all the DNA electrophoresis experiments, 1 kb plus™ DNA Ladder (catalog # 10787-018) was used. Electrophoresis was performed in either minicell® Primo™ EC320 or midicell® Primo ™ EC330 (Holbrook, New York, USA) connected to a EC 250-90 power pack (E-C Apparatus Corporation, Markham, Ontario, Canada). A constant 100 volts in 1 x TBE buffer was applied to run the gel. The electrophoresis was stopped until the bromophenol blue dye marker front had migrated approximately two-thirds the length of the gel. DNA bands were visualized by staining with 0.5 µg/mL ethidium bromide (catalog # 15585011) and exposure to ultraviolet light within

gel documentation system (FluorChem™, Alpha Innotech) and the gel images were saved in TIFF format for future reference.

2.10. Agarose gel purification

PCR products were purified using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, USA, catalog # A9281). In the case of purifying the PCR product from an agarose gel, the gel portion containing the band of interest was sliced using a sterilized scalpel and transferred into a microcentrifuge tube. The solution containing the PCR product was mixed with an equal volume of binding solution. Ten μL membrane binding solution per 10 mg of the gel slice was added and vortexed. The tube was incubated at 50-65 °C water bath until the gel slice was completely dissolved. The membrane binding solution containing either the PCR product or the dissolved gel mixture was transferred onto a SV mini-column and incubated for 1 minute at room temperature. The tube was centrifuged at 16000 x g for 1 minute. The mini-column was washed with 700 μL of membrane wash solution and then centrifuged as above. The sample bound within the mini-column was washed a second time with 500 μL of membrane wash solution and centrifuged as above for 5 minutes. To remove any residual ethanol, the column matrix was dried by centrifuging at 16000 x g for 1 minute at room temperature. The purified PCR product was eluted from the mini-column in 50 μL of nuclease-free water and stored at -20 ℃ for DNA sequencing and/or other downstream applications.

2.11. Storage media for recombinant bacteria

Bacterial stock cultures containing the desired plasmids were stored in 50% pre-sterilized glycerol at -80 °C for future experiments. Typically, aliquots containing 0.5 mL bacterial culture

and 0.5 mL of 50% glycerol were mixed aseptically and vortexed until homogeneous. The respective plasmids, however, were also stored as DNA preparations in 50 µL of nuclease-free water and kept at -20 °C.

2.12. Recombinant protein expression

Small scale cultures of transformed and untransformed strains (negative control) were grown at 37 °C in fluted conical flasks (Pyrex) containing 50 mL LB medium to OD_{600} ~ 0.6 at which time IPTG (catalog # 15529019) at either high (0.5-1 mM) or low (0.1-0.3 mM) concentrations were added to induce recombinant protein synthesis. As part of the screening process in order to identify optimum conditions, the induced cultures were then grown at RT (~22.5 °C), 28 °C, and 37 °C for at least 4 hours to a maximum of 16 hours (overnight). For large scale protein expression, four large fluted conical flasks (Pyrex) containing 500 mL of protein expression media (usually LB) were prepared and the recombinant protein was expressed based on the pre-determined conditions gathered from the small scale expression.

The PURExpress *In Vitro* Protein Synthesis Kit (NEB, catalog # E6800S) which is a cell-free transcription/translation system was utilized in order to check the *in vitro* expression of the ORFs encoding HEases. According to the manufacturer's protocol, 7.5 µL of 'solution B' was thoroughly mixed with 10 µL of 'solution A' followed by the addition of 1-5 µg purified RNA containing a proper ribosome binding site (RBS) for efficient translation. One unit of murine RNase inhibitor (NEB, catalog # M0314) was also included in the reaction mixture to remove any unwanted activity of RNase. The reaction mixture was supplemented with nuclease-free water to make a volume up to 25 µL. After 3-4 hours of incubation at 37 °C, 2.5 µL of the reaction sample was mixed with 2.5 µL of the 2 x protein loading dye (65.8 mM Tris-HCl, pH

6.8, 26.3% (w/v) glycerol, 2.1% SDS, 0.01% Bromophenol blue) and subjected to 12.5% SDS-PAGE (refer section 2.14).

2.13. Extraction and purification of recombinant protein

The large scale protein expression culture (see section 2.12) was poured in a 250 mL style-3120 Nalgene centrifugal bottle (Sigma-Aldrich, Oakville, Ontario, Canada) and centrifuged at 7000 x g for 10 minutes using a SLA-1500 rotor (radius size: 13.59 cm) and Sorvall RC5B PLUS superspeed centrifuge machine (GMI Inc., Ramsey, Minnesota). The cell pellet was harvested and stored at -80 °C overnight. The frozen pellet was thawed on ice and resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 % (w/v) glycerol, 6 mM β-mercaptoethanol) at a ratio of 5 ml of buffer to 1 g of cells (wet weight). Cells were homogenized twice using the French press at 1200 psi and the resulting lysate was centrifuged using a SS-34 rotor (radius size: 10.7 cm) at 17000 x g for 20 minutes at 4 °C to pellet cell debris. The centrifugation step was repeated until a clear cell lysate was achieved. Thereafter, the lysate (about 8 ml) was added to 3 ml of Ni-NTA resin (Qiagen, Toronto, Ontario, Canada) and incubated at 4 °C with agitation for 20 minutes. The sample along with the Ni-NTA slurry was loaded onto a Ni-NTA super flow column (Qiagen) and the washing steps were as listed: wash 1: 30 ml of washing buffer (WB) (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 % (w/v) glycerol, 6 mM β-mercaptoethanol); wash 2: 30 ml of WB buffer with 25 mM of imidazole; and wash 3: 30 ml of WB buffer with 50 mM of imidazole. The protein was eluted in Elution buffer (EB i.e. WB supplemented with 250 mM imidazole) and collected in 10 fractions (1 mL each). After confirming the presence of purified protein among these 10 fractions by performing denaturing SDS polyacrylamide gel (12.5%) electrophoresis (see section 2.14.), the fractions showing the

desired protein were pooled. The slide-a-lyzer dialysis cassettes (Millipore) of desired molecular weight cut-off (MWCO) were used to remove the excess imidazole. The protein sample (3 mL) was injected into the dialysis cassette and usually suspended in 2 L of dialysis buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 6 mM β-mercaptoethanol) according to the manufacturer's protocol. The dialysed sample was concentrated by the Amicon concentrator (model 8050) using an YM-10 membrane (Millipore) to a final volume of 1ml. Usually, the purified, concentrated protein (HEase) was immediately assayed for its endonuclease activity before storing. However, for long term storage, the pure protein was mixed with 200 µL of protein storage buffer (50 mM Tris-HCl pH 8.0, 400 mM NaCl, 0.5 mM DTT, 10% (w/v) glycerol) and kept in a -80 °C freezer.

2.14. Resolving proteins on SDS-PAGE

A discontinuous sodium dodecyl sulfate (SDS) polyacrylamide gel (PAGE) consisting of a 5% stacking gel and a 12.5% resolving gel was cast in vertical slabs of dimensions 10 x 10 cm and 0.5 mm thickness. All the following ingredients (except Tris-HCl and SDS) and protein gel electrophoresis system were purchased from BioRad (Mississauga, Ontario, Canada). For preparing 3 mL of 5% stacking gel, which is sufficient for casting two gels of the above dimensions, the following ingredients (in mL) were added in a particular order as mentioned - H_2O: 2.062, 30% Acrylamide/Bis solution (catalog #161-0158): 0.5, 1 M Tris-HCl (pH 6.8): 0.375, 10% SDS: 0.03, 10% Ammonium persulfate (APS, catalog # 161-0700): 0.03, TEMED (catalog # 161-0800): 0.003. Similarly, the following ingredients (in mL) were added in a particular order to prepare 10 mL of 12.5% resolving gel- H_2O: 3.296, 30% Acrylamide/Bis solution: 4.0, 1.5 M Tris-HCl (pH 8.8): 2.5, 10% SDS: 0.1, 10% APS: 0.1, TEMED: 0.004. The gel was cast and set up according to the manufacturer's instructions. A vertical mini-Protean II

electrophoresis system connected to a FB 154 power pack (ThermoFisher Scientific) was used to resolve the denatured protein samples. Typically, the samples loaded onto the gel contained 8-12 μg of protein and were mixed with equal volumes of 2 x protein gel-loading buffer and boiled for 10 minutes before loading. BLUeye Prestained Protein ladder (catalog # PM007-0500K) was purchased from Froggabio (North York, Ontario, Canada) which was used in all SDS-PAGE analysis of protein fractions for approximate weight determination. A constant 150 volts was used to run the gels in 1 x running buffer (0.025 M glycine, 0.192 M Tris base, and 0.1% SDS, pH ~8.3) until the dye reached the bottom of the glass plates. The gels were retrieved from the plates and stained with a gel staining solution containing 0.5 g/L Coomassie Brilliant Blue R-250 (dissolved in 30% ethanol, 10% acetic acid) and destained with repeated changes of destaining solution (15% methanol, 7% acetic acid) until the background was mostly clear. Both staining and destaining were performed on a Mistral multimixer shaking platform from Lab-line Instruments (Dubuque, Iowa, USA). Another destaining solution containing 7% acetic acid and 1% glycerol was used to soak the gels for 1 hour before mounting between two pieces of cellophane membrane. A plastic frame firmly secured the cellophane sheets containing the gel which was dried overnight at room temperature.

2.15. *In vitro* endonuclease assay

In order to check the activity of HEases, *in vitro* endonuclease assay reactions were performed. The reaction mixture contained: 25 μg/mL substrate or the negative control plasmid or the substrate PCR product, 5 μL Reaction Buffer #3 (100 mM NaCl, 50 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$) supplemented with 1 mM DTT (catalog # D1532), 9-10 μg purified HEase and H_2O to achieve a final volume of 50 μL. Cleavage reactions were incubated at 37 °C and 10 μL

aliquots were withdrawn at 0, 30, 60, 90 and 120 minutes. These aliquots were treated with the addition of 2 µL of 200 mM EDTA (pH 8.0) and 1 µL of proteinase K (1 mg/mL, catalog # BP1700-100) in order to stop the endonuclease reactions. The aliquots were further incubated for 30 minutes at 37 °C. The products of these assays were resolved on a 1% agarose gel.

2.16. Cleavage site mapping assay

The protocol for cleavage site mapping is according to Bae *et al*. (2009). Briefly, 40 µL of linearized plasmid (25 µg/mL) was treated with 10 units of T4 DNA polymerase (5u/µL) that included 20 µL of 5 x T4 DNA polymerase buffer, 20 µL dNTP mixture (0.5 mM) and H_2O to achieve the final volume of 100 µL. The reaction was incubated at room temperature (~ 22.5 °C) for 20 minutes and then placed on ice for 5 minutes. The reaction was terminated by heating the mixture at 70 °C for 10 minutes. Enzymes possessing the LAGLIDADG motif cleave substrate DNA in a manner that generates 4-nt, 3′-OH overhangs. The reason for treating the linearized substrate plasmid with T4 DNA polymerase (catalog # EP0061) is due to the enzyme's ability to remove exonucleolytically 3′-OH overhangs, thus generating blunt DNA ends. The DNA was recovered from the reaction mixture with the Wizard® SV Gel and PCR Clean-Up system (see section 2.10). The linearized blunt-ended plasmid DNA was then religated in 50 µL ligation mixture containing 20 µL (0.25 µg) of the T4 DNA polymerase treated cleaved plasmid, 2 µL of T4 DNA ligase (1u/µL, catalog # 15224017), 10 µL of 5 x ligase buffer and H_2O to achieve the final volume of 50 µL. The reaction was incubated at room temperature for 2 hours and thereafter the ligation mixture was diluted 5-fold and 10 µL of this dilution was used to transform *E. coli* NEB 5-alpha cells. Potential transformants were platted on LB agar plates supplemented with 100 µg/mL ampicillin, incubated at 37 °C until the colonies were visible.

Single colonies were inoculated in 5 mL LB media supplemented with appropriate amount of ampicillin and grown overnight. Plasmids were recovered from the overnight cultures with the Presto^{TM} Mini Plasmid Kit (see section 2.7) and the plasmid DNAs were sent to either the NRC DNA Technologies Unit (Saskatoon, Saskatchewan, Canada) or MICB DNA sequencing facility (University of Manitoba, Cancer Care, McDermot Avenue, Winnipeg, Canada) for Sanger cycle sequencing using the M13F forward and M13R reverse primers (see Table 2.1) for both treated and untreated substrates. The chromatograms from these sequencing reactions were aligned manually and compared with the GeneDoc program (version 2.7; Nicholas *et al*., 1997); in particular the HEase and T4 DNA polymerase treated substrate plasmid sequence with the untreated substrate plasmid sequences. Nucleotides missing in the sequence of the HEase treated plasmid when compared to the original untreated substrate plasmid derived sequence allowed for the determination of the HEase cleavage site.

2.17. cDNA synthesis

The ThermoScript RT-PCR kit (catalog # K044) was used to make cDNA from the RNA extracted from bacterial cells. For the first strand synthesis, a 20 µL reaction mix was prepared containing 1 µg RNA, 0.5 µM of the reverse primer, 0.1 M DTT, 4 µL of 5 x cDNA synthesis buffer, 1 mM of each dNTP, 40 units RNaseOUT and 15 units of Thermoscript reverse transcriptase. Reverse transcription was performed at 55 °C for 1 hour and stopped by heating the reaction mixture to 85 °C for 10 minutes. Finally, in order to remove any RNA contaminants, 1 µL of RNase H (2 units) was added to the reaction mixture followed by incubation at 37 °C for 20 minutes.

Table 2.1. Primer list

Forward primers:

Primer name	Primer sequence	Comment
F1-T7	5′-*GAATTC*<u>*TAATACGACTCACTATAGGGA*</u> *ACTATCAAACTCCGGGG*-3′	T7 promoter sequence underlined; Chapter 3
Primer ‘A’	5′-*TAGAGGACTATGCATGTCC*-3′	Chapter 3
Primer ‘C’	5′-*ACAGCATGCAGCAAAAGCGG*-3′	Chapter 3/4
M13F	5′-*GTAAAACGACGGCCAG*-3′	Chapter 3/4
rns-F1	5′-*CGTGCCAGCAGTCGCGG*-3′	Chapter 3
TwinHEG-F	5′-*ATGTGGTTATCCCGCATTTCG*-3′	Chapter 4
TwinORFnew-F	5′-*ATAAGAACGATCTGGAAGTCCTGC*-3′	Chapter 4
cytbi3pMAL-F	5′-*ATTAATATCACTATTCTGGTTAGCGG*-3′	Chapter 5
malE-F	5′-*GGTCGTCAGACTGTCGATGAAGCC*-3′	Chapter 5

Table 2.1. Primer list (continued)

Reverse primers:

Primer name	Primer sequence	Comment
mtsr-2	5′-*CGAGTGGTTAGTACCAATCC*-3′	Chapter 3
Primer ‘B’	5′-*TTCCTCAGTAAGATGGCC*-3′	Chapter 3
M13R	5′-*CAGGAAACAGCTATGAC*-3′	Chapter 3/4
TwinHEG-R	5′-*TTGAAGTTTTCGTTCTTGATGCC*-3′	Chapter 4
Primer ‘D’	5′-*TGTATAACATCTCAGCCGACTGCC*-3′	Chapter 3/4
cytbi3pMAL-R	5′-*TATCTCTGGATCCCATTACAGTTTCGGGTAGC ACAG*-3′	BamHI sequence underlined; Chapter 5
malE-R	5′-*TGTCCTACTCAGGAGAGCGTTCAC*-3′	Chapter 5

Chapter 3

Biochemical characterization of a twintron (nested intron) encoded homing endonuclease

3.0. Abstract

The small ribosomal subunit gene residing in the mitochondrial DNA of the thermophilic fungus *Chaetomium thermophilum* var. *thermophilum* La Touche DSM 1495 is interrupted by a twintron at position mS1247. The mS1247 twintron represents the first mixed twintron found in fungal mtDNA, composed of an external group I intron encoding a LAGLIDADG open reading frame that is interrupted by an internal group II intron. Splicing of the internal group II intron reconstitutes the open reading frame and thus facilitates the expression of the encoded homing endonuclease. The cleavage assays suggest that the twintron (nested intron) encodes an active homing endonuclease that could potentially mobilize the twintron to *rns* genes that have not yet been invaded by this mobile composite element.

The work presented in this chapter has been published.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68.

Conceived and designed the experiments: TKG, GH. Performed the experiments: TKG. Analyzed the data: TKG, GH. Contributed reagents/materials/analysis tools: GH. Wrote the paper: TKG, GH.

3.1. Introduction

Homing endonucleases (HEases) are encoded by homing endonuclease genes (HEGs) which can be embedded within composite mobile elements such as group I introns and group II introns, archaeal introns, as well as inteins (Dujon, 1989; Belcour *et al*., 1997; Gimble, 2000; Tocchini-Valentini *et al*., 2011). HEGs can be components of mobile elements that mimic introns but can in many cases with the help of host factors splice from the primary transcripts or in the case of inteins from the protein precursor (Belfort, 2003; Lang *et al*., 2007). HEases are named based on conserved amino acid motifs and the LAGLIDADG and GIY-YIG families of HEases are most frequently encountered among fungal mitochondrial group I introns (Haugen *et al*., 2006; Stoddard, 2006; Hausner, 2012). Enzymes possessing the LAGLIDADG motif cleave substrate DNA in a manner that generates 4-nucleotide, 3′-OH overhangs and their DNA recognition sequences are generally asymmetrical and long (12-40 bp; Belfort and Roberts, 1997). HEases promote the mobility of the HEG or the mobility of composite elements that encodes HEases by cleaving a target sequence in cognate alleles that lack HEGs or intron/intein insertions (Dujon and Belcour, 1989).

Amlacher *et al*. (2011) determined the nuclear and mitochondrial genomes (mtDNA) for the thermophilic fungus *Chaetomium thermophilum var*. *thermophilum* La Touche (strain DSM 1495; Microfungus Collection and Herbarium; http://straininfo.net/strains/418186). Although the optimal growth is around 45 °C, this fungus can tolerate temperatures up to 60 °C although being eukaryotic (http://eol.org/pages/1016560/details). This thermophilic fungus has been isolated from soil and compost heaps and has been reported from many locations including British Isles, Himachal Pradesh (India), Egypt, Zambia (Wang *et al*., 2012; Busk and Lange, 2013). Besides its ecological importance as a cellulolytic fungus, it is viewed as a source of thermostable

enzymes which are subjected to better structural and biochemical studies than comparable mesophilic fungi (Amlacher *et al*., 2011). For example, as lower temperature precipitates yeast proteins, studying the nuclear pore complex proteins in *C. thermophilum* proved advantageous since the isolation of proteins was more abundant, more soluble and thermostable than the yeast proteins (Amlacher *et al*., 2011).

Comparative sequence analysis of the mitochondrial small ribosomal subunit (*rns*) gene among species of Ascomycota and related taxa indicated that the *rns* gene appears to be a reservoir for mobile introns (Hafez *et al*., 2013). Interestingly, the *rns* gene has also been shown to harbour atypical complex intronic arrangements like the ones commonly described as twintrons (Copertino and Hallick, 1991; Hafez *et al*., 2013).

A twintron (nested intron) is basically an intron-within-intron arrangement that is excised by sequential splicing reactions where the internal intron is removed prior to the excision of the external intron. Since the discovery of this composite arrangement (group II intron embedded in another group II intron) within the *psbF* locus of *Euglena gracilis* chloroplast DNA (Copertino and Hallick, 1991), several categories of twintrons have been characterized. A twintron can have a simple architecture where an external intron is interrupted by one internal intron of either the same (Copertino and Hallick, 1991) or different intron group (Copertino *et al*., 1998; Hafez *et al*., 2013). A complex arrangement where an external intron interrupted by multiple internal introns of same (Drager and Hallick, 1993) or different intron groups has also been observed (Suzuki *et al*., 2013). Even though a majority of these twintrons have been characterized within the *Euglena* chloroplast genome (Copertino and Hallick, 1991; Thompson *et al*., 1997), these elements have also been found in *Pyrenomonas salina* (cryptomonad algae, Maier *et al*., 1995), *Didymium iridis* (Einvik *et al*., 1998a) and *Drosophila* (Scamborova *et al*., 2004). Recently, a

novel twintron has been uncovered within the *rns* gene of the fungal mitochondrial genome at position mS917 of the *Cryphonectria parasitica*, where a group ID intron encoding a LAGLIDADG ORF invaded another ORF-less group ID intron (Hafez *et al*., 2013).

Examination of *C. thermophilum* mtDNA sequence showed that the *rns* gene contained another novel twintron at position mS1247 (Hafez *et al*., 2013). Characterization of the *rns* twintron in two strains of *C. thermophilum* var. *thermophilum* (strains DSM 1495 and UAMH 2024/CBS 141.64) indicated that this twintron is composed of an external group I intron that has been invaded by a group II intron (Figure 3.1). In particular, the internal group II intron inserted within the external group I intron LAGLIDADG open reading frame (ORF). This posits a unique possibility whereby splicing of the internal group II intron would allow the ORF to be reconstituted and thus allowing for the expression of the encoded HEase. This might offer a window for engineering a HEase that based on the splicing competency of an intron might have an *in vivo* switch for regulating HEase activity/expression. As HEases require long DNA recognition sites they cut infrequently within a genome and therefore are useful for DNA engineering (Stoddard, 2011; Prieto *et al*., 2012). The value of native HEases towards engineering site specific genome editing tools has been demonstrated by Takeuchi *et al*. (2011) on work based on HEGs inserted within the mtDNA *rps3* gene in *Ophiostoma novo-ulmi* subspecies *americana* and *Letographium truncatum* (Gibb and Hausner, 2005; Sethuraman *et al*., 2009).

In this study, we examined the S1247 twintron encoded putative HEase (I-CthI) from *C. thermophilum*, mapped its cleavage site within the *rns* gene and provided some characterization of the HEase protein and demonstrated by an *in vitro* splicing assay that the group II intron

indeed splices in a manner that allows for the interrupted HEase ORF to be reconstituted during RNA processing.

3.2. Materials and Methods

The methods exclusively related to this chapter have been detailed in this section. For common materials and methods used in this chapter (appropriately mentioned in the text), the readers are directed to Chapter 2 (General Materials and Methods).

3.2.1. *In vitro* RNA splicing assay

To demonstrate the splicing competency of the group II intron component of the S1247 twintron (nested intron), a splicing assay involving *in vitro* transcription was performed following the protocol of Salman *et al*. (2012). A segment of the *rns* gene containing the S1247 twintron (nested intron) was amplified by PCR using the F1-T7 forward primer containing a T7-promoter sequence (see Table 2.1, T7 promoter sequence underlined) and the mtsr-2 reverse primer (see Table 2.1). The PCR reaction was performed with PCR reagents from ThermoFisher Scientific according to the manufacturer's recommendations. The reaction conditions were as follows: initial denaturing for 3 minutes at 93 °C, followed by 25 cycles of 93 °C for 1 minute, 55 °C for 1 minute and 72 °C for 2 minutes. The PCR reactions were evaluated by submarine agarose gel electrophoresis on a 1% gel as described in section 2.9. The F1-T7/mtsr-2 PCR product was used as the template for the *in vitro* transcription reaction (also see Figure 3.2A for schematic representation of the protocol).

A 20 µL *in vitro* transcription reaction mixture contained the following ingredients: 0.1 µg DNA template (F1-T7/mtsr-2 PCR product), 0.5 mM of each NTP, 2 µL of T7 transcription buffer (20 mM NaCl, 40 mM Tris-HCl pH 7.8, 6 mM $MgCl_2$, 2 mM spermidine, 10 mM DTT) and 20 units of T7 RNA Polymerase-Plus Enzyme Mix (Catalog # AM2716). The reaction mixture was incubated for 2 hours at 37 °C. For splicing, the addition of NaCl and $MgCl_2$ to a

final concentration of 1.2 M NaCl and 60 mM respectively to the transcription reaction buffer was necessary in order for efficient splicing of the group II intron. However, initial experiments on *in vitro* splicing competency of internal group II intron were carried out without addition of salts (NaCl and $MgCl_2$) in the transcription reaction buffer. Template DNA was removed by the addition of 2 units of DNaseI (catalog #AM2222) and incubated at 37 °C for 15 minutes; the reaction was stopped by adding 1 µL EDTA (50 mM) followed by a 10 minutes incubation at 65 °C. In order to confirm the elimination of the DNA template, the forward primer 'A' and the reverse primer 'B' (see Table 2.1) were applied to perform a standard PCR reaction as described above using 2 µL from the *in vitro* transcription reaction mixture as the template (also see Figure 3.2B).

To evaluate if splicing occurred, cDNA was generated from the *in vitro* transcribed RNA by reverse transcriptase (RT) PCR as described in section 2.17. For the synthesis of the first strand of cDNA, 0.5 µM of the reverse primer (mtsr-2) was used in this reaction. In order to characterize the twintron derived transcripts, several primers were designed (also see Figure 3.2A) to recover potential splicing intermediates. The forward primer 'A' and the reverse primer 'B' were based on the external intron's upstream and downstream sequences with regards to the group II intron location. Primer 'C' and primer 'D' (see Table 2.1) were designed to amplify a 400 bp segment of the internal group II intron and these primers were designed to provide a positive control. To amplify potential splicing intermediates, the protocol recommended in the One *Taq*® Hot start DNA polymerase kit (NEB) was followed. A 2 µL of cDNA (0.2 µg) was added to a 50 µL PCR reaction mixture containing1 µM of each forward primer 'A' and reverse primer 'B'. The PCR reaction conditions were as follows: initial denaturation at 94 °C for 2 minutes, followed by 30 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds, 68 °C for 2

minutes, followed by final extension time of 10 minutes at 68 °C. PCR products were resolved on a 1% agarose gel as described in section 2.9 (also see Figure 3.2C and 3.2D). The amplicons obtained were excised from the agarose gel and purified using the Wizard® SV Gel and PCR Clean-Up system as described in section 2.10. The gel extracted DNA fragments were sent to the NCR DNA Technologies Unit (Saskatoon, Saskatchewan, Canada) for cycle sequencing, utilizing the primers used for obtaining the amplicons (also see Figure 3.2A).

3.2.2. Construction of *E. coli* expression vector for the I-CthI HEase

Based on sequences obtained by Amlacher *et al*. (2011) and Hafez *et al*. (2013) (Genbank accessions: JN007486 and JX139037 respectively) the mS1247 twintron (nested intron) ORF was reconstructed by removing the group II intron sequence from the external intron's ORF sequence. The genetic code for the HEase ORF was optimized for expression in *E. coli* and the HEase sequence was synthesized by Genscript. The synthesized ORF was inserted into the pET28 b (+) plasmid as a BamHI/NdeI fragment with an N-terminal 6 x Histidine-tag; this construct was named I-CthI-pET28 b (+) (Figure 3.3A). The construct was transformed into *E. coli* BL21 (DE3) as described in section 2.5 for protein expression, purification and biochemical studies.

Ten mL of LB media supplemented with 100 µg/mL kanamycin and 0.25% (w/v) glucose was inoculated with 100 µL of *E. coli* BL21 (DE3) transformed with I-CthI-pET28 b (+) and incubated overnight with agitation at 37 °C. Ten mL from the overnight culture was used to inoculate 1 L of LB medium supplemented with appropriate concentration of kanamycin and 25% (w/v) glucose as described above. The general method for protein overexpression has been described in section 2.12. The overexpression conditions for this protein has been indicated here.

The culture was grown at 37 °C with agitation (rpm 210) till the OD_{600} reached ~ 0.65 and expression of the protein was induced with 0.5 mM IPTG. The culture was then shifted to 28 °C and incubated overnight (~16 hours) with agitation (rpm 210). The recombinant protein was extracted, purified, resolved on denaturing gel and dialysed as described in section 2.13 (also see Figure 3.3B and 3.3C).

3.2.3. Endonuclease assay

The putative HEase was evaluated for activity by performing endonuclease assays. A substrate plasmid containing the HEase recognition site was designed as follows. A segment of the *C. thermophilum rns* sequence (JN007486) was constructed that is composed of sequences flanking the mS1247 twintron (321 nucleotide upstream and 148 nucleotide downstream of the twintron) based on predictions of Hafez *et al*. (2013). This 469 bp sequence was synthesized by GenScript and inserted into the EcoRV site within the pUC57 vector (2.7 kb). The substrate plasmid was named Cth-rns.pUC57 and its size is 3.1 kb (see Figure 3.4A).

To serve as the negative control, a *C. thermophilum rns* segment that contains the twintron plus the flanking exon sequences was PCR amplified as described in section 2.4 using the rns-F1 forward and mtsr-2 reverse primers (see Table 2.1; Hafez *et al*., 2013). The resulting PCR product (2.7 kb) was resolved on a 1% agarose gel and extracted from the gel using the Wizard® SV Gel and PCR Clean-Up system as described in 2.10. The purified 2.7 kb DNA fragment was cloned into the pCR4 TOPO vector using the TOPO TA Cloning Kit ingredients and protocols recommended by the manufacturer (catalog # K4575-01SC).

Finally the HEase was challenged with a PCR product that contained the putative I-CthI target site. This template was generated to rule out the possibility that the putative HEase cuts in

the pUC57 vector sequence, the DNA sequence containing only the substrate region was PCR amplified from the Cth-rns.pUC57 (i.e. substrate plasmid) using M13F forward and M13R reverse primers (see Table 2.1) using the following conditions: initial denaturation 94 °C for 1 minute, followed by 30 cycles of 94 °C for 30 seconds, 52 °C for 30 seconds, 68 °C for 30 seconds and a final elongation time of 5 minutes at 68 °C. The substrate and control plasmids were transformed into *E. coli* DH5α and the plasmids were purified with the Wizard ®Plus Minipreps DNA purification kit as described in section 2.10. *In vitro* endonuclease assay was performed as described in section 2.15 (also see Figure 3.4B and 3.4C).

3.2.4. I-CthI cleavage site mapping

The cleavage mapping assay was performed as described in section 2.16 (also see Figure 3.5).

3.2.5. Temperature profile and thermal stability of the I-CthI protein

To test the effect of temperature on I-CthI cleavage activity, the endonuclease assay was performed as described in section 2.15, but the reactions were incubated at a temperate range from 25 °C to 85 °C at 10 degree intervals for 1 hour. The reactions were terminated as described previously and the products of these assays were resolved on a 1% agarose gel (see Figure 3.6A). Temperature stability was also examined by circular dichroism spectropolarimetry. The CD spectra were acquired on a JASCO J-810 spectropolarimeter-fluorometer calibrated with (+)-10-camphorsulfonicacid and purged with nitrogen (N_2) at 20 L/minute. CD spectra of the protein samples were measured in the far UV region (180-250 nm) using 0.05-0.10 cm path length quartz cuvettes at the initial setting at 35 °C, a scan rate of 10 nm/minute, and a response time of

8 seconds. CD spectra were corrected by baseline subtraction and were converted to mean residue ellipticity (MRE) according to the following formula: $[\Theta]_M = M\,\Theta/10lcn$ where the units for $[\Theta]_M$ are 10^3 deg cm^2/dmol, M is the molecular mass of His_6-HEase (34.8 kDa or 1.2 mg/mL), Θ is the measured ellipticity in millidegrees, l is the path length of the cuvette in cm (0.1 cm), c is the protein concentration in g/L and n is the number of amino acid residues in the protein (305). Temperature was controlled during thermal denaturation experiments of the His_6-HEase sample using the Peltier device connected to the spectropolarimeter. Spectra were recorded for the following temperate range: from 25 °C to 85 °C at 10 degree intervals for 30 minutes at each tested temperature (see Figure 3.6B).

The structure of the twintron encoded HEase was predicted with the online Protein Homology/analogY Recognition Engine V 2.0 (PHYRE2) program (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index; Kelley and Sternberg, 2009; also see Figure 3.6C).

3.2.6. Co-crystallization trials of I-CthI bound to its cognate target site

The co-crystallization trials with the purified I-CthI HEase along with its linearized substrate DNA (target site) was undertaken by our collaborative partner, Dr. Barry Stoddard's group in the Fred Hutchinson Cancer Research Centre in Seattle, Washington, USA. Based on their progress report with somewhat limited details, large scale protein overexpression and purification using the heparin column followed by gel filtration were conducted. The details on the use of overexpression vector, media, buffers for the protein purification and the ratio of protein: DNA concentration in the cocrystallization mixture were not provided but essentially followed the methods described in Takeuchi *et al.* (2011). In order to perform co-crystallization

studies, several crystal trays were set up (hanging drop vapour diffusion method) where the purified protein, I-CthI (~10 mg/mL) was incubated with the substrate DNA target site (23-25 bp; Takeuchi *et al*., 2011). The crystal trays contained polyethylene glycol (PEG) Suite (1- 48) with either divalent cations (25 mM $CaCl_2$, 500 mM $CaCl_2$) or without divalent cations. In order to reduce disulfide bonds between cysteine residues in the protein structure, 10 mM DTT was added to the purified protein and flash froze prior to setting up the crystal trays. In addition, the Wizard Classic crystallographic screens (Rigaku, Washington, USA) were used.

3.2.7. Phylogenetic analysis of the twintron encoded ORF and related LAGLIDADG HEases

The online resource BLASTp (Altschul *et al*., 1997) was used to retrieve sequences that were related to the mS1247 twintron ORF. The LAGLIDADG ORF amino-acid sequences were aligned with the online PRALINE multiple sequence alignment program (Simossis and Heringa, 2005) and the alignment was further refined with the GeneDoc program (Nicholas *et al*., 1997).

For phylogenetic analyses, only those segments of the alignment where all sequences could be aligned unambiguously were retained. Phylogenetic estimates were generated by the programs contained within the Molecular Evolutionary Genetic Analysis program package (MEGA 5.2; Tamura *et al*., 2011) and the MrBayes program v3.1 (Ronquist and Huelsenbeck, 2003). With MEGA phylogenetic trees were generated with the Maximum parsimony (MP), Neighbor joining (NJ), and Maximum likelihood (ML) methods. MEGA 5.2 was also used for determining the best fit substitution model for ML analysis; thus for ML analysis the WAG +G+F model was applied and for all programs the bootstrap option was selected (1000 replicates) in order to obtain estimates for the confidence levels for the major nodes present

within the phylogenetic trees (Felsenstein, 1985).

The MrBayes program was used for Bayesian analysis and the parameters for amino acid alignments were the mixed model setting. The Bayesian inference of phylogenies was initiated from a random starting tree and four chains were run simultaneously for 2000000 generations; trees were sampled every 100th generations. The first 40% of trees generated were discarded ("burn-in") and the remaining trees were used to compute the posterior probability values and majority rule consensus tree. Phylogenetic trees were drawn with the TreeView program (Page, 1996; also see Figure 3.7) using the MrBayes tree files, and the phylogenetic tree was annotated with Corel Draw^{TM} (Corel Corporation Ltd., Ottawa, Canada).

3.3. Results

3.3.1. *In vitro* splicing of the internal group II intron reconstitutes the LAGLIDADG ORF

Initial attempts using RT-PCR with whole cell RNA as a template to recover precursor *rns* transcripts failed and we could only recover mature *rns* transcripts (as previously characterized in Hafez *et al*. (2013) (Figure 3.1). In order to demonstrate that the internal group II intron could splice, we set up *in vitro* transcription assays based on a segment of the *rns* gene that contained the twintron region (Figure 3.2A). These assays allowed for the generation of splicing competent group II intron RNAs, however only under high salt concentrations (1.2 M NaCl and 60 mM $MgCl_2$), we could recover splicing intermediates along with processed RNAs (Figure 3.2B, C and D). The splicing reaction products were analyzed by RT-PCR utilizing primers 'A' and 'B' (based on external intron sequences that flank the internal group II intron). Among the observed cDNAs was a dominant band at 1.1 kb (Figure 3.2D), the expected size for cDNAs from transcripts where the group II was spliced out. Along with the 1.1 kb fragment we observed what appears to be cDNAs generated from unspliced versions (1.9 kb) and various shorter fragments. Those were not further investigated. As a positive control cDNAs were generated based on the internal group II with primers 'C' and 'D' generating a 400 bp fragment (Figure 3.2C and D); serving as a negative control was a sample without the RT step to ensure all DNA was removed from the *in vitro* transcription assay (Figure 3.2B). The 1.1 kb cDNA PCR product was excised from the agarose gel and submitted for DNA sequence analysis and the resulting data were compared with the genomic version of the mS1247 twintron. Comparative sequence analysis within the GeneDoc program showed that the 1.1 kb cDNA was the result of the group II intron being spliced out and the joining of the flanking external intron sequences. Also as predicted previously by *in silico* analysis the "internal intron"/ external intron splice

junction (Hafez *et al*., 2013) corresponds to a phase 0 position with regards to the coding region of the group I intron ORF and the splicing of the group II intron allows the group I ORF located within the P 9.1 loop to be reconstituted into a continuous reading frame potentially encoding a functional HEase.

3.3.2. Overexpression and purification of the twintron encoded homing endonuclease

The codon optimized HEase ORF was overexpressed in *E. coli*, however optimization with regards to IPTG concentration and temperature during induction was required. The transformed *E. coli* culture was grown overnight and the recombinant HEase protein expressed at all temperatures tested (28 °C and 37 °C) including RT (~22 °C) except at 16 °C. Overall induction with 0.5 mM IPTG followed by incubation at 28 °C was determined to be the best expression condition. Upon harvesting and lysing the cells the desired protein migrated on an SDS-PAGE at around 29 kDa close to the predicted size of the twintron (nested intron) encoded HEase 32 kDa (Figure 3.3B). The purification of the protein was achieved by affinity chromatography involving Ni-NTA Superflow resin. A step up gradient with buffers containing 25 mM and 50 mM imidazole was used to remove the background proteins while a buffer containing 250 mM imidazole was used in protein elution; the purification of the protein was monitored by SDS-PAGE (Figure 3.3C). The desired fractions were pooled, dialyzed and the protein concentrated to a final concentration of 2.2 mg/mL. The HEase protein was purified in sufficient concentrations to pursue endonuclease and cleavage assays plus to conduct some investigations with regards to the thermal stability of the twintron (nested intron) encoded protein.

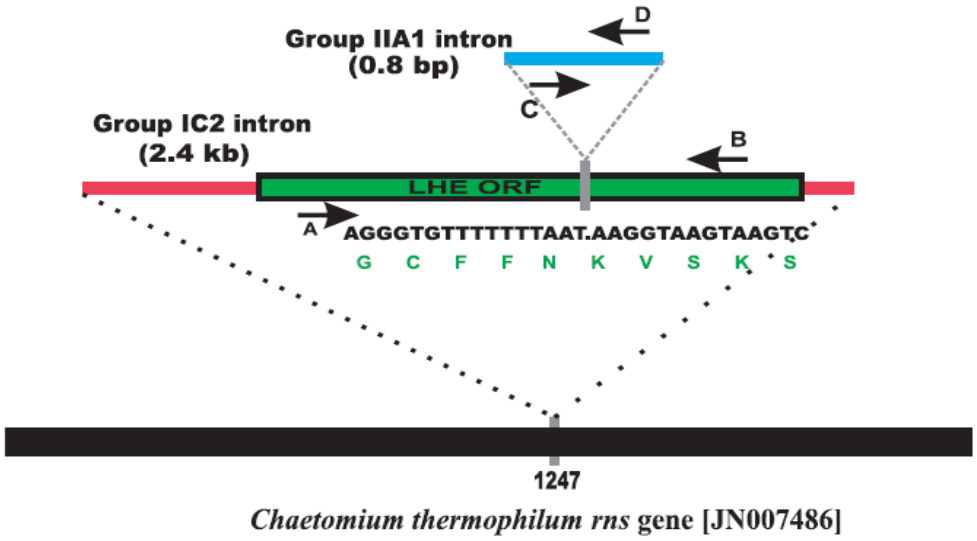
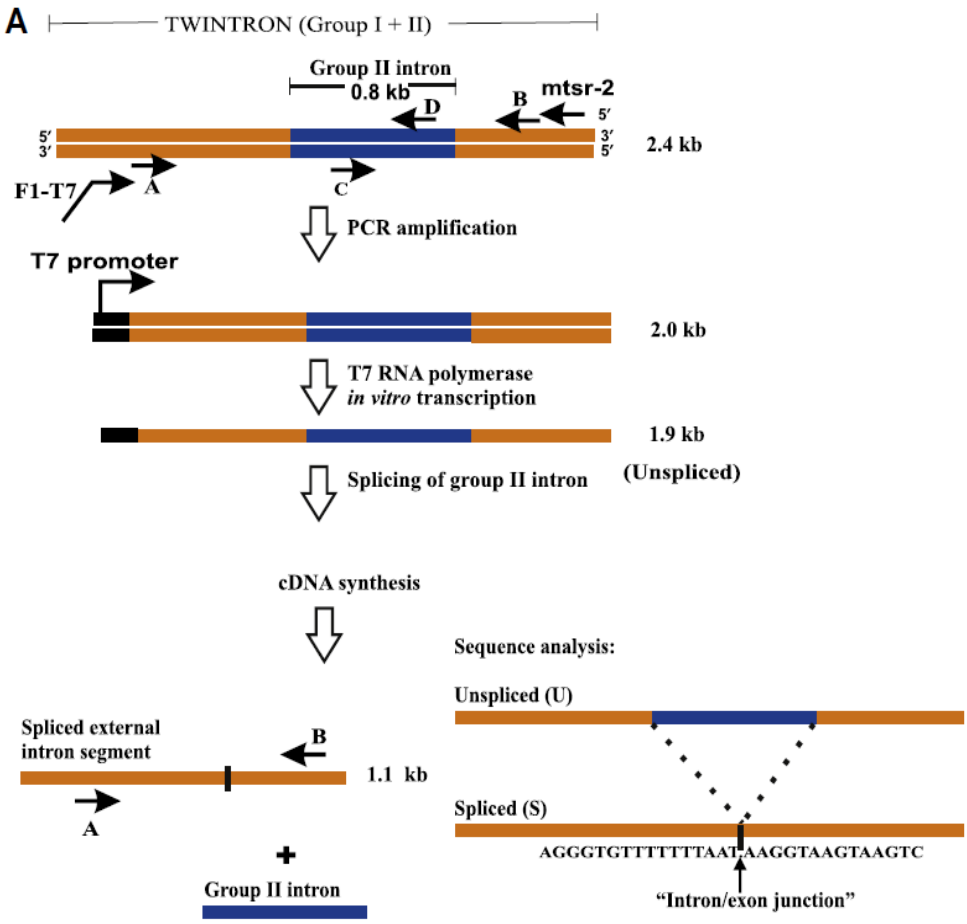


Figure 3.1. A schematic representation of the mtDNA *rns* gene (black line) of *C. thermophilum* strains DSM 1495 and UAMH 2024 (Hafez *et al*., 2013) and the twintron at S1247. At the S1247 position a group IC2 intron (orange line) that encodes a double motif LAGLIDADG type HEase (green line) is interrupted by an ORF-less group IIA1 intron (blue line). The arrows represent the position of primers (not to scale) within the twintron utilized in this study. A segment of the HEase ORF sequence is shown in black letters to illustrate the location of the group II intron insertion (represented by a single dot). The corresponding amino acid sequence is provided below the nucleotide sequence. The position of the group II intron is referred to as a phase 0 intron, as its position does not disrupt a codon.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).


A
TWINTRON (Group I + II)
Group II intron
0.8 kb
D
B
mtsr-2
5′
5′
3′
3′
5′
2.4 kb
F1-T7
A
C
PCR amplification
T7 promoter
2.0 kb
T7 RNA polymerase
in vitro transcription
1.9 kb
Splicing of group II intron
(Unspliced)
cDNA synthesis
Sequence analysis:
Unspliced (U)
Spliced external
intron segment
B
1.1 kb
A
Spliced (S)
AGGGTGTTTTTTTAATAAGGTAAGTAAGTC
Group II intron
"Intron/exon junction"

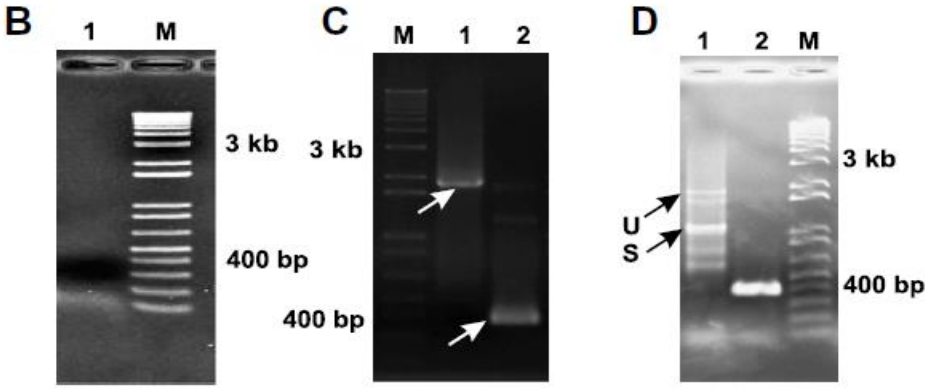

B
1
M
3 kb
400 bp
C
M
1
2
3 kb
400 bp
D
1
2
M
3 kb
U
S
400 bp

Figure 3.2. *In vitro* RNA splicing assay to determine the group II intron splice junction within the group I intron ORF. (**A**) A schematic representation of the *in vitro* transcription and splicing assay. Based on the location of the primers the DNA template generated with primers F1-T7 and mtsr-2 is 2.0 kb and the initial transcript based on utilizing the T7 promoter is expected to be approximately 1.9 kb (based on expected RT-PCR products recovered with primers A and B). The mature (spliced) transcript (as recovered by RT-PCR utilizing primers A and B) would be 1.1 kb as the internal group II intron consists of about 800 nucleotides. The predicted splice junction is shown in black letters and it is based on the position where the group I ORF is interrupted by the group II intron (see Figure 3.1). For panels B, C and D, 1% agarose gels are shown where RT-PCR products are resolved that were obtained from the *in vitro* transcription assays. *In vitro* RNA splicing assay to determine the group II intron splice junction within the group I intron ORF. (**B**) Lane 1 shows an RT-PCR reaction (forward primer A and reverse primer B) performed without a reverse transcriptase step, to confirm that all DNA (template) has been degraded after the completion of the *in vitro* transcription assay. (**C**) RT-PCR analysis (with primers 'A' and 'B') showing that under standard *in vitro* splicing conditions the group II intron failed to splice from the *in vitro* transcribed RNA precursor. The RT-PCR recovered amplicon was 1.9 kb indicating that the group II remained within the group I intron (lane1, white arrow). Lane 2 represents a positive control; by utilizing primers 'C' and 'D' the presence of the group II intron RNA (400 bp, white arrow) can be demonstrated. (**D**) *In vitro* splicing assay performed in the presence of high salt which was added to the standard *in vitro* transcription buffer. Here various amplicons were recovered by RT-PCR utilizing primers 'A' and 'B', but one dominant fragment was noted at 1.1 kb. Based on RT-PCR cDNA sizes the DNA fragment labeled 'U' (1.9 kb) represents the unspliced precursor and the DNA fragment labeled 'S' (1.1 kb) represents the spliced version of the transcript. The 1.1 kb fragment was recovered from the agarose gel and DNA sequence analysis confirmed the group II intron splice junction is as shown in panel **A**. The presence of the group II intron RNA (as in panel **C**, a 400 bp cDNA is indicated by a black arrow) is shown in lane 2. In all the agarose gels depicted 'M' represents the lane that includes the 1 kb plus^{TM} DNA ladder.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

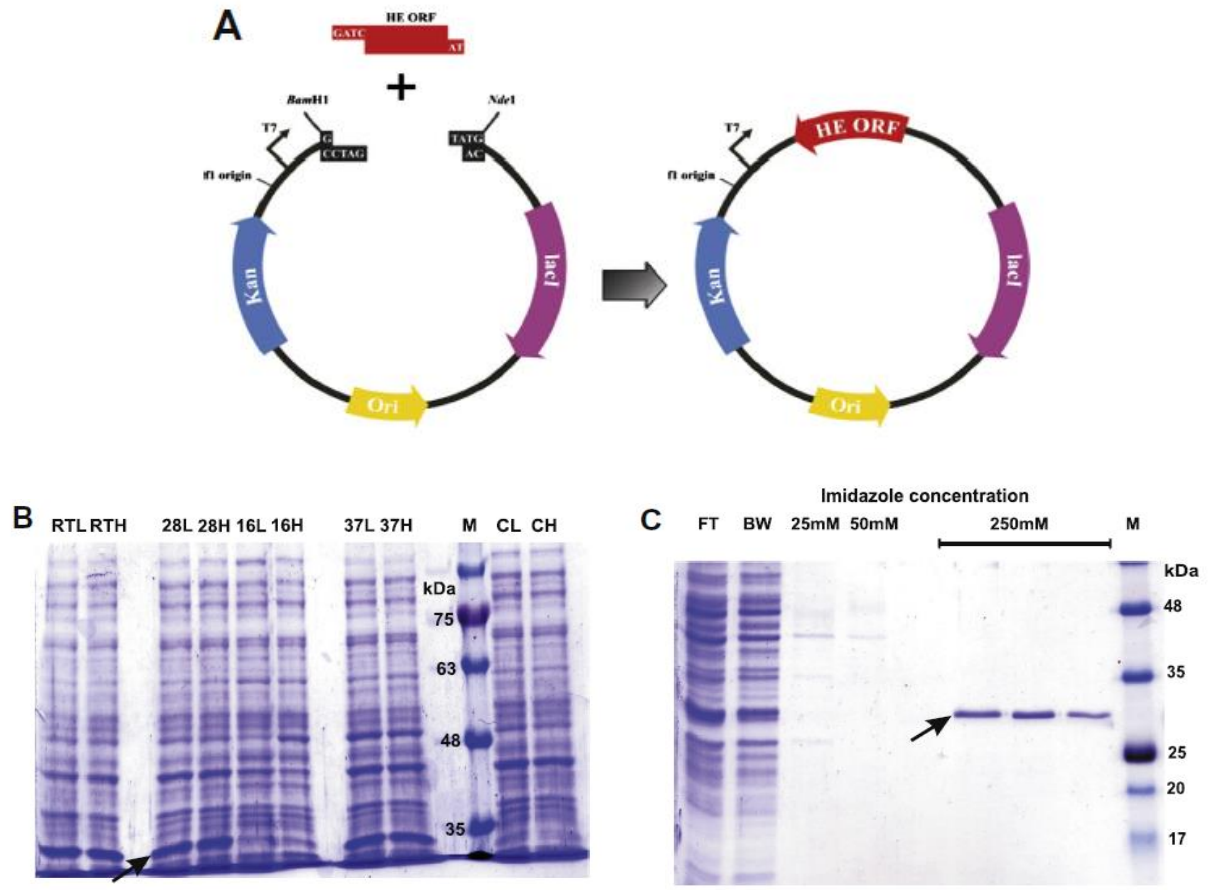


Figure 3.3. An overview of the expression plasmid is shown in (**A**); here the codon optimized (for *E.coli*) HEase ORF was cloned in the pET28 b (+) vector which provides a 6 x His-tag and is suitable for protein overexpression. (**B**) HEase protein overexpression and purification; 10% SDS-PAGE showing small scale protein expression trials using IPTG induced cells grown at various temperatures exposed to high (H= 1mM) or low (L= 0.5mM) IPTG concentrations; RT= Room temperature (~23 °C), 28, 16 and 37 °C. Lanes labeled C (i.e., CL and CH) represent protein expression profiles for the control plasmid, pET28 b (+). The arrow indicates the expected overexpressed protein (~29 kDa). (**C**) 12.5% SDS-PAGE showing the purification of the HEase protein using Ni-NTA resin. Lanes labeled FT shows the "Flow through" fraction and the lane labeled BW shows the "Binding wash" elution. The desired protein eluted at 250 mM imidazole (see arrow). 'M' in both the SDS-PAGE represents the BLUeye prestained protein ladder.

Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission License number: 3842180422262).

3.3.3. The mS1247 twintron encoded I-CthI is an active endonuclease

In vitro endonuclease assays were performed by incubating the purified I-CthI with circular and linearized versions of the substrate plasmid that contains the putative *rns* target site (Figure 3.4A). The endonuclease activity of the enzyme was tested at five different time points (0, 30, 60, 90 and 120 minutes) and at 60 minutes I-CthI completely linearized the circular substrate plasmid (3.1 kb; Figure 3.4B). Cleavage activity was already observed at time '0' time point, however time '0' means that the HEase was added and thereafter the stop buffer plus proteinase K was added. Therefore there was a short time period that allowed the enzyme to digest the substrate. Moreover, when a PCR derived DNA fragment (700 bp) containing the *rns* target site was used as the substrate, the enzyme cleaved the PCR product yielding two fragments (400 and 300 bp, Figure 3.4B). So both circular (supercoiled) and linear substrate molecules are cleaved by I-CthI. However, I-CthI did not cleave the non-substrate plasmid even after one hour of incubation at 37 °C (Figure 3.4C). LAGLIDADG HEases tend to generate cohesive termini by generating staggered cuts with 3′-OH, 4 nucleotide single stranded overhangs. T4 DNA polymerase can hydrolyze 3′ overhangs and thus blunt the cleaved DNA fragment and therefore one can indirectly characterize 3′ overhangs by comparative sequence analysis with uncut substrate DNA molecules (Bae *et al*., 2009). The T4 DNA polymerase treated and religated I-CthI-cleaved substrate plasmid sequence when compared with the sequence of the uncut substrate plasmid showed that a 5′ -AAGA- 3′ segment was removed from the sense strand (Figure 3.5A). So the cleavage site mapping experiment showed that I-CthI cleaves 8-nt downstream of the twintron insertion site (sense strand) or 4-nt downstream of position S1247 at the antisense strand (Figure 3.5B).

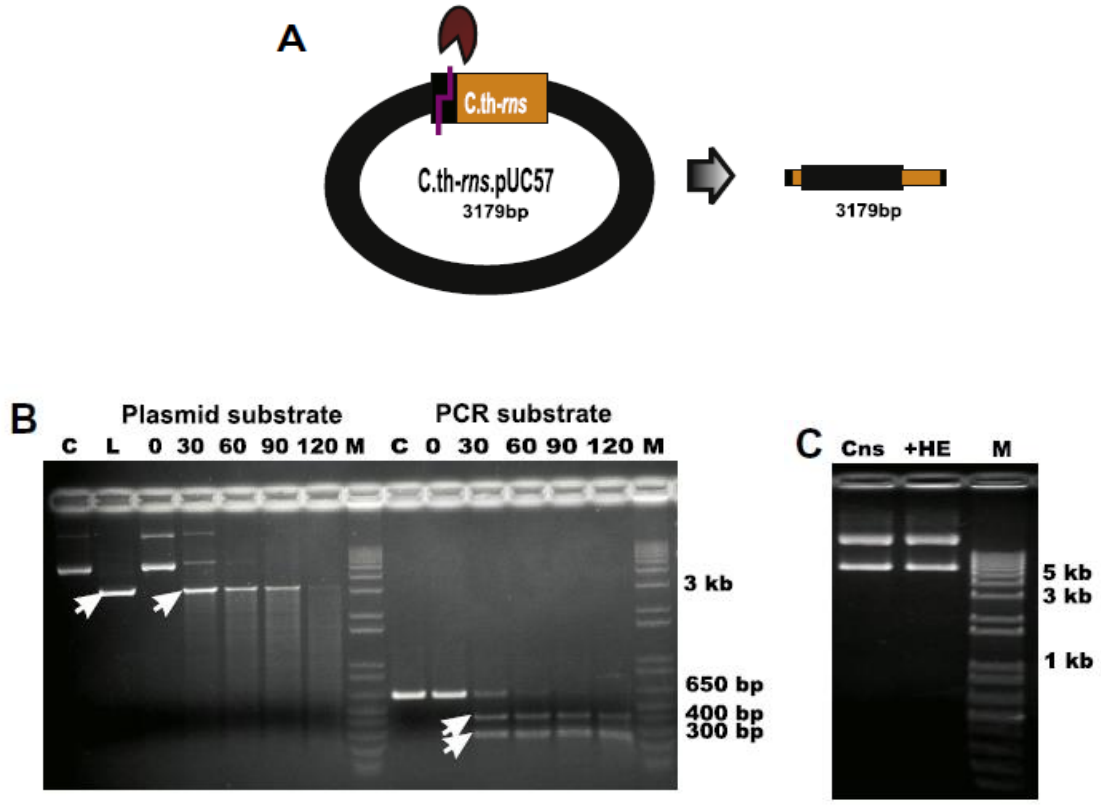


Figure 3.4. (**A**) Schematic overview of the *in vitro* endonuclease assay. An intron-less section (400 bp) of the *rns* gene was constructed and cloned in pUC57 (2.7 kb) to serve as the substrate plasmid (3.1 kb). A second construct with the target site interrupted by the intron (at mS1247) was used as the negative control. (**B**) *In vitro* endonuclease cleavage assay with the *C. thermophilum* twintron encoded HEase; a 1% agarose gel showing the results of the endonuclease assay. Lane C = uncut control substrate (in pUC57); lane L = linearized substrate plasmid (cleaved with BamHI); numbers on top of each lane represent incubation times in minutes at 37 °C. In all instances 1 µg plasmid was treated with the 5 µL aliquot of HEase (9 µg) and an arrow shows the linearized substrates. Similarly, the 650 bp PCR substrate was cleaved into two fragments (see arrows) when incubated with the same concentration of the HEase. (**C**) *In vitro* endonuclease cleavage assay with the *C. thermophilum* twintron encoded HEase when challenged with the negative control plasmid. In lane C_{ns} (control/non-substrate plasmid) no HEase was added and in the lane '+HE', the non-substrate plasmid was incubated with the same concentration of HEase as above for one hour at 37 °C. No cutting was observed and only high molecular weight supercoiled and concatenated plasmid DNAs were observed. For gels depicted in (**B**) and (**C**) the lane denoted as 'M' shows the 1kb plus™ DNA ladder.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

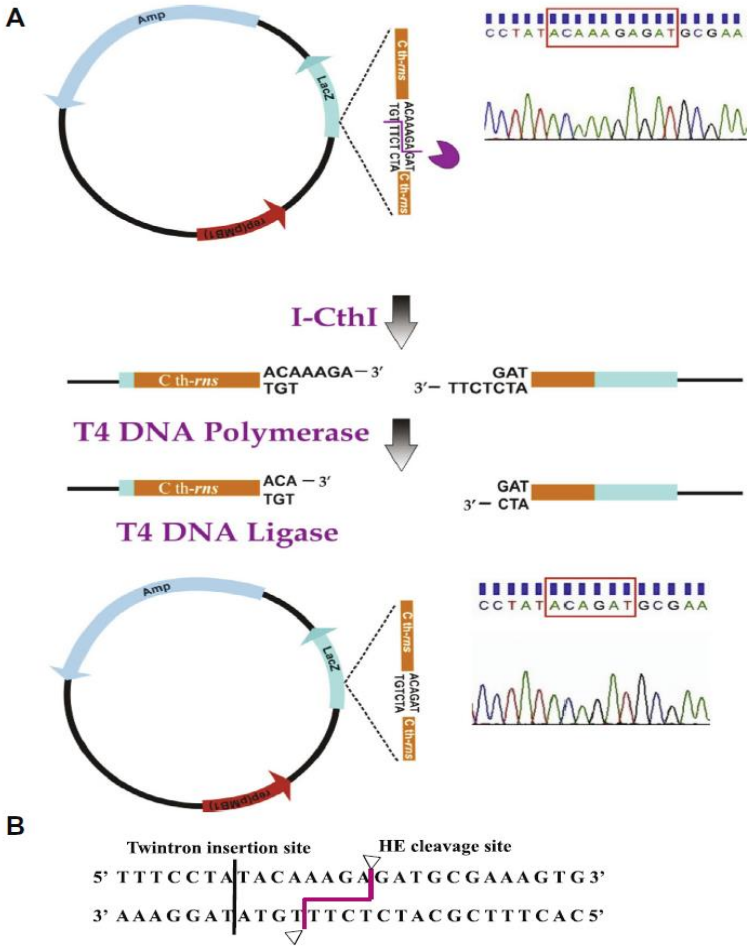


Figure 3.5. (**A**) Cleavage site mapping for the *C. thermophilum* twintron (nested intron) encoded HEase. The cleavage site was mapped by comparing uncut with I-CthI treated substrate DNAs. Cleavage by I-CthI generates a staggered cut with 4 nucleotide 3′-OH overhang in the substrate plasmid at the enzymes target site. The cleaved ends were blunted using T4 DNA polymerase. The religated plasmid was sequenced and compared to the sequence of the untreated substrate plasmid in order to map the cleavage site by scanning for a 4 bp deletion in the T4 DNA polymerase treated cleaved substrate plasmid. (**B**) Schematic representation of the I-CthI cleavage site near the twintron insertion sequence. Proposed cleavage sites are indicated by open triangles; and a vertical line represents the twintron insertion site. The HEase cleavage site is 8-nt downstream of the twintron insertion site with regards to the sense strand or 4-nt downstream with regards to the antisense strand.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

3.3.4. The effect of temperature on I-CthI endonuclease activity and stability

I-CthI was challenged with the substrate plasmid at a temperature range starting from 25 °C in 10 °C increments until 85 °C (Figure 3.6A). The enzyme showed activity at 25, 35 and 45 °C; albeit at 25 °C and 45 °C there appeared to be evidence for the present of considerable amount of uncut substrate. No visible cleavage activity was noted at 55 °C and above (Figure 3.6A).

The stability of the I-CthI protein over the same temperature range as was used to evaluate its endonuclease activity was evaluate by circular dichroism (CD) spectropolarimetry. Circular dichroism allows for the prediction of protein secondary structural features, by splitting plane polarized light into its left and right components and by monitoring the differences in absorbance between the two components (Kelly *et al*., 2005). When the sample is subjected to different temperatures during the CD analysis, one can evaluate the thermal stability of the protein by monitoring changes in its secondary structure. Far-UV CD spectra's of the I-CthI protein over a temperature range from 25 °C to 85 °C (Figure 3.6B) showed a shift of the bands at 222 nm and 208 nm towards less negative values, reflecting the reduced fraction of α-helical segments (Greenfield, 2006). In addition, we noted that there was a shift of bands at 210 nm towards the negative direction possible showing an increase in disorder within the protein (Kelly *et al*., 2005; Greenfield, 2006). These changes were particular noticeable at 55 °C and above. This loss of α-helical segments correlates with the data from the endonuclease assays that showed a reduction of endonuclease activity at 45 °C and a lack of detectable cutting activity at or above 55 °C.

The online program PHYRE2 was used to examine the secondary and tertiary structure of I-CthI (Figure 3.6C). The program showed that the protein is comprised of ten alpha helices

(39%) and nine beta strands (27 %), the confidence key of these regions was found to be 100% when compared with the crystal structure of another homing endonuclease I-OnuIi2 (PDB accession number: c3qqyA). There were also a few disordered regions in the proteins tertiary structure starting from amino acid 134 to 164 and the overall disorderness was estimated to be 22%.

3.3.5. Co-crystallization trials

The PEG Suite containing either 25 mM $CaCl_2$ or absence of divalent cations did not generate any crystals. However, PEG 1- 48 in the presence of 500 mM $CaCl_2$ generated only one promising quasi structure with the 23 bp DNA target site. The Wizard crystallographic screen also yielded only one crystal with the 25 bp DNA target site. None were suitable for detailed characterization. Further structural characterization was not possible with this protein as it was reported to be highly soluble and did not precipitate or crystallize in any of crystallographic screens.

3.3.6. Phylogenetic relationship of the I-CthI HEase

The programs utilized to infer phylogenetic relationships for the aligned LAGLIDADG data set yielded similar tree topologies and in all cases the trees received moderate node support values although deeper nodes received poor support values (Figure 3.7). However, strong support was noted for the node that unites the I-CthI twintron ORF with other mS1247 group I intron encoded ORFs suggesting that most likely these introns share a common ancestor and the group II intron was probably inserted more recently within the *C. thermophilum* ancestral mS1247 intron. Also Bayesian, ML and NJ analysis provided significant support for suggesting

that the mS1247 intron ORFs share a common ancestry to group I intron ORFs located within protein coding genes such as *nad4L* and *cox1*. It is also worthwhile to note that within this data set only *C. thermophilum* can be classified as a thermophile.

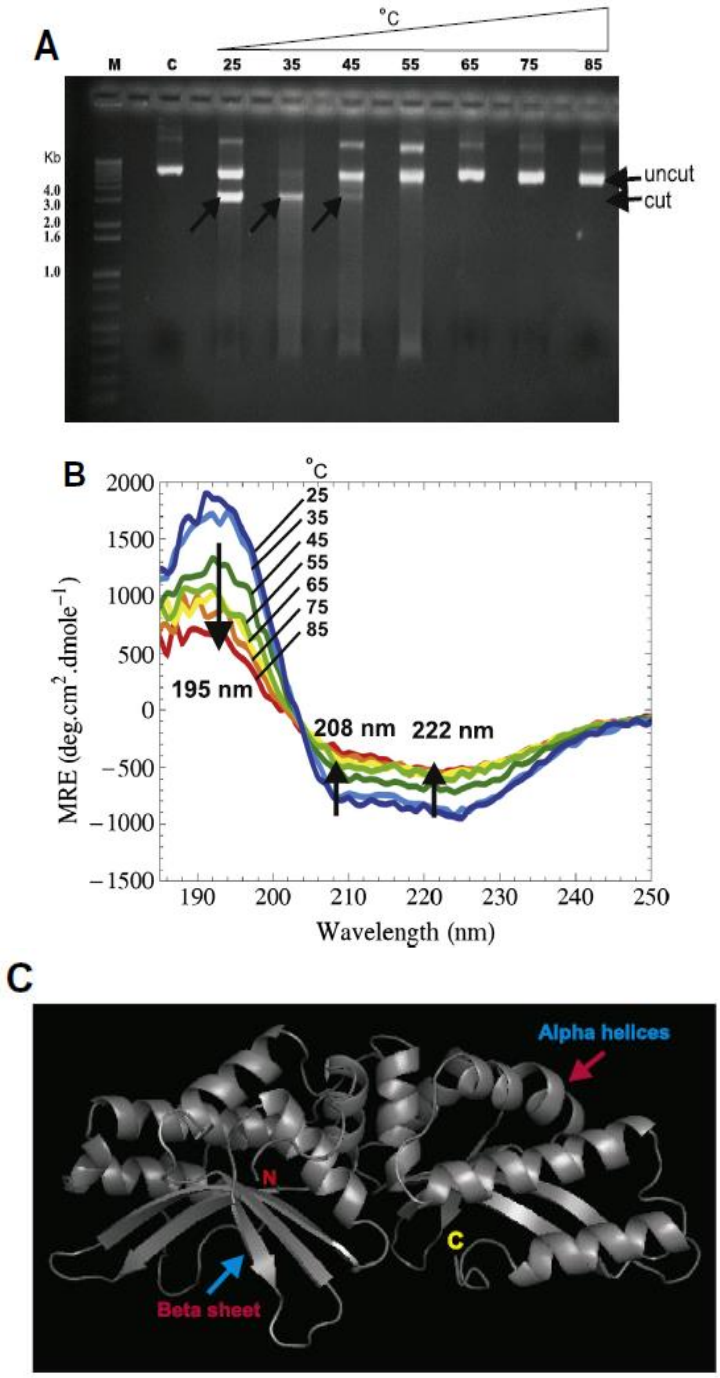


Figure 3.6. Effect of temperature on I-CthI endonuclease activity. (**A**) 1% agarose gel showing the effect of temperature on the *in vitro* endonuclease activity when the reactions were incubated at a temperate range from 25 °C to 85 °C in 10 degree intervals for 1 hour. The lane marked 'M' contains the 1 kb plus™ DNA ladder and the lane marked 'C' contains a control where no HEase

was added to the substrate plasmid. The enzyme appears to cut most efficiently at 35 °C while the cleavage activity diminishes as the temperature rises (linearized substrate shown by arrow). (**B**) Temperature stability for I-CthI examined by circular dichroism spectropolarimetry. Spectra were recorded for the following temperate range: from 25 °C to 85 °C in 10 degree intervals for 30 minutes and the corresponding CD spectra were plotted. The values recorded between 208 nm and 222 nm slowly move towards the 0 base line as the temperature increases. This region of the spectra is indicative for alpha helices and a shift towards less negative values corresponds to the loss of the structural integrity in those alpha helices. Also note that the decrease in the spectras at 195 nm over the tested temperature range is indicative of an increase in structural disorder, i.e. accumulation of random coils in the proteins secondary structure. (**C**) An *in silico* model for the I-CthI protein generated by the PHYRE2 program. The program identified the double motif LAGLIDADG I-OnuI (PDB: c3qqyA) HEase protein as a template for folding I-CthI. The model shows the symmetrical nature of this protein and the alpha helices and beta sheets along with amino terminal (N) and carboxyl terminal (C) have been marked. The beta sheets arrange in a configuration that forms the DNA binding surface of the HEase and the LAGLIDADG motifs contribute toward the active site of these enzymes (reviewed in Stoddard, 2006).
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

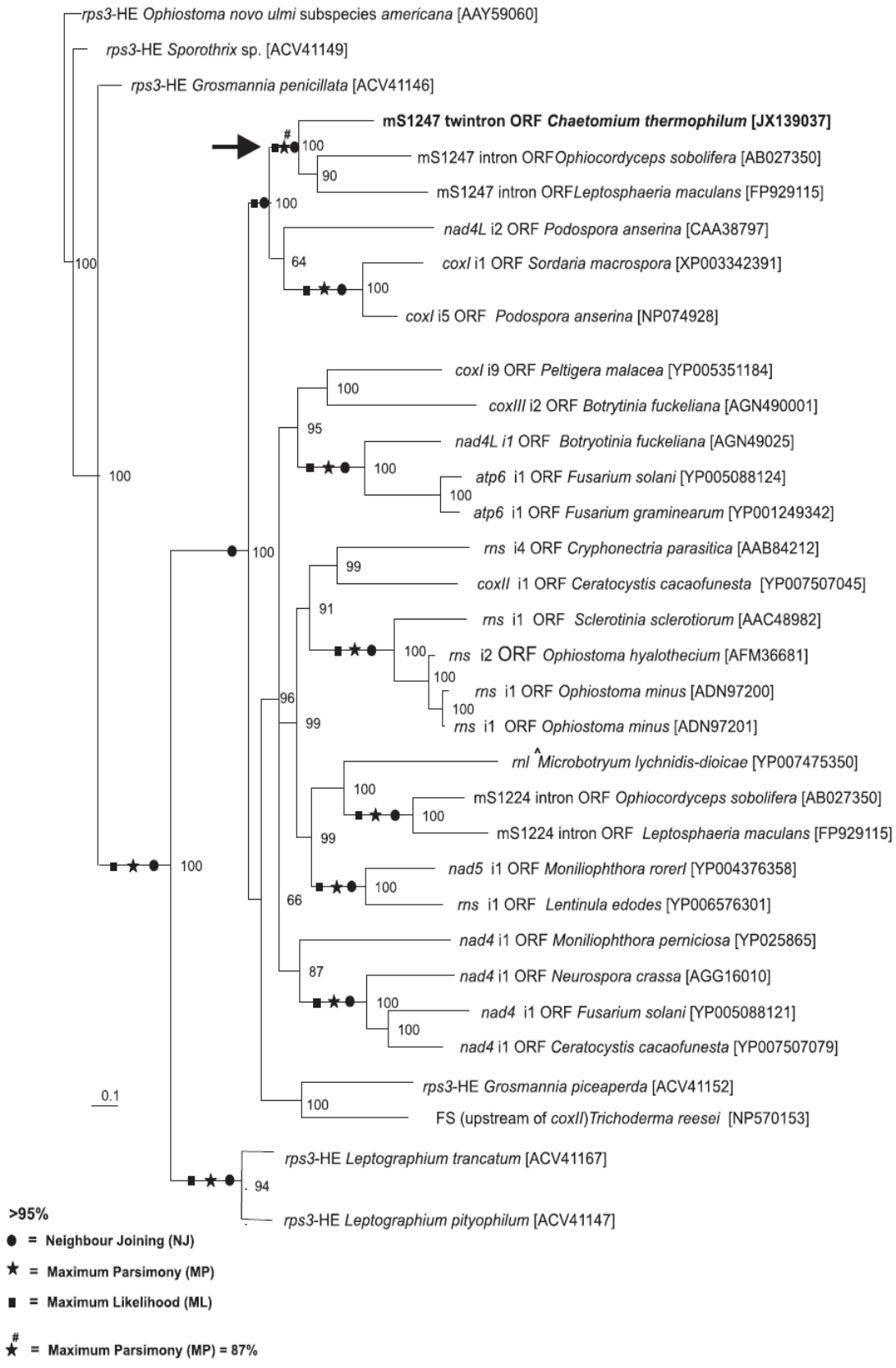

rps3-HE Ophiostoma novo ulmi subspecies americana [AAY59060]
rps3-HE Sporothrix sp. [ACV41149]
rps3-HE Grosmannia penicillata [ACV41146]
mS1247 twintron ORF Chaetomium thermophilum [JX139037]
mS1247 intron ORF Ophiocordyceps sobolifera [AB027350]
mS1247 intron ORF Leptosphaeria maculans [FP929115]
nad4L i2 ORF Podospora anserina [CAA38797]
coxI i1 ORF Sordaria macrospora [XP003342391]
coxI i5 ORF Podospora anserina [NP074928]
coxI i9 ORF Peltigera malacea [YP005351184]
coxIII i2 ORF Botrytinia fuckeliana [AGN490001]
nad4L i1 ORF Botryotinia fuckeliana [AGN49025]
atp6 i1 ORF Fusarium solani [YP005088124]
atp6 i1 ORF Fusarium graminearum [YP001249342]
rns i4 ORF Cryphonectria parasitica [AAB84212]
coxII i1 ORF Ceratocystis cacaofunesta [YP007507045]
rns i1 ORF Sclerotinia sclerotiorum [AAC48982]
rns i2 ORF Ophiostoma hyalothecium [AFM36681]
rns i1 ORF Ophiostoma minus [ADN97200]
rns i1 ORF Ophiostoma minus [ADN97201]
rnl ^Microbotryum lychnidis-dioicae [YP007475350]
mS1224 intron ORF Ophiocordyceps sobolifera [AB027350]
mS1224 intron ORF Leptosphaeria maculans [FP929115]
nad5 i1 ORF Moniliophthora rorerl [YP004376358]
rns i1 ORF Lentinula edodes [YP006576301]
nad4 i1 ORF Moniliophthora perniciosa [YP025865]
nad4 i1 ORF Neurospora crassa [AGG16010]
nad4 i1 ORF Fusarium solani [YP005088121]
nad4 i1 ORF Ceratocystis cacaofunesta [YP007507079]
rps3-HE Grosmannia piceaperda [ACV41152]
FS (upstream of coxII)Trichoderma reesei [NP570153]
rps3-HE Leptographium trancatum [ACV41167]
rps3-HE Leptographium pityophilum [ACV41147]
0.1
>95%
= Neighbour Joining (NJ)
= Maximum Parsimony (MP)
= Maximum Likelihood (ML)
= Maximum Parsimony (MP) = 87%

Figure 3.7. Phylogenetic tree showing the phylogenetic position of the mS1247 twintron encoded LAGLIDADG ORF. The tree topology is based on Bayesian analysis, solid circles, squares and astrices represents nodes that received bootstrap support values >95 % in NJ, ML, and MP analysis. Posterior Probability (PP) supportive values are recorded at the nodes and these were obtained from a 50% majority Bayesian consensus tree. The branch lengths are based on Bayesian analysis and are proportional to the number of substitutions per site. GenBank accession numbers [in square brackets] are listed next to species names. The outgroup designated for this analysis was the HEase ORF inserted within the mtDNA *rps3* locus in *Ophiostoma novo-ulmi* subsp. *americana*.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

3.4. Discussion

3.4.1. The twintron (nested intron) encoded split ORF encodes an active homing endonuclease

Characterization of the mS1247 twintron encoded reconstituted ORF showed that it encodes a functional HEase that cuts downstream of the twintron insertion site. The results also showed that the twintron encoded ORF, which is interrupted by a group II intron, is functional when the internal group II intron is removed. The twintron encoded HEase was able to cleave both plasmid and linear (PCR product) substrates that contained the *rns* target site. So this HEase has the potential of providing this composite element with the ability to insert into cognate alleles that lack an insertion.

We also showed that the internal group II intron has the potential to self-splice under laboratory conditions in the presence of high salt concentration. This is commonly observed in group II intron *in vitro* splicing assays (Toor *et al*., 2006; Mullineux *et al*., 2010; Fedorova, 2012) and one typically assumes that under cellular conditions either intron-encoded or host factors or both facilitate the formation of splicing competent group II intron RNA ribozyme configurations (Bonen and Vogel, 2001; Lambowitz and Zimmerly, 2011; Hausner, 2012). Nested introns or twintrons, combinations of self-splicing introns embedded within another potentially self-splicing intron have been noted in protists rDNA, fungal mtDNAs and algal chloroplast genomes (Drager and Hallick, 1993; Einvik *et al*., 1998a; Hafez *et al*., 2013). Analogous spiceosomal version of nested introns/twintrons also exits in Fungi and Metazoan nuclear genomes (Flipphi *et al*., 2013; Janice *et al*., 2013). These nested elements may offer insights on the evolution of complex nuclear introns (Janice *et al*., 2013; Suzuki *et al*., 2013) or in the case of group I and group II introns the evolution of composite organellar mobile elements

and novel ribozymes (Einvik *et al*., 1998b; Khan and Archibald, 2008; Moreira *et al*., 2012; Pombert *et al*., 2012). A similar situation to the mS1247 twintron has been noted in the *rnl-rps3* (ribosomal protein 3) locus in *Grosmannia piceiperda*. Here a twintron (mL2449) has an internal group I intron interrupting the external group I intron's rps3 open reading frame, but upon splicing of the internal intron, the rps3 ORF is reconstituted thus allowing for the potential expression of the RPS3 protein (Rudski and Hausner, 2012). Obviously the evolution of nested elements requires compatibility and co-evolution among its various constituents in order to allow for efficient splicing and expression of protein products; raising questions with regards to their mode of evolution as a simple neutral evolution model (Goddard and Burt, 1999; Gogarten and Hilario, 2006) may not explain their persistence within a population. Such complex multicomponent associations would be eliminated almost immediately if mutations could accumulate quickly due to lack of selection within the ORFs and ribozyme components; clearly this needs to be investigated in more detail in future studies.

With regards to the mS1247 internal group II intron, this element might benefit from this association by gaining a neutral location thus minimizing impact to the host genome plus the ORF less group II intron still has the ability to be mobilized as part of the twintron unit by the DNA based mobility mechanism that drives group I intron homing. Although this needs experimental investigations one can speculate that the splicing of the group II introns may be a regulatory step that can control the amount of HEase protein that can be translated from the processed group I intron RNA. The survival of these elements depends on their ability to spread into cognate (or new) sites that lack introns and by minimizing their impact on the host genome (Gillham, 1994; Goddard and Burt, 1999; Edgell *et al*., 2011). Therefore the production of the

HEase is important but one would expect that excess production would be a drain on the host system.

3.4.2. Origin of the twintron

The origin of the group II intron at this stage is speculative; blastn searches did not reveal the presence of similar introns within the NCBI data base. Various scenarios can be envisioned as to the origin of the internal intron of the mS1247 twintron. It has been demonstrated that group II introns that lack ORFs may not have lost their ability for homing or ectopic integration into new sites. For example Moran *et al*. (1995) noted that in yeast, a mtDNA group II intron lacking reverse transcriptase activity could move based on a DNA-level recombination mechanism. It is also possible that the mS1247 internal intron was mobilized by retrotransposition facilitated by a trans-acting factor provided by a related reverse transcriptase-encoding group II intron. With regards to invading ectopic sites, two mechanisms have been proposed, one involving reverse splicing of the intron RNA into an ectopic site within another RNA or reverse splicing into an ectopic site within DNA (Zimmerly *et al*., 1995a). Reverse splicing within RNA would require two additional steps, reverse transcription into cDNA and the integration of the cDNA into the host genome by recombination (Mueller *et al*., 1993; Zimmerly *et al*., 1995a, b; Bonen and Vogel, 2001; Cousineau *et al*., 2000). So far strong experimental evidence for reverse splicing into RNA is still lacking but evidence for ectopic integration of group II intron RNA into DNA sites that resemble the intron's native homing sites (IBS1 and IBS2) has been demonstrated (Yang *et al*., 1996; Yang *et al*., 1998; Dickson *et al*., 2001).

Assuming that the internal group II intron is a recent addition to the mS1247 intron ORF locus sequence comparisons between the three currently available examples may hint at the

possibility of ectopic integration (Figure 3.8). The sequence upstream of the site where the group II intron has inserted within the *C. thermophilum* twintron shows some conservation. So this sequence among the mS1247 intron ORFs may indeed have had sufficient resemblance in *C. thermophilum* mS1247 intron to the native homing site of the internal group II intron. This would have allowed for the group II intron EBS1 and EBS2 sequences to interact with ORF based IBS1 and IBS2 sequences and facilitate the insertion of the intron by either reverse splicing at the RNA level or, by means of a trans-acting reverse transcriptase, reverse splicing at the DNA level (Figure 3.8).

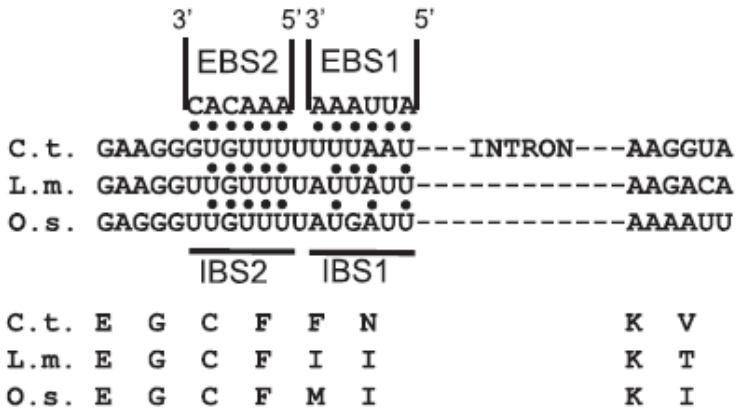


Figure 3.8. Comparison of the internal group II intron "exon" binding sequences (EBS) and corresponding intron binding sequences of the mS1247 internal group II intron with sequences found in the ORFs located in non-twintron versions of the mS1247 intron [(*Ophiocordyceps sobolifera* (O.s.) and *Leptosphaeria maculans* (L.m.); Genbank accessions ABO27350 and FP929115 respectively] at the site where the internal intron is found in *C. thermophilum* (C.t). The comparison indicates potential base parings (indicated by filled circles) between the internal intron's EBS sequences and its IBS sequences and sequences found in other alleles (mS1247 intron ORFs).Translation of the putative IBS region sequence is also shown as it relates to the LAGLIDADG ORF encoded by the mS1247 group I introns.
Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68. (Elsevier Publications. Image reproduced with permission. License number: 3842180422262).

3.4.3. A homing endonuclease with a possible "on" switch.

Currently several gene targeting endonucleases are being developed based on scaffolds that include group II intron based targetrons, zinc-fingers, TALENs, the CRISPR/cas9 system and HEases (Karberg *et al*., 2001; Stoddard, 2011; Hafez and Hausner, 2012; Gaj *et al*., 2013; Ran *et al*., 2013; Marton *et al*., 2013). The advantage of native HEases is their high degree of specificity and that they can be uncovered by exploring mobile introns, i.e. they do not need to be synthesized *de novo*. In general HEases require long DNA recognition sites and therefore cut infrequently within a genome; this makes them useful for DNA engineering (i.e. genome editing) (Gimble, 2005, 2007). HEases are currently employed to induce mutations and for gene replacement strategies (Storici *et al*., 2003; Gimble, 2005; Marcaida *et al*., 2010; Siegl, 2010). With regards to gene replacements the strategy employed is to generate a double stranded break in the targeted gene and by co-transforming simultaneously the cells with a segment of DNA that shares homology with the target sequence and therefore homologous recombination would allow for gene replacement (Stoddard, 2011).

Usually HEase assays originating from non-thermophiles are performed at 30 ^{o}C to 37 ^{o}C (Kowalski and Debyshire, 2002; Sethuraman *et al*., 2009) and the twintron encoded HEase demonstrated endonuclease activity over a wide temperature range from 25 ^{o}C up to 45 ^{o}C but it appears to lose activity above 45 ^{o}C; the latter based on the CD spectropolarimetry temperature assays is probably due to loss of structural integrity. Overall the temperature range for I-CthI endonuclease activity is similar to that employed for many restriction enzymes so I-CthI could be readily utilized as a rare cutting endonuclease under standard laboratory conditions.

In biotechnological applications it is sometimes desirable to control HEase activity. The twintron ORF studied herein offers a system where the internal group II intron could be the key

to engineer an endonuclease with an "on switch". Splicing of the group II intron could be a regulatory step that allows for the maturation of the HEase transcript and thus translation of the ORF. This may have applications in bacterial systems which are more emendable to group II intron splicing (Toor *et al*., 2006; Yao and Lambowitz , 2007; Yao *et al*., 2013) unlike eukaryotic system where so called debranching enzymes can degrade transcripts that have complex folds such as those generated by group II intron RNAs (Mastroianni *et al*., 2008). In general controlling the cleavage activity of HEases is considered a desirable feature in order to optimize these elements as tools for gene replacements as it would allow for controlling HEase activity based on cellular conditions and thus reduce or delay potential toxic effects on the cell when HEase activity is not desirable (Posey and Gimble, 2002).

Double motif LAGLIDADG HEase are compact monomeric endonucleases and the genes that encode them can be recovered from many microbial and organellar genomes (Hafez and Hausner, 2012; Hafez *et al*., 2014) thus bioprospecting for HEGs might be an avenue of acquiring protein scaffolds that can be used directly or modified in order to develop tools for genome editing by targeting specific sequences (Baxter *et al*., 2012). The I-CthI is a promising addition to the currently characterized fungal derived HEases for future applications in biotechnology (Taylor *et al*., 2012). The mS1247 twintron also demonstrates how composite mobile elements can evolve by different categories of mobile introns inserting into one another.

I would like to thank Dr. Mohamed Hafez for discussions and sharing ideas about the mS1247 intron and Dr. Joe O'Neil (Department of Chemistry, University of Manitoba, Canada) for generously allowing access to his laboratory and assistance in the circular dichroism spectropolarimetry work. Finally, I would like to thank Dr. Steven Zimmerly (Biological Sciences, University of Calgary, Canada) for his suggestions on group II intron *in vitro* splicing assays.

Chapter 4

Using group II introns for attenuating the *in vitro* and *in vivo* expression of a homing endonuclease

4.0. Abstract

In *Chaetomium thermophilum* DSM 1495 within the mitochondrial DNA (mtDNA) small ribosomal subunit (*rns*) gene, a group IIA1 intron interrupts an open reading frame (ORF) encoded within a group I intron (mS1247). This arrangement offers the opportunity to examine if the nested group II intron could be utilized as a regulatory element for the expression of the homing endonuclease (HEase). Constructs were generated where the codon-optimized ORF was interrupted with either the native group IIA1 intron or a group IIB type intron. This study showed that the expression of the HEase (*in vivo*) in *Escherichia coli* can be regulated by manipulating the splicing efficiency of the HEase ORF-embedded group II introns. Exogenous magnesium chloride ($MgCl_2$) stimulated the expression of a functional HEase but the addition of cobalt chloride ($CoCl_2$) to growth media antagonized the expression of HEase activity. Ultimately the ability to attenuate HEase activity might be useful in precision genome engineering, minimizing off target activities, or where pathways have to be altered during a specific growth phase.

The work presented in this chapter has been published.
Guha TK, Hausner G. 2016. Using Group II Introns for Attenuating the *In Vitro* and *In Vivo* Expression of a Homing Endonuclease. PLoS ONE. **11**(2): e0150097.

Conceived and designed the experiments: TKG, GH. Performed the experiments: TKG. Analyzed the data: TKG, GH. Contributed reagents/materials/analysis tools: GH. Wrote the paper: TKG, GH.

The detailed protocol for inserting ribozyme based switch in homing endonuclease genes is provided in Chapter 7: Appendices (S.7.2).

4.1. Introduction

Homing endonucleases (HEases) are site-specific DNA cleaving enzymes that are encoded by homing endonuclease genes (HEGs) which are frequently found embedded within archaeal introns, composite mobile genetic elements such as group I introns, group II introns, inteins (Gimble, 2000; Hausner, 2012; Edgell *et al*., 2000; Toor and Zimmerly, 2002; Dujon, 1989) and sometimes HEGs are freestanding (Dujon, 1989; Belfort and Perlman, 1995; Mueller *et al*., 1996; Stoddard, 2005; Dalgaard *et al*., 1997). The LAGLIDADG family (LHEases) are frequently encoded within fungal mitochondrial group I introns (Haugen and Bhattacharya, 2004; Silva *et al*., 2011) and HEases in general recognize long asymmetrical 12-40 bp of DNA sequences as their target sites and cleave in a manner that generates four nucleotide 3′-OH overhangs (Belfort and Perlman, 1995; Mueller *et al*., 1996; Stoddard, 2006). As LHEases require long DNA target sequences they cut infrequently within a genome (Stoddard, 2006) and this feature has been utilized for various applications in genome editing as it relates to agriculture (Gao *et al*., 2010), population control of disease vectors (Deredec *et al*., 2008; Windbichler *et al*., 2007; Chan *et al*., 2011), and human health (Davé *et al*., 2009; Grizot *et al*., 2006; Takeuchi *et al*., 2011). In this study we are testing the possibility of utilizing mitochondrial group II intron sequences as an on/off "switch" system that provides the opportunity for temporal control of HEase activity in *Escherichia coli*.

Previously a molecular switch was developed that controlled the endonuclease activity of PI-SceI *in vitro*; here two cysteine amino-acid residue pairs were separately inserted into the HEase DNA binding loops to allow for disulfide bond formation that lock the endonuclease into a nonproductive conformation (Posey and Gimble, 2002). This essentially is a redox switch and the activity of the protein could be controlled by adding or removing a reducing agent (Posey and Gimble, 2002). Other strategies suggested for regulating endonuclease activity included

manipulating metal ion cofactors (such as Mg^{+2}) or developing temperature sensitive versions of HEases (Muir *et al*., 1997). However, in order to be a true on/off "switch" it was suggested that manipulating cellular concentrations of ion cofactors might be difficult and using different temperatures might pose a problem with some cell lines or some temperature sensitive HEases once misfolded could not be reactivated (i.e. refolded) (Posey and Gimble, 2002). Redox switches as developed for PI-SceI have potential for *in vitro* applications but are not practical for *in vivo* applications as the targeted cell would probably suffer damage if oxidizing conditions were applied.

Recently we characterized a twintron-like arrangement (or 'nested' intron) at position S1247 in the mitochondrial DNA (mtDNA) small ribosomal subunit (*rns*) gene of *Chaetomium thermophilum* var. *thermophilum* La Touche (strain DSM 1495). The external intron encodes a double-motif LAG HEase ORF (I-CthI) which is interrupted by an internal ORF-less group IIA1 intron (Hafez *et al*., 2013). The group IIA1 intron was shown by *in vitro* self-splicing experiments to be excised and thus generating a transcript where the LAGLIDADG HEase (LHEase) ORF is contiguous and could allow for the expression of a functional endonuclease (see Chapter 3; Guha and Hausner, 2014). Splicing of the internal group II intron could be a regulatory step that allows for the maturation of the HEase transcript and translation of the open reading frame. Splicing of group II introns requires the intron RNA to assume a splicing competent tertiary fold that includes interactions between intron and flanking exon sequences. Folding of the intron RNA is in part facilitated by base pair complementarities (Lambowitz and Zimmerly, 2004). In addition various intron and/or host genome encoded factors tend to assist *in vivo* RNA folding and splicing of group II intron from transcripts (Lambowitz and Zimmerly, 2004; Olga and Nora, 2007). Group II intron RNAs are ribozymes composed of six helical

regions, referred to as domains I through VI, emerging from a central wheel and these domains interact to form a conserved tertiary structure that brings together distant sequences to form an active site. The active site binds the splice sites and the branch-point nucleotide residue and uses bound Mg^{+2} ions to activate the appropriate bonds for catalysis (Michel *et al*., 2009; Lambowitz and Zimmerly, 2011). Terbium cleavage assay determined domain V (DV) to be an important region of the active site as it contains the catalytic triad AGC and an AY bulge, both of which bind Mg^{+2} (Gordon and Piccirilli, 2001). Experimental results from phosphorothioate substitutions at the splice sites suggested that group II introns either use separate active sites with different Mg^{+2} ions to catalyze the two splicing (transesterification) steps or a single active site which is rearranged between the steps (Pyle and Lambowitz, 2006; Zhang and Doudna, 2002). Moreover, X-ray crystallographic structure of the catalytic core of an *Oceanobacillus iheyensis* IIC intron revealed that two helices of DV are bent to bring the AC bulge near the CGC triad, thus juxtaposing the phosphate backbones of the most conserved DV sequences, rather than stacking coaxially (Toor *et al*., 2008). Nine potential Mg^{+2} binding sites were assigned in or near DV for this intron (Zimmerly and Semper, 2015). Hence, a potential trigger for splicing might involve the availability of Mg^{+2} inside *E. coli* cells and this property has been applied in this research to develop a mechanism for the *in vivo* regulation of HEase (I-CthI) activity.

In order to evaluate if group II intron sequences could be manipulated as regulatory elements that could prevent or at least attenuate HEase expression, constructs have been generated where a pET28b (+) expression vector contains the I-CthI sequence that is interrupted by its native group IIA1 intron sequence (Hafez *et al*., 2013). The HEase sequence was optimized for expression in *E. coli* and for maintaining the intron binding sequences (IBS1 and IBS2) which are located upstream of the intron insertion site. The IBS elements are needed for

splicing as they interact with the corresponding exon binding sequences (EBS1 and EBS2) present within the intron in part to establish a splicing competent fold (Zimmerly and Semper, 2015). A second construct was designed where the native IIA1 intron was replaced by a mitochondrial group IIB intron (no ORF) [rI1 of *Scenedesmus obliquus*] along with its corresponding IBS1 sequence (Hollander and Kück, 1999).

It has been suggested that the *in vivo* control of HEase activity would allow for more precise temporal inactivation or modification of genes (Hafez and Hausner, 2012; Stoddard, 2014) for example to shift metabolic processes during a particular growth phase of the bacteria that are being manipulated for the production of certain metabolites or proteins (Hoefel *et al*., 2012). Therefore, this nested intron arrangement offers the opportunity to examine if the group II intron could be utilized as a regulatory element for the expression of the I-CthI HEase.

4.2. Materials and Methods

The methods exclusively related to this chapter have been detailed in this section. For the common materials and methods used in this chapter (which are appropriately mentioned in the text), the readers are directed to Chapter 2 (General Materials and Methods).

4.2.1. Design of the *Escherichia coli* expression vectors and substrate

An expression plasmid with the codon-optimized version of the I-CthI ORF along with its native internal group IIA1 intron (GenBank accession number: JX139037.1) was synthesized by Genscript. However, the IBS1 and IBS2 sequences which are upstream of the intron insertion site (IBS1: 5′ TGTTTT 3′, IBS2: 5′ TTTAAT 3′) and thus located within the HEase ORF were not modified in order to maintain the splicing potential of the group IIA1 intron. The synthesized ORF sequence (1722 bp) was inserted into the pET28b (+) plasmid as a NheI/BamHI fragment. The vector provides the ORF with an N-terminal 6 x histidine-tag; this construct was named I-CthI-[IIA1]-pET28b (+).

In order to expand the concept of the "intron based on switch" as a potential regulatory element for HEase expression, a non-native group IIB intron (rI1) from *Scenedesmus obliquus* (GenBank accession number X17375.2; Kück *et al*., 1990) was inserted in a suitable position within the I-CthI ORF. The intron was inserted at a location that allowed for maintaining the IBS/EBS interactions with minimal modification of the HEase coding sequence; fortuitously for the rI1 intron only the IBS1 sequence is essential for splicing (Hollander and Kück, 1999). Prior to the insertion of this intron, a suitable region in the HEase ORF sequence was located, that matched the required IBS1 (*aac* coding for arginine and *agg* coding for asparagine) element for this group IIB intron. However, the presence of this sequence at the carboxyl terminal of the

HEase ORF made it less suitable. An alternative approach was undertaken. A sequence located further upstream, *cgcaac* encoding asparagine and arginine, was rearranged to *aaccgc* essentially introducing two conservative amino acid replacements. The *cgc* nucleotides were further modified to *agg* which did not change the amino acid composition of the protein (see Figure 4.1A). The codon-optimized ORF (1530 bp) including the intron was synthesized and inserted as a NdeI/BamHI fragment into the pET28b (+) plasmid (same as for the native intron).The construct was named I-CthI-[IIB]-pET28b (+). Both of the constructs were transformed into chemically competent *E. coli* BL21 (DE3) as described in section 2.5 for testing the *in vivo* group II intron splicing competency and for additional biochemical studies.

A substrate plasmid (Cth-rns.pUC57) was constructed in order to evaluate the activity of I-CthI (see Chapter 3, section 3.2.3; Guha and Hausner, 2014).

4.2.2. *In vivo* RNA splicing assay

Reverse Transcriptase PCR (RT-PCR) was employed to examine *in vivo* splicing intermediates for the HEase ORF group II introns. Here the HEase gene derived transcripts were studied to verify if splicing occurred and if splicing in *E. coli* maintained the expected intron/exon junctions. One hundred µL of chemical competent *E. coli* BL21 (DE3) were transformed with either I-CthI-[IIA1]-pET28b (+) or the empty pET28b (+) vector and inoculated in 10 mL of LB media supplemented with 100 µg/mL kanamycin and 0.25% (w/v) glucose and incubated overnight with agitation at 37 °C. Five hundred µL from the overnight culture was used to inoculate 50 mL of LB medium supplemented with kanamycin (kan) and glucose (as described above). Additionally, each of the culture flasks was supplemented with either 1 mM, 5 mM, 10 mM or 20 mM magnesium chloride ($MgCl_2$). Initially, exogenous $MgCl_2$

concentrations up to 100 mM were tested. Even though such high concentrations of $MgCl_2$ was not detrimental to the bacterial cells (as evident from checking $O.D._{600}$), there was no appreciable difference in the splicing activity as compared to the culture containing 20 mM $MgCl_2$. Therefore, $MgCl_2$ concentrations up to 20 mM were considered for further studies. A culture flask with no exogenously added $MgCl_2$ was used as the negative control. A second set of negative controls consisted of cultures in which: a) 10 μM of cobalt chloride ($CoCl_2$) was added to the LB media or b) the LB media was supplemented with both 10 μM $CoCl_2$ and 5 mM $MgCl_2$. The cultures were grown at 37 °C with agitation till the $O.D._{600}$ reached 0.65. Ten mL of the bacterial cells from each of the above cultures were centrifuged for 3 minutes at 7000 x g. The cells were lysed and RNA was extracted using the GENEzol TriRNA Pure kit (FroggaBio, North York, Ontario; catalog # GZX100) following the manufacturer's protocol. To ensure complete removal of any contaminating DNA, the RNA was treated with 2 units of DNaseI and incubated at 37 °C for 15 minutes; the reaction was stopped by adding 1 μL EDTA (50 mM) followed by 10 minute incubation at 65 °C. Furthermore, in order to confirm the elimination of the DNA, 2 μL of the reaction mixture was applied to perform a standard PCR reaction using the forward primer TwinHEG-F and the reverse primer TwinHEG-R (see Table 2.1; also see Figure 4.1B).

ThermoScript RT-PCR system was used to make cDNA from the RNA extracted from bacterial cells grown in the presence of various concentrations of $MgCl_2$ as described in section 2.17. In order to characterize the expression vector derived HEase transcripts several primers were designed to recover potential splicing intermediates. The forward primer TwinHEG-F and the reverse primer TwinHEG-R are based on the codon-optimized ORF upstream (290 nucleotides) and downstream sequences (248 nucleotides) with regards to the group IIA1 intron

insertion site (see Figure 1B). Primer‘C’ and primer‘D’ (see Table 2.1; see Chapter 3; Guha and Hausner, 2014) were used to amplify a 400 bp segment of the internal group IIA1 intron and this segment served as the positive control for detecting the presence of the group IIA1. In order to examine the splicing potential and splicing intermediates for constructs featuring the rI1 intron, the same protocol was applied as above. Since the location of the group IIB intron within the ORF is further upstream compared to the native intron, one additional forward primer TwinORFnew-F (see Table 2.1) was designed in order to capture the splicing products that involved the rl1 intron. The reverse primer TwinHEG-R remained the same for the group IIB intron splicing analysis. The RT-PCR amplicons obtained were excised from the agarose gel and purified using the Gel/PCR DNA fragments Extraction Kit and instructions from the manufacturer (FroggaBio, North York, Ontario; catalog # DF100). The gel extracted DNA fragments were sent to the MICB DNA sequencing facility (University of Manitoba, Cancer Care, McDermot Avenue, Winnipeg, Canada) for Sanger cycle sequencing utilizing the primers used for obtaining the amplicons (see Figure 4.1B).

4.2.3. *In vitro* and *in vivo* protein expression and purification

For both types of constructs I-CthI-[IIA1]-pET28b (+) and I-CthI-[IIB]-pET28b (+), protein expression was first evaluated with an *in vitro* translation assay. RNA extracted from *E.coli* BL21 cells grown under different Mg^{+2} concentrations were further subjected to *in vitro* translation as described in section 2.12 and analyzed for the presence of the desired protein at approximately 29 kDa.

For *in vivo* protein overexpression in *E. coli* and protein purification, the same protocol was applied for both types of constructs I-CthI-[IIA1]-pET28b (+) and I-CthI-[IIB]-pET28b (+).

In order to check for the expression of the HEase protein *in vivo*, the remaining 40 mL of culture (first 10 mL from each 50 mL culture were removed for RNA extraction to perform *in vivo* RNA splicing assay; see section 4.2.2) was induced with IPTG to a final concentration of 0.5 mM when the $O.D._{600}$ of the cells reached 0.65. The cultures were then shifted to 28 °C and incubated with agitation for 4 hours. Cells were resuspended in the lysis buffer as described in section 2.13 and sonicated in short pulses of 15 seconds using the Sonic Dismembrator model 300 (ThermoFisher Scientific). Eight μg of crude lysate from each of the induced samples were subjected to 12.5% SDS PAGE and the protein gel was stained as described in section 2.14. This gel was analyzed for the presence of the desired protein band. Furthermore, Ni-NTA super flow column was used to purify the protein following the methods described in section 2.13.

4.2.4. *In vitro* endonuclease assay

The I-CthI HEases as expressed from constructs containing either the native group IIA1 intron or the rI1 group IIB intron were evaluated for activity by performing *in vitro* endonuclease assay as described in section 2.15. Two non-substrate sequences were also challenged with the HEase preparations to evaluate the specificity and purity of the HEases extracts. A *C. thermophilum rns* segment containing the mS1247 twintron plus flanking exon sequences (GenBank accession number: JN007486.1) cloned in the pUC57 vector and a fragment of the cytochrome oxidase (*cox*) gene *from Annulohypoxylon stygium* (GenBank accession number: NC_023117.1) cloned into the pUC57 vector served as negative controls. These non-substrates (1 μg) were challenged with the same concentration of the HEase protein and incubated for two hours at 37 °C. In one set of reactions, in order to rule out the possibility that addition of $CoCl_2$ has any inhibitory effect on the I-CthI HEase's activity, 10 μM of $CoCl_2$ was added to the *in*

vitro endonuclease reaction buffer and the above protocol was followed. Moreover, to test the effect of the activity of the protein in the presence of only $CoCl_2$ (and not in combination with $MgCl_2$), the protein was incubated with the substrate while 10 μM of $CoCl_2$ was added to the endonuclease reaction buffer minus 10 mM $MgCl_2$.

4.2.5. Cleavage site mapping assay

The cleavage mapping assay was performed as described in section 2.16 (also see Figure 4.8).

4.2.6. Evaluating the role of $MgCl_2$ in stimulating HEase expression

In order to evaluate if intron splicing could be manipulated as a potential "on switch", the addition of exogenous Mg^{+2} was investigated with regards to expression of the HEases. An *in vivo* endonuclease assay was established to evaluate the expression of functional HEases at various Mg^{+2} concentrations. Here two compatible plasmids were maintained in *E.coli* BL21 (DE3) based on antibiotic selection [kanamycin (kan) and chloramphenicol (cam)]. The I-CthI-[IIA1]-pET28b (+) - kan (ColE1 origin of replication) construct (7.1 kb) allowed for the expression of the HEase ORF and a second plasmid with chloramphenicol and the appropriate HEase target site sequence served as the substrate plasmid. For constructing the substrate plasmid, the Cth-*rns*.pUC57 was digested with BamHI and XbaI and a 469 bp containing the *rns* segment with the HEase target site was cloned in the pACYC184 plasmid [ATCC 37033 (American Type Culture Collection, Manassas, VA, USA; p15A origin of replication].The substrate plasmid was named Cth-*rns*.pACYC184 - cam (4.6 kb). Excision of the intron would permit the expression of an active HEase from the I-CthI-[IIA1]-pET28b (+) - kan plasmid and

this could be detected by plating cells on various media with different antibiotics, ultimately the presence of the HEase should lead to the loss of the Cth-*rns*.pACYC184 containing the cam resistance marker (see Figure 4.2). Chloramphenicol irreversibly binds to the bacterial ribosome, particularly to a receptor site on the 50S subunit, inhibiting the peptidyl transferase which results in the prevention of amino acid transfer to the growing peptide chains, ultimately leading to inhibition of protein formation. Therefore, once the cam resistance marker is lost, the cells become susceptible to the antibiotic during the *in vivo* endonuclease activity.

For the *in vivo* endonuclease assay, cotransformed *E. coli* BL21 (DE3) cells were grown overnight in duplicates in culture tubes containing 5 mL LB media plus the appropriate antibiotics. One percent glucose was added to the media containing the HEase-cotransformed construct to prevent leaky expression from the T7 promoter. A 0.5 mL aliquot from the 5 mL overnight cultures was used to inoculate 50 mL LB broth cultures supplemented with 100 μg/mL kan, 60 μg/mL cam, 1% glucose and 5 mM $MgCl_2$. For convention, we refer to this culture flask as 'LB+Mg^{+2}'. Another culture flask designated 'LB' which contained no added $MgCl_2$ was inoculated with the same amount of overnight culture and this served as the negative control for this experiment. The cells were grown at 37 °C with vigorous shaking (210 rpm) and the cultures were either induced with 0.5 mM IPTG when the $O.D._{600}$ reached ~ 0.56 or not induced. The cultures were then shifted to 28 °C for the production of the HEase. After 4 hours, both the induced and uninduced cultures from 'LB+Mg^{+2}' and 'LB' were diluted to 10^{-6} and 100 μL of the diluted cultures were plated on LB agar plates containing 60 μg/mL cam (done in triplicate). Plates were incubated at 37 °C until the colonies were clearly visible; colonies were counted approximately after 18 hours of incubation. For this experiment, three technical and two biological replicates were performed. In addition, to ensure that proteins expressed by the empty

pET28b (+) vector (without HEase ORF intron containing construct) were not involved in the endonuclease activity, 50 ng of the vector was also cotransformed along with Cth-*rns*.pACYC184 into 100 µL of chemically competent *E.coli* BL21 (DE3) cells and the above protocol was followed. To evaluate the effect of 5 mM exogenous $MgCl_2$ on the *in vivo* splicing potential of the rI1 group IIB intron, the I-CthI-[IIB]-pET28b (+) - kanamycin construct was cotransformed with the substrate plasmid and the *in vivo* endonuclease protocol was performed as described above.

In order to antagonize the stimulatory effect of $MgCl_2$ on splicing of group II introns, 10 µM of cobaltous chloride ($CoCl_2$) was added to the culture media along with 5 mM $MgCl_2$. It has been previously shown that $CoCl_2$ perturbs the import of Mg^{+2} in *E. coli* cells (Nelson and Kennedy, 1971; Nelson and Kennedy, 1972; Truong *et al*., 2013). Moreover, in order to negate the possibility that $CoCl_2$ can promote splicing of the group IIA1 or group IIB, *E.coli* BL21 (DE3) cells containing each of the constructs I-CthI-[IIA1]-pET28b (+) and I-CthI-[IIB]-pET28b (+) were exposed to 10 µM of $CoCl_2$ in the culture media. The cultures containing both the salts ($MgCl_2$ and $CoCl_2$) as well as $CoCl_2$ alone were either uninduced or induced with 0.5 mM IPTG when the $O.D._{600}$ reached ~ 0.5. The cultures were further incubated at 28 °C for the production of the HEase. After 4 hours, the cultures were diluted to 10^{-6} and 100 µL of the diluted cultures were plated on each of the LB agar chloramphenicol selection plates (done in triplicate).The plate assays were performed as described above in order to evaluate the splicing of the internal group II introns in the presence of $CoCl_2$. For statistical analysis, unpaired student's t test was performed to determine the significance of the results obtained. Graphpad Prism 6.01 statistical analysis software was used to calculate the Student's t test and the respective bar graphs were drawn using the same software (see Figure 4.9).

4.3. Results

4.3.1. Exogenous Mg^{+2} induces *in vivo* splicing of group IIA1 and group IIB introns

To demonstrate the splicing competency of a group IIA1 and IIB intron in the presence of various added concentrations of exogenous Mg^{+2}, RNA was extracted from *E. coli* BL21 (DE3) cells containing the constructs I-CthI-[IIA1]-pET28b (+) or I-CthI-[IIB]-pET28b (+). The splicing reaction products for the native intron were detected with RT-PCR utilizing primers TwinHEG-F and TwinHEG-R. Among the observed cDNAs obtained from RNA extracted from bacterial cells grown in the presence of various concentrations, only cells grown at 5 mM and 10 mM $MgCl_2$ showed evidence of alternate products. A PCR product near the 500 bp marker, the expected size (538 bp) for cDNAs from transcripts where the group IIA1 intron spliced out, was further investigated (Figure 4.3A). Bacterial cells grown in either lower (0 mM, 1 mM) or higher (20 mM) $MgCl_2$ did not show any evidence for splicing and only the full length unspliced PCR product (1296 bp) was recovered. A 400 bp PCR product was observed when internal group IIA1 intron specific primers ('C' and 'D') were used and this served as the positive control showing the presence of the intron in all the samples examined. PCR amplicons derived from the unspliced and spliced cDNAs were gel excised and submitted for DNA sequence analysis and the resulting data were compared with the control non-spliced DNA template. Comparative sequence analysis showed that the 556 bp RT-PCR product was the result of the group IIA1 intron being spliced out and the joining of the flanking exon segments. Based on a previous study (Guha and Hausner, 2014) it was expected that the RT-PCR product obtained from a spliced transcript should be 538 bp in length. The additional nucleotides noted was due to a shift of the 5′ splice junction which was 18 nucleotides downstream from the original IBS1 and IBS2 elements. This is probably due to a cryptic/alternate splice site that was utilized in *E. coli*.

However, the 3′ splice site was found to be consistent with previous studies on this intron in its "native" mitochondrial environment (see Chapter 3; Guha and Hausner, 2014).

Similar results were obtained for demonstrating the splicing competency of the group IIB intron RNA extracted from bacterial cells grown in the presence of increasing concentrations of $MgCl_2$ in the LB media yielded a RT-PCR product of 658 bp representing the spliced version of the HEase transcript. However, no splicing was observed when the cells were grown in the absence or in the presence of 1 mM and 20 mM concentrations of external $MgCl_2$ (Figure 4.3B). Sequence analysis of the RT-PCR product revealed that the splicing of this group IIB intron occurred as predicted based on its IBS1 sequence (Kück *et al*., 1990) and the splicing followed the conventional intron/splice sites yielding a continuous I-CthI ORF.

To demonstrate the splicing competency of group IIA1 and IIB intron in the presence of $CoCl_2$ or both $MgCl_2$ and $CoCl_2$ in the growth media, RNA was extracted from *E.coli* BL21 cells containing the constructs I-CthI-[IIA1]-pET28b (+) or I-CthI-[IIB]-pET28b (+). RT-PCR results showed that the bacterial cells grown in either 10 μM of $CoCl_2$ or the addition of both 10 μM of $CoCl_2$ and 5 mM $MgCl_2$ in the growth media did not show any evidence for splicing and only the full length unspliced PCR product was recovered (Figure 4.3C).

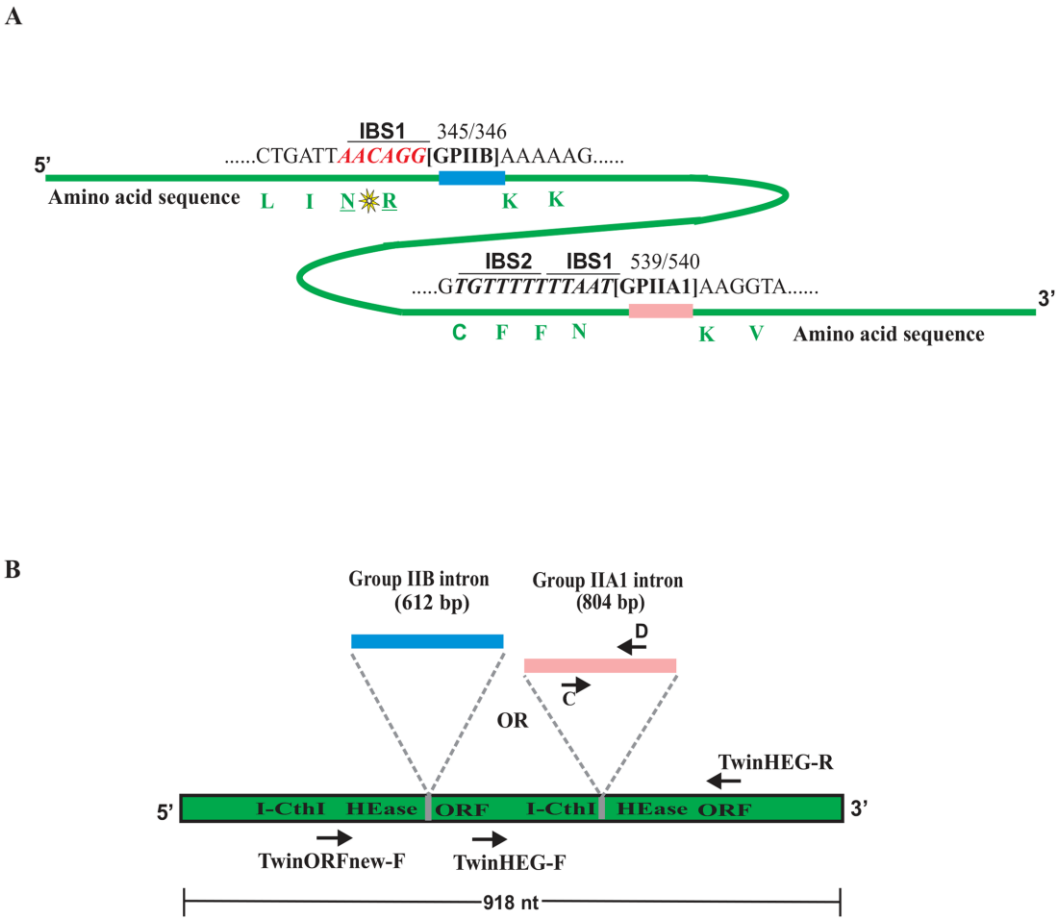


Figure 4.1. (**A**) Homing endonuclease ORF and location of introns. Schematic representation (not drawn to scale) for the location of the native group IIA1 intron (GPIIA1, shown in pink) and the non-native group IIB intron (GPIIB, shown in blue) within the I-CthI HEase ORF sequence (shown in green). Number 539 and 540 represent the exact location (insertion site) of the GPIIA1 in the ORF sequence. The IBS1 and the IBS2 elements are in italics (underlined), both are located upstream from the intron insertion site. The corresponding amino acids (marked in green) have been indicated to their respective codons. In another construct, GPIIB was inserted in the I-CthI HEase ORF sequence at position 345/346 that allowed for maintaining the IBS/EBS interactions with minimal modification of the HEase coding sequence (see text for details). The conservative amino acid substitution (indicated with a yellow star) between Arginine (R) and Asparagine (N) residues has been introduced to form the correct IBS1 (marked in red, italics and underlined) for supporting GPIIB splicing. The corresponding amino acids (marked in green) have been indicated to their respective codons. (**B**) Overview of primer location for RT-PCR. Diagram (not drawn to scale) showing the relative location of the RT-PCR primers within I-CthI HEase ORF sequence utilized for detecting splicing products during the *in vivo* RNA splicing assay (Guha and Hausner, 2016).

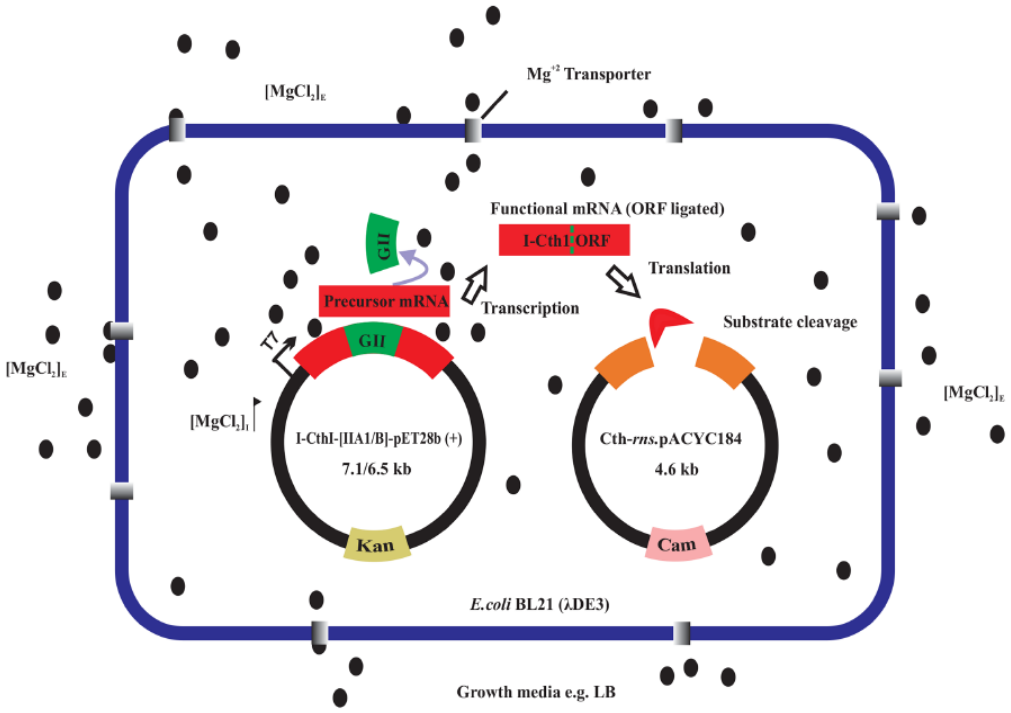


Figure 4.2. Impact of $MgCl_2$ on splicing and the expression of a homing endonuclease. An *in vivo* endonuclease assay was established where two compatible plasmids were maintained in *E. coli* BL21 (DE3) based on antibiotic selection [kanamycin (kan) and chloramphenicol (cam)]. Cells were grown in absence and presence of either 5 mM or 10 mM added $MgCl_2$ (E = external) and induced with 0.5 mM IPTG ($O.D._{600}$ = 0.56). The internal concentration (I) of $MgCl_2$ increases probably due to the activities of magnesium transporters. Free Mg^{+2} ions are available to bind to the catalytic center of the group II intron [domain V (not shown)] and initiate efficient splicing and religation of the ORF. The expressed HEase cleaves the target site in the substrate plasmid resulting in the loss of cam resistance marker. Cells grown in the absence of $MgCl_2$ fail to splice out the internal group II thus yielding non-functional HEase thereby resulting in the maintenance of the substrate plasmid. The *E. coli* genome is not shown for simplicity (Guha and Hausner, 2016).

4.3.2. The alternate splice site for the group IIA1 does not affect I-CthI functionality

Sequence analysis of the cDNA (556 bp) from the group IIA1 intron construct derived transcripts had revealed a cryptic splice site which is 18 nucleotides downstream of the original (native) IBS sequences. The alternate IBS sequences can potentially H-bond with sequences that are near or overlap with the native (original) EBS sequences (Figure 4.4A). In order to evaluate if the addition of the extra six amino acids could affect the active site or the protein's tertiary structure, the online program Protein Homology/analogY Recognition Engine V2.0 (PHYRE2) (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id+index) (Kelley *et al*., 2015) was used to map the position of these newly added amino acids onto the predicted tertiary structure of I-CthI. The program showed that the protein is composed of ten alpha helices (39%) and nine beta stands (26%), the confidence key of these regions were found to be 100% when compared to the crystal structure of the HEase I-SmaMI (PDB accession number: c4loxA). The additional six amino acids V, R, R, C, G and Y are located in a linker region of the HEase protein and did not appear to disrupt the active sites (LAGLIDADG motifs) or beta sheets required for making contact with the DNA target site (Figure 4.4B). The HEase protein derived from the group IIA1 intron containing construct was purified and challenged with its substrate. The enzyme completely linearized the circular substrate plasmid (3.1 kb) within 90 minutes. However, when this protein was challenged with two different non-substrates to test its specificity, even after two hours of incubation at 37 °C, the protein failed to cleave the non substrates (Figure 4.4C). It was also noted that the addition of 10 µM $CoCl_2$ in the *in vitro* endonuclease reaction buffer did not inhibit the functionality of the I-CthI HEase and this protein could cleave its substrate in 90 minutes at 37 °C. However, when the substrate was incubated just in the presence of 10 µM

$CoCl_2$ without $MgCl_2$ in the endonuclease reaction buffer, the protein did not initiate any cleavage activity (Figure 4.5).

4.3.3. *In vitro* and *in vivo* translation show evidence of HEase protein production under specific magnesium concentration

SDS PAGE analysis showed the presence of the protein I-CthI at the desired location (~ 29 kDa) only from *in vitro* translation assays that used RNA extracted from cells grown in the culture media supplemented with either 5 mM or 10 mM $MgCl_2$. RNA extracted from cells grown in the absence or at 1 mM and 20 mM $MgCl_2$ failed to yield the desired protein in the *in vitro* translation assays. This was observed for both of the tested constructs I-CthI-[IIA1]-pET28b (+) and I-CthI-[IIB]-pET28b (+) when the cells were grown at the same concentrations of $MgCl_2$ (Figure 4.6A).

The codon-optimized HEase ORF (from both constructs) expressed in *E.coli* BL21 cells only when the cells were grown in the culture media supplemented with 5 mM or 10 mM $MgCl_2$ (Figure 4.6B). The recovery and purification of the HEase protein was achieved by affinity column chromatography according to the method detailed in section 2.13. This purified protein was pooled and concentrated (3 mg/mL) in order to perform the *in vitro* endonuclease assay.

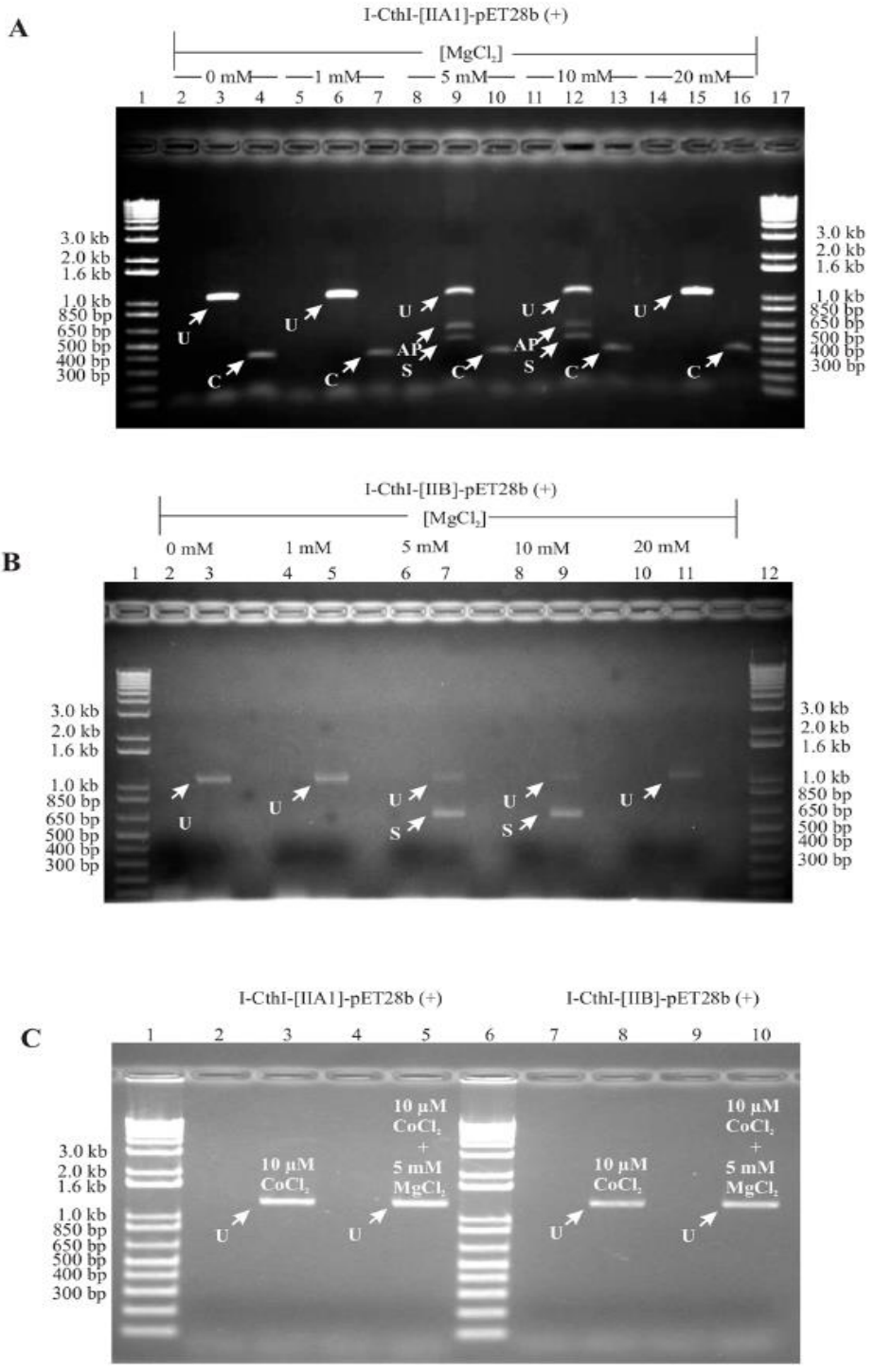

A
I-CthI-[IIA1]-pET28b (+)
[MgCl2]
0 mM
1 mM
5 mM
10 mM
20 mM
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
3.0 kb
2.0 kb
1.6 kb
1.0 kb
850 bp
650 bp
500 bp
400 bp
300 bp
U
C
AP
S
B
I-CthI-[IIB]-pET28b (+)
[MgCl2]
0 mM
1 mM
5 mM
10 mM
20 mM
1 2 3 4 5 6 7 8 9 10 11 12
C
I-CthI-[IIA1]-pET28b (+)
I-CthI-[IIB]-pET28b (+)
1 2 3 4 5 6 7 8 9 10
10 µM CoCl2
10 µM CoCl2 + 5 mM MgCl2

Figure 4.3. (**A**) A mtDNA group IIA1 intron can splice in *E. coli*. A 1% agarose gel showing the RT-PCR results (Primers: TwinHEG-F/R) for *in vivo* group IIA1 intron splicing under various concentrations of external $MgCl_2$ in the culture media. Lanes 2, 5, 8, 11, 14 show the results for standard PCR (TwinHEG-F/R) on RNA samples testing for the absence of genomic DNA prior to cDNA synthesis. Lane 3, 6 (see white arrows) and 15 show the PCR product (~1.1 kb) that represents the unsliced transcript (U) in the absence (0 mM), presence of low (1 mM) or high (20 mM) concentration of $MgCl_2$ in the culture media respectively. Lanes 9 and 12 (see white arrows) show the PCR product (556 bp) that indicates splicing occurred (S) and possible alternative products (AP) are also indicated in lanes. Lanes 4, 7, 10, 13 and 16 (see white arrows) are positive controls (C; Primers C and D) that show the presence of the internal group IIA1 (400 bp) in all the samples examined. Lane 1 and 17 contain the 1kb plus^{TM} DNA ladder. (**B**) A mtDNA group IIB intron can splice in *E. coli*. A 1% agarose gel showing the RT-PCR results (Primers: TwinORFnew-F/TwinHEG-R) for *in vivo* group IIB intron splicing under various concentrations of external $MgCl_2$ in the culture media. Lanes 2, 4, 6, 8 and 10 (see white arrows) show the results for standard PCR (Primers: TwinORFnew-F/TwinHEG-R) on RNA samples testing for the absence of genomic DNA prior to cDNA synthesis. Lanes 3, 5 and 11 (see white arrows) show the PCR product (~1.2 kb) that represent unspliced transcripts (U) in the absence (0 mM), presence of low (1 mM) or high (20 mM) concentration of $MgCl_2$ in the culture media respectively. Lanes 7 and 9 (see white arrows) show a PCR product that represents spliced transcripts (658 bp) in the presence of 5 mM and 10 mM $MgCl_2$ in the culture media. Lane 1 and 12 contain 1 kb plus^{TM} DNA ladder. **(C)** Effect of $CoCl_2$ and/or $MgCl_2$ on intron splicing. A 1% agarose gel showing the RT-PCR results for *in vivo* splicing of group IIA1 intron and group IIB intron when 10 µM $CoCl_2$ alone or 10 µM $CoCl_2$ in combination with 5 mM $MgCl_2$ were added to the LB growth media. Lanes 2 and 4 show the results for standard PCR on RNA samples testing for the absence of genomic DNA prior to cDNA synthesis. Lane 3 and 5 show that splicing cannot be detected (see band marked with U) when I-CthI-[IIA1]-pET28b (+) [BL21] was grown with the addition of 10 µM $CoCl_2$ in one LB culture media and combination of 10 µM $CoCl_2$ and 5 mM $MgCl_2$ in the other LB media respectively. For the second construct I-CthI-[IIB]-pET28b (+) [BL21], lanes 7 and 9 represent the results for standard PCR on RNA samples testing for the absence of genomic DNA prior to cDNA synthesis. Lanes 8 and 10 that splicing cannot be detected (see band marked with U) when I-CthI-[IIB]-pET28b (+) [BL21] was grown with the addition of 10 µM $CoCl_2$ in one LB culture media and combination of 10 µM $CoCl_2$ and 5 mM $MgCl_2$ in the other LB media respectively. In lanes where splicing occurred the RT-PCR products are labelled with S. Lanes 1 and 6 represent the 1kb plus^{TM} DNA ladder (Guha and Hausner, 2016).

4.3.4. I-CthI ORF interrupted with either a group IIA1 or IIB introns results in the expression of an active HEase

The primary objective of this work was to evaluate if group II introns can be utilized as regulatory elements for expressing a mtDNA fungal HEase in *E. coli*. Two types of group II introns were utilized, group IIA1 and IIB introns. As previously mentioned *in vitro* endonuclease assays showed that functional HEase were expressed in *E. coli* under conditions that favour group II intron splicing. It appears that both types of mitochondrial group II introns can splice in *E. coli* under suitable Mg^{+2} concentrations ultimately yielding HEase that can linearize their substrates in one hour at 37 °C; and non-substrates were not cleaved even after two hours of incubation indicating that the enzyme is still highly specific for its cleavage site (**Figure 4.7**).

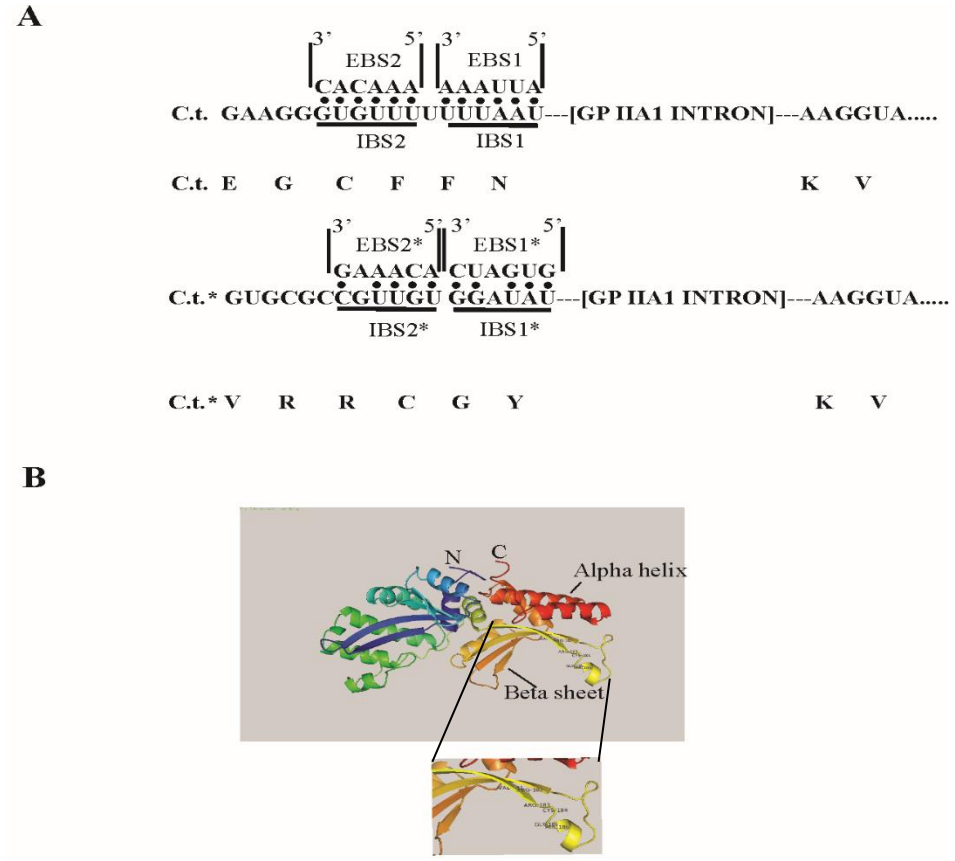

A
3' EBS2 5'
CACAAA
3' EBS1 5'
AAAUUA
C.t. GAAGGGUGUUUUUUUUAAU---[GP IIA1 INTRON]---AAGGUA.....
IBS2
IBS1
C.t. E G C F F N K V
3' EBS2* 5'
GAAACA
3' EBS1* 5'
CUAGUG
C.t.* GUGCGCCGUUGU GGAUAU---[GP IIA1 INTRON]---AAGGUA.....
IBS2*
IBS1*
C.t.* V R R C G Y K V
B
N
C
Alpha helix
Beta sheet
C
mins
M C L 0 30 60 90 120 M
5.0 kb
4.0 kb
3.0 kb
2.0 kb
1.6 kb
1.0 kb
650 bp

Figure 4.4. (**A**) Intron and exon binding sites for the mS1247 nested group IIA1 intron. Watson-Crick base pairing (shown by solid black dots) between the newly discovered cryptic (marked by asterisk sign) splice site sequence (IBS1* and IBS2*) and corresponding exon binding sequences (EBS1* and EBS2*) of the mS1247 internal group IIA1 intron. The original IBS1, IBS2 and EBS1, EBS2 for mS1247 twintron from *C. thermophilum* are indicated. (**B**) An *in silico* model for the expressed I-CthI protein. An *in silico* model for the I-CthI protein derived from the I-CthI-[IIA1]-pET28b (+) construct generated by the PHYRE2 program. The program identified the double motif LAGLIDADG I-SmaMI (PDB: c4loxA) HEase protein as a template for folding I-CthI. Alpha helices and beta sheets along with amino terminal (N) and carboxyl terminal (C) have been marked. The LAGLIDADG motifs contribute towards the active site of the enzyme while the beta sheets arrange in a configuration that forms the DNA binding surface. The extra six amino acids (V, R, R, C, G and Y) were not present in any of the active sites of the HEase instead they are located in a linker region between the two beta sheets near the carboxyl terminal of the protein. The linker region showing the extra six amino acids has been magnified for better illustration. The amino acid positions are also mentioned. (**C**) *In vitro* endonuclease assay for I-CthI. A 1% agarose gel showing the *in vitro* endonuclease assay with *C. thermophilum* HEase ORF intron containing construct I-CthI-[IIA1]-pET28b (+). Lane C and L represent uncut control substrate plasmid and linearized substrate plasmid (cleaved with BamHI) respectively. Numbers on the top of each lane represent incubation time in minutes at 37 °C. For each of the above endonuclease assays, 1 μg of substrate DNA was treated with 24 μg of the purified HEase (3 mg/mL). The arrow shows the linearized band at 3.1 kb when the substrate was incubated for 90 minutes. C_{ns} represents the negative control plasmid while C_{ns} + E represents the negative control plasmid incubated with the same concentration of the HEase for 120 minutes at 37 °C. Another negative control plasmid (Cox_s) was used to examine the specificity of this twintron encoded I-CthI. Cox_s is the substrate for another HEase encoded from the intron of the *cox* gene from *Annulohypoxylon stygium* while Cox_s + E represent the same plasmid incubated with purified I-CthI (encoded from the above construct) for 120 minutes at 37 °C. For both the negative controls, no endonuclease activities were observed. Lane denoted with M represents the 1 kb plus^{TM} DNA ladder (Guha and Hausner, 2016).

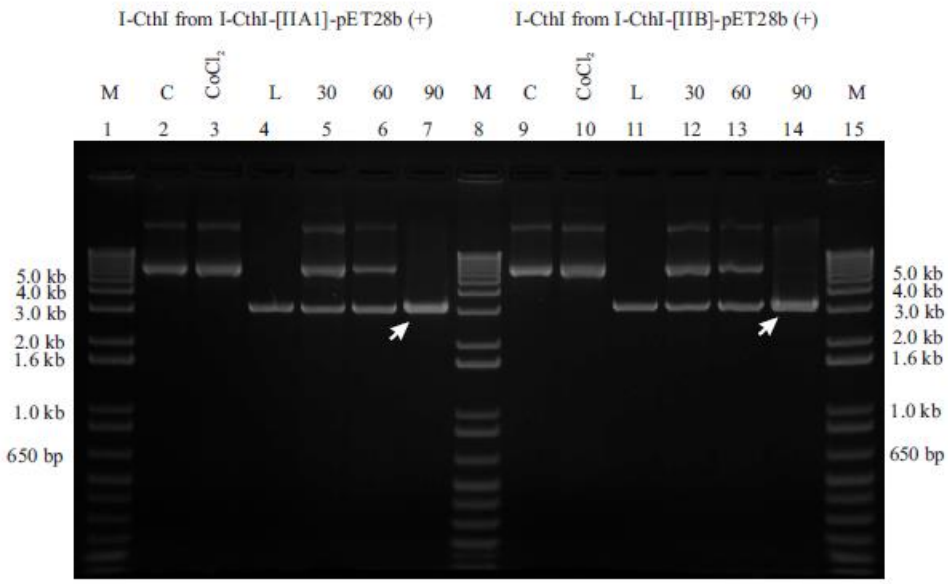


Figure 4.5. $CoCl_2$ does not affect I-CthI endonuclease activity. A 1% agarose gel showing the effect of addition of 10 µM $CoCl_2$ during the *in vitro* endonuclease assay with construct I-CthI-[IIA1]-pET28b (+) and I-CthI-[IIB]-pET28b (+) encoded I-CthI HEase. Lane 2 represents the control uncut substrate (C) plasmid. Lane 3 shows the endonuclease activity of I-CthI on the substrate plasmid in the presence of 10 µM $CoCl_2$ alone in the endonuclease reaction buffer without 10 mM $MgCl_2$. The reaction incubation time is 60 minutes. Lane 4 represents the linearized substrate (L) when treated with BamHI. Numbers on the top of the lanes (30, 60, 90) represent incubation time in minutes at 37 °C. The white arrow shows the linearized band at 3.1 kb. The same order (i.e. lanes 9 through 14) was maintained for the endonuclease activity of the I-CthI HEase derived from I-CthI-[IIB]-pET28b (+) [BL21] construct. Lanes 1, 8 and 15 contain the 1 kb plus^{TM} DNA ladder (Guha and Hausner, 2016).

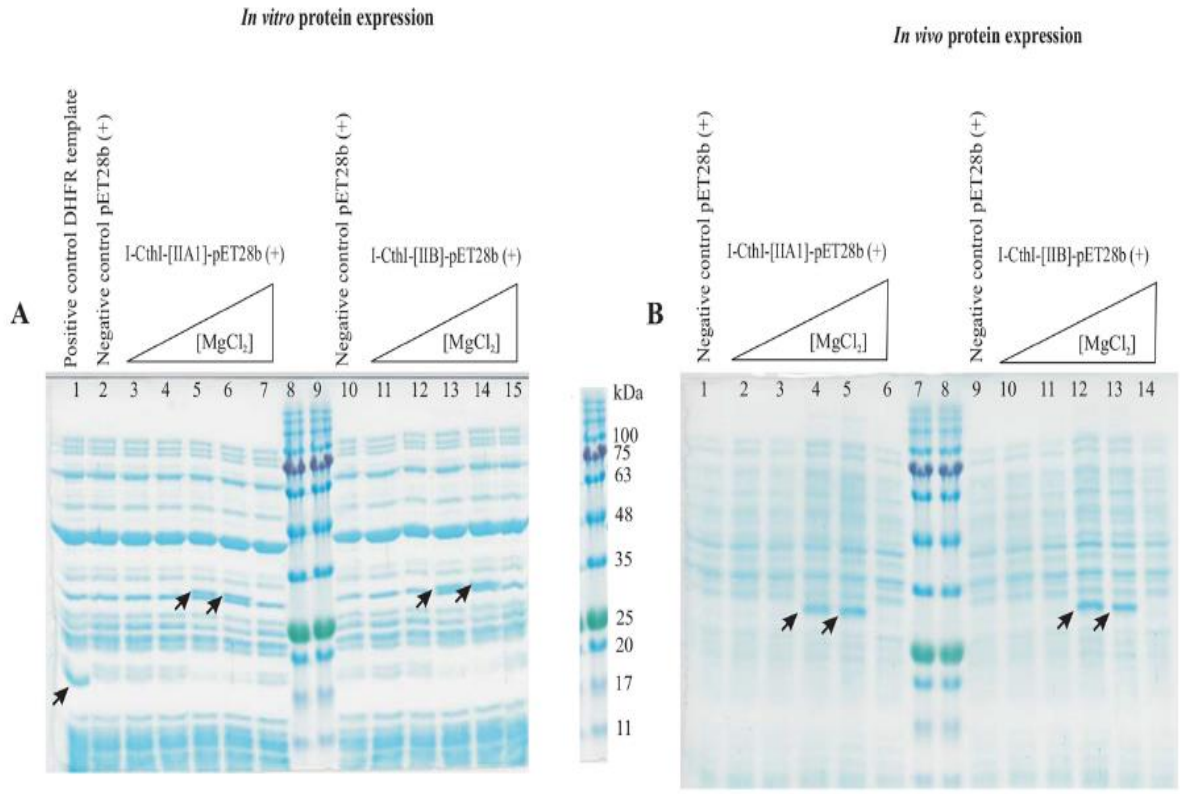


Figure 4.6. (**A**) The effect of $MgCl_2$ on *in vitro* protein expression. A 12.5% SDS-PAGE (stained with Coomassie Brilliant Blue) showing *in vitro* protein expression for constructs I-CthI-[IIA1]-pET28b (+) [left] and I-CthI-[IIB]-pET28b (+) [right] in the presence of various concentrations of external $MgCl_2$ in the culture media. Lane 1 represents the *E.coli* dihydrofolate reductase (marked with arrow) when 125 ng/µL was used as the template (positive control) for the PURExpress *In Vitro* Protein Synthesis kit. Lanes 2 and 10 show the *in vitro* protein expression profiles when empty pET28b (+) vectors (without the above constructs) were used as the negative control. Lanes 3 and 11 represent the *in vitro* protein expression profile when RNA (extracted from the culture in the absence of $MgCl_2$) was used as the template. Lanes 4 through 7 represent the protein expression profiles when RNA (extracted from the cultures in the presence of 1 mM, 5 mM, 10 mM and 20 mM respectively) was used as the template for the *in vitro* protein synthesis. The expression of the protein (I-CthI) has been marked with arrows. For *in vitro* expression from the I-CthI-[IIB]-pET28b (+) construct, lanes 12 through 15 follow the same order as depicted for the I-CthI-[IIA1]-pET28b (+) construct (i.e. lanes 4-7). Lanes 8 and 9 represent the Blueye prestained protein ladder. (**B**) The effect of $MgCl_2$ on *in vivo* protein expression. A 12.5% SDS-PAGE showing *in vivo* protein expression for constructs I-CthI-[IIA1]-pET28b (+) [left] and I-CthI-[IIB]-pET28b (+) [right] in the presence of various concentrations of external $MgCl_2$ in the culture media. Lanes 1 and 9 represent the *in vivo*

protein expression profiles from the empty pET28b (+) vector (without the constructs). Lanes 2 through 6 represent the protein expression profiles when I-CthI-[IIA1]-pET28b (+) [BL21] was grown under increasing concentrations of external $MgCl_2$ starting from 0 mM, 1 mM, 5 mM, 10 mM and 20 mM. Lane 10 through 14 represent the protein expression profiles when I-CthI-[IIB]-pET28b (+) (BL21) was grown under increasing concentrations of external $MgCl_2$. Lanes 10 through 14 follow the same order as for the protein expression profiles when I-CthI-[IIA1]-pET28b (+) [BL21] was grown under increasing concentrations of external $MgCl_2$ (i.e. lanes 2-6). The overexpressed I-CthI (migrate at ~29 kDa) has been marked with arrows. Lanes 7 and 8 represent the Blueye prestained protein ladder (Guha and Hausner, 2016).

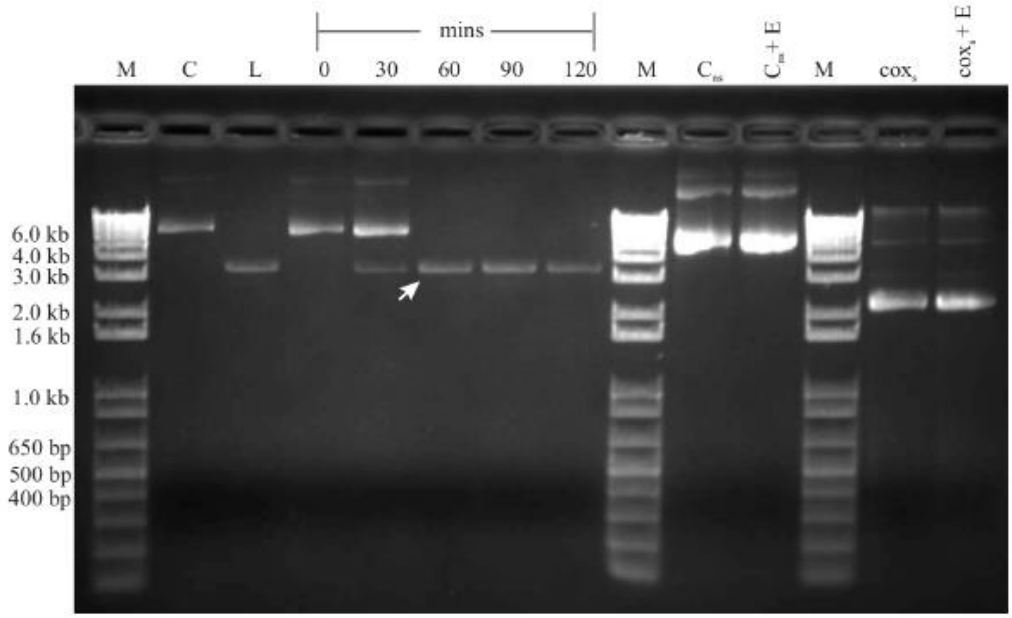
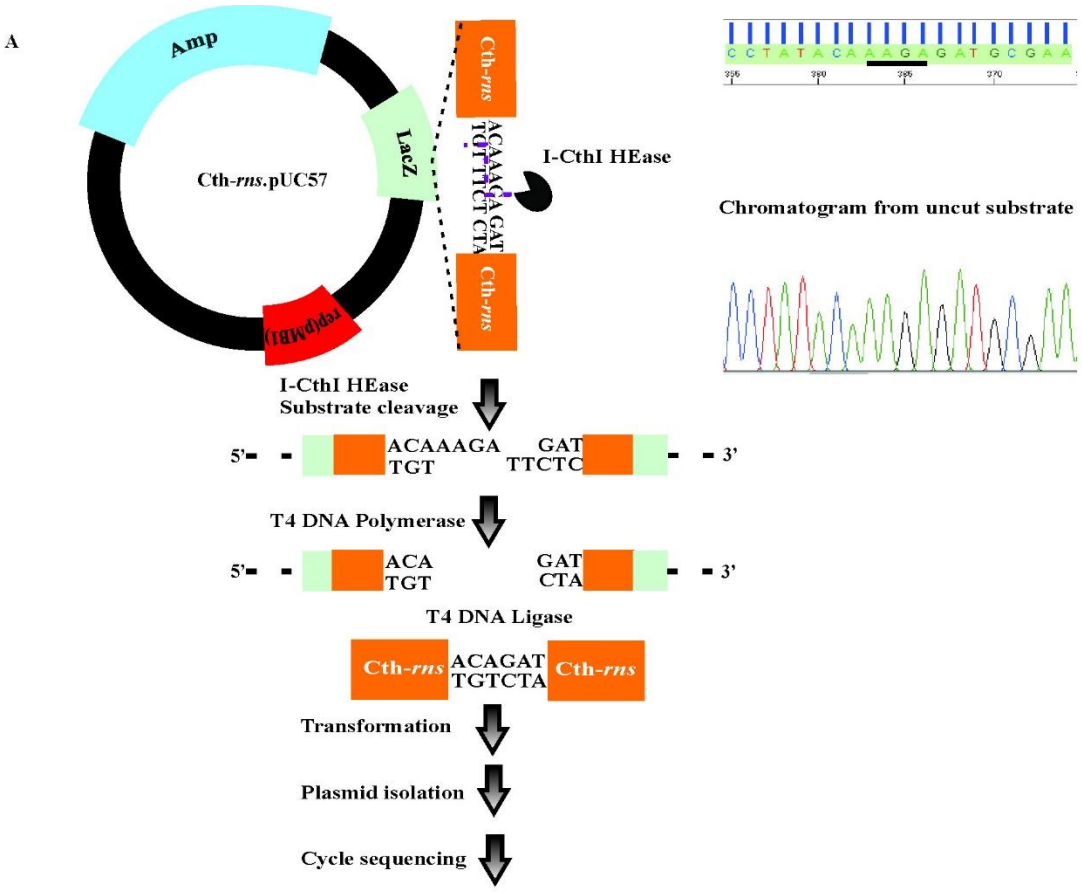
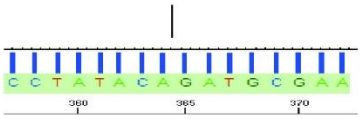


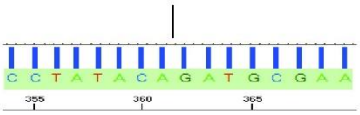
Figure 4.7. *In vitro* endonuclease assay. A 1% agarose gel showing the *in vitro* endonuclease assay with construct I-CthI-[IIB]-pET28b (+) encoded HEase. Lanes C and L represent uncut control substrate plasmid and linearized substrate plasmid (cleaved with BamHI). Numbers on the top of each lane represent incubation time in minutes at 37 °C. For each of the above endonuclease assay, 1 µg of substrate DNA was treated with 8 µL of the purified HEase (2.5 mg/mL). The arrow shows the linearized band at 3.1 kb when the substrate was incubated for 60 minutes. Two negative controls were applied in this assay. C_{ns} and Cox_s were used to examine the specificity of this twintron encoded I-CthI. Cox_s is the substrate for another HEase encoded a *cox* gene intron from *Annulohypoxylon stygium* while Cox_s + E represents the assay with the same plasmid incubated with purified I-CthI (encoded from the above stated construct) for 120 minutes at 37 °C. C_{ns} represents the negative control plasmid previously used (Guha and Hausner, 2014) in characterizing the I-CthI HEase; this pUC57 based construct contains the *rns* gene plus the mS1247 intron. C_{ns} + E represents the negative control plasmid incubated with the same concentration of the HEase for 120 minutes at 37 °C. For both the negative controls, no endonuclease activities were observed. Lanes marked with M represent the 1 kb plus™ DNA ladder (Guha and Hausner, 2016).

4.3.5. Endonuclease cleavage mapping of HEases derived from ORFs interrupted by group II introns shows cleavage sites have not changed

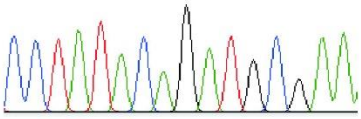
LAGLIDADG HEases tend to generate cohesive termini by generating staggered cuts with four nucleotide 3′-OH single stranded overhangs. The T4 DNA polymerase treated and religated I-CthI cleaved substrate plasmid sequences when compared with the sequence of the uncut substrates showed that a 5′-AAGA-3′ segment was removed from the sense strand. The endonuclease cleavage mapping data is shown in Figure 4.8A. So the cleavage mapping site experiments showed that the intron contained HEases ORFs (containing either the group IIA1 or IIB intron) ultimately allowed for the expression of a functional I-CthI protein that cleaves 8 bp downstream of the mS1247 nested intron insertion site (sense strand) or 4 bp downstream of position S1247 at the antisense strand (Figure 4.8B). These results are in agreement with previous experiments utilizing I-CthI constructs that did not contain introns within the HEase ORF (Guha and Hausner, 2014); so the presence of the introns and subsequent RNA processing events in *E. coli* did not alter the target site specificity of the HEase.



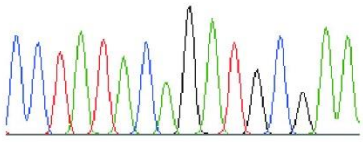




Chromatogram from I-CthI HEase (IIA1 construct) treated and religated substrate



Chromatogram from I-CthI HEase (IIB construct) treated and religated substrate



B

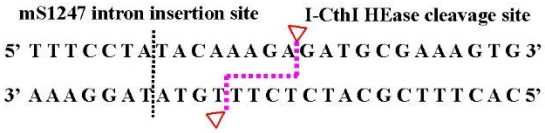


Figure 4.8. (**A**) Endonuclease cleavage mapping for HEases derived from ORFs interrupted by group II introns. The cleavage site was mapped by comparing uncut substrate with I-CthI treated substrate DNAs. Cleavage by I-CthI generates a staggered cut with a 4 nucleotide 3′-OH overhang in the substrate plasmid at the enzyme's target site. T4 DNA polymerase was used to blunt the cleaved ends. The religated plasmid was sequenced and compared to the sequence of the untreated substrate plasmid in order to map the cleavage site by scanning for a 4 bp deletion in the T4 DNA polymerase treated cleaved substrate plasmid. (**B**) Schematic representation of the I-CthI cleavage site near the twintron (nested intron) insertion sequence. Proposed cleavage sites are indicated by open triangles; and a vertical line represents the twintron (nested intron) insertion site. The HEase cleavage site is 8-nt downstream of the twintron (nested intron) insertion site with regards to the sense strand or 4-nt downstream with regards to the antisense strand (Guha and Hausner, 2016).

4.3.6. *In vivo* endonuclease assays for HEase activity in the presence of $MgCl_2$ and/or $CoCl_2$

In vivo endonuclease assays were performed to evaluate the effect of the addition of either $MgCl_2$ and/or $CoCl_2$ on the expression and functionality of the I-CthI HEase. The results (in cfu/mL) of the *in vivo* endonuclease assays for HEase activity are depicted in Tables 4.1, 4.2, and 4.3. Since three technical and two biological replicates (i.e. six independent values) were performed for each of the assay plates, the mean value for the bacterial colony forming units (cfu/mL) are provided along with their respective standard deviations (σ).

First, in order to check whether the protein expressed from the I-CthI-[IIA1]-pET28 b (+) construct is functional (i.e., can it cleave) or is toxic to the *E.coli* BL21 genome, I-CthI-[IIA1]-pET28b (+) [BL21] was either uninduced or induced with 0.5 mM IPTG in the presence of 5 mM exogenously added $MgCl_2$ (Table 4.1, left panel). When the uninduced culture was plated on kan plates, the bacterial colony count was calculated 3.2×10^{10} cfu/mL, $\sigma = 1.8 \times 10^9$ and for the induced culture, the count was 3.0×10^{10} cfu/mL, $\sigma = 2.0 \times 10^9$. Moreover, a bacterial lawn was observed in the absence of the antibiotic. These data showed that I-CthI is not toxic to *E. coli* and it does not appear to have any target specificity within the *E.coli* genome. The plate assay results from the negative control pET28b (+) vector cotransformed with Cth-*rns*.pACYC184 substrate are also listed (Table 4.1, right panel). The results showed that even in the absence or in the presence of 0.5 mM IPTG, when the cells were plated on cam plates, the bacterial colony count was 4.4×10^{10} cfu/mL, $\sigma = 2.2 \times 10^9$ and 4.2×10^{10} cfu/mL, $\sigma = 1.8 \times 10^9$ respectively. Bacterial lawn was also observed when no antibiotic was applied. These data suggested that the proteins encoded from the empty pET28b (+) vector were not detrimental to the substrate plasmid carrying the cam resistance marker.

The plate assay results from the *E. coli* cells cotransformed with I-CthI-[IIA1]-pET28b

(+) and Cth-*rns*.pACYC184 and grown in the absence (LB, left panel) or presence of 5 mM exogenously added $MgCl_2$ ($LB+Mg^{+2}$, right panel) are depicted in Table 4.2. The results in the left panel show that when the growth media had no exogenously added $MgCl_2$, even with or without induction with 0.5 mM IPTG, colonies were observed when plated on LB agar-cam plates. The colony count for the uninduced and the induced cultures were 3.0×10^{10} cfu/mL, $\sigma = 1.3 \times 10^9$ and 2.9×10^{10} cfu/mL, $\sigma = 2.0 \times 10^9$ (Table 4.2) respectively indicating that probably absence/inadequate Mg^{+2} concentration in the growth media (irrespective of IPTG induction) did not allow the *in vivo* excision of the internal group II intron thereby not yielding functional HEase. In contrast, the presence of 0.5 mM IPTG and in the presence of 5 mM $MgCl_2$ in the growth media (right panel) allowed for the expression of a functional HEase which probably happened due to the splicing out of the internal intron. This functional protein could have cleaved the target site, thereby degrading the cam resistance substrate plasmid. The bacterial colony count was 2.3×10^9 cfu/mL, $\sigma = 1.3 \times 10^9$ (Table 4.2). It is worthwhile to mention that there is an approximately 12.6 fold decrease in the cfu/mL (cells which were induced and grown in exogenous $MgCl_2$) when compared to cfu/mL of the cells grown under inducible conditions but in the absence of exogenous $MgCl_2$. Student's t test performed on the cfu/mL indicated significant difference (decrease; P value < 0.0001) when compared to the cells grown under inducible conditions but in the absence of exogenous $MgCl_2$ (Figure 4.9A).

The results for $CoCl_2$ acting as a possible antagonist for the uptake of $MgCl_2$ are provided in Table 4.3. The left panel of the table shows the results of the *in vivo* endonuclease assay in the presence of only 10 µM of $CoCl_2$ in the culture media. This set of assays was performed to rule out the possibility that $CoCl_2$ was involved in splicing of the internal intron. Even with or without induction with 0.5 mM IPTG, bacterial colonies were observed when

plated on LB agar-cam plates. The colony count for the uninduced and the induced cultures were 2.2 x 10^{10} cfu/mL, σ = 0.9 x 10^9 and 2.5 x 10^{10} cfu/mL σ = 1.2 x 10^9 respectively indicating that $CoCl_2$ was not involved in facilitating the excision of the group II introns thus yielding a non-functional protein. The right panel shows the results of the *in vivo* endonuclease assay in the presence of both 10 µM of $CoCl_2$ and 5 mM $MgCl_2$ in the culture media. Indeed bacterial cells with colony count of 2.8 x10^{10} cfu/mL, σ = 1.3 x 10^9 was observed (Table 4.3). Student's t test performed on the cfu/mL indicated significant difference (increase; P value <0.0001) when the cells were grown with the addition of both 5 mM $MgCl_2$ and 10 µM $CoCl_2$ compared to the cells grown with the addition of only 5 mM $MgCl_2$ under inducible conditions (Figure 4.9A). Assuming that $CoCl_2$ antagonizes $MgCl_2$ uptake one would observe that under both inductive and non-inductive conditions, bacterial colonies would be visible when plated on LB cam plates as the substrate plasmid carrying the antibiotic resistance marker (cam) was not targeted hence maintained.

The plate assay results from the second cotransformed construct I-CthI-[IIB]-pET28b (+) and Cth-*rns*.pACYC184 are provided in Supplemental Tables S7.1 and S7.2 (see Chapter 7: Appendices). For this construct, 19 fold decrease in the cfu/mL (cells which were induced and grown in exogenous $MgCl_2$) was observed when compared to cfu/mL of the cells grown under inducible conditions but in the absence of exogenous $MgCl_2$. Student's t test performed on the cfu/mL indicated significant difference (decrease; P value < 0.0001) when the cells were grown under inducible conditions and in the presence of exogenous $MgCl_2$ (Figure 4.9B). Student's t test performed on the cfu/mL indicated significant difference (increase; P value <0.0001) when the cells were grown with the addition of both 5 mM $MgCl_2$ and 10µM $CoCl_2$ compared to the cells grown with the addition of only 5 mM $MgCl_2$ under inducible conditions (Figure 4.9B).

Plate assay (two biological and three technical replicates)	**I-CthI-[IIA1]-pET28b (+) [BL21]**	**pET28b (+) + Cth-*rns*.pACYC184 [BL21]**
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	Not applicable	3.0×10^{10} cfu/mL $\sigma = 0.8 \times 10^{9}$
Plate 'C' No induction (cam) 5 mM $MgCl_2$	3.2×10^{10} cfu/mL $\sigma = 1.8 \times 10^{9}$	4.4×10^{10} cfu/mL $\sigma = 2.2 \times 10^{9}$
Plate 'D' 0.5 mM IPTG (cam) 5 mM $MgCl_2$	3.0×10^{10} cfu/mL $\sigma = 2.0 \times 10^{9}$	4.2×10^{10} cfu/mL $\sigma = 1.8 \times 10^{9}$

Table 4.1. *In vivo* activity of I-CthI expressed from I-CthI-[IIA1]-pET28b (+). *In vivo* endonuclease activity of the HEase expressed from the I-CthI-[IIA1]-pET28b (+) [BL21] construct or pET28b (+) empty vector which are challenged with the substrate plasmid Cth-rns.pACYC184 [BL21]; results reported cfu/mL. Three technical and two biological replicates were performed for each of the constructs. The numbers represent the mean of six independent cfu/mL. Standard deviations are also indicated for each of the above observations.

	0 mM $MgCl_2$ in LB media	**5 mM $MgCl_2$ in LB media**
Plate assay (two biological and three technical replicates)	**I-CthI-[IIA1]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]**	**I-CthI-[IIA1]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]**
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	3.3×10^{10} cfu/mL $\sigma = 3.7 \times 10^{9}$	3.1×10^{10} cfu/mL $\sigma = 1.4 \times 10^{9}$
Plate 'C' No induction (cam)	3.0×10^{10} cfu/mL $\sigma = 1.3 \times 10^{9}$	2.8×10^{10} cfu/mL $\sigma = 1.2 \times 10^{9}$
Plate 'D' 0.5 mM IPTG (cam)	2.9×10^{10} cfu/mL $\sigma = 2.0 \times 10^{9}$	2.3×10^{9} cfu/mL $\sigma = 1.3 \times 10^{9}$

Table 4.2. Effect of 5 mM $MgCl_2$ on the *in vivo* activity of I-CthI-[IIA1]. *In vivo* endonuclease activity of I-CthI-[IIA1]-pET28b (+) + Cth-*rns*.pACYC184 [BL21] cotransformed constructs presented in cfu/mL. This table depicts the plate assay results of the above construct under different conditions, one is without added $MgCl_2$ and the other is with the addition of 5 mM $MgCl_2$. Three technical and two biological replicates were performed for each of the constructs and the numbers represent the mean of six independent cfu/mL. Standard deviations are also indicated for each of the above observations.

	10 µM $CoCl_2$ in LB media	**10 µM $CoCl_2$ + 5 mM $MgCl_2$ in LB media**
Plate assay (two biological and three technical replicates)	**I-CthI-[IIA1]-pET28b (+) + Cth-rns.pACYC184 [BL21]**	**I-CthI-[IIA1]-pET28b (+) + Cth-rns.pACYC184 [BL21]**
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	2.9×10^{10} cfu/mL $\sigma = 2.6 \times 10^{9}$	2.3×10^{10} cfu/mL $\sigma = 1.5 \times 10^{9}$
Plate 'C' No induction (cam)	2.2×10^{10} cfu/mL $\sigma = 0.9 \times 10^{9}$	2.1×10^{10} cfu/mL $\sigma = 1.6 \times 10^{9}$
Plate 'D' 0.5 mM IPTG (cam)	2.5×10^{10} cfu/mL $\sigma = 1.2 \times 10^{9}$	2.8×10^{10} cfu/mL $\sigma = 1.3 \times 10^{9}$

Table 4.3. Effect of $CoCl_2$ on the *in vivo* activity of I-CthI-[IIA1]. Cobalt chloride antagonism on the possible uptake of magnesium in *E.coli* cells as shown by the *in vivo* endonuclease activity of the HEase expressed from I-CthI-[IIA1]-pET28b (+) and challenged with the substrate plasmid Cth-rns.pACYC184 [BL21]; results are presented in cfu/mL. This table depicts the plate assay results of the above constructs under different conditions, one is with the addition of 10 µM $CoCl_2$ and the other is with the addition of both 10 µM $CoCl_2$ and 5 mM $MgCl_2$ in the LB media. Three technical and two biological replicates were performed for each of the constructs and the numbers represent the mean of six independent cfu/mL. Standard deviations are also indicated for each of the above observations.

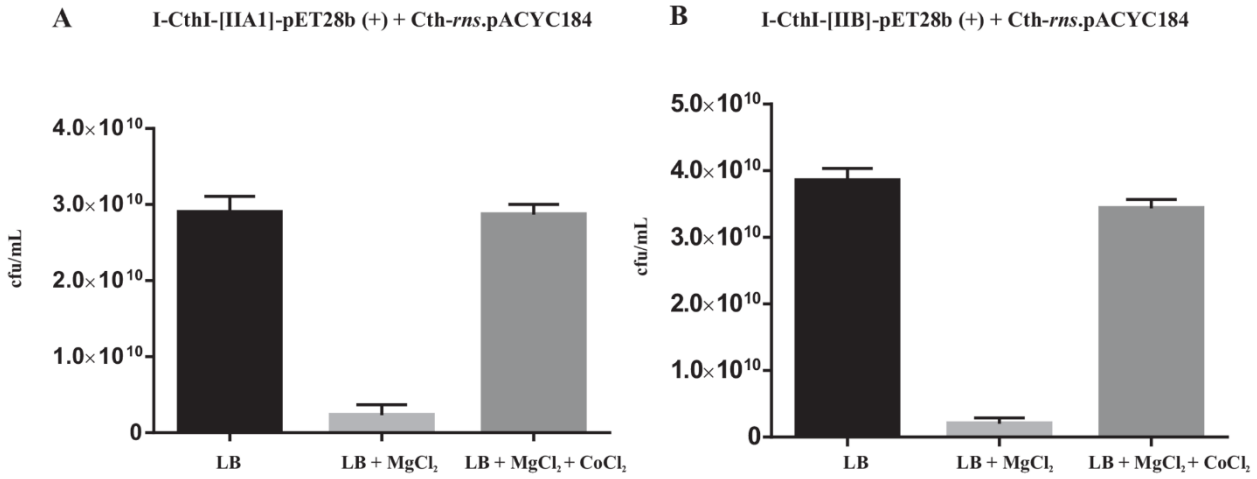


Figure 4.9. The bar graphs summarizing the results of the *in vivo* endonuclease assay. The bar graph shown in panel (**A**) represents the Student's t test performed for assessing the significance of the decrease ($P < 0.0001$) in the number of viable colonies (cfu/mL) for I-CthI-[IIA1]-pET28b (+) and Cth-*rns*.pACYC184 cotransformed construct when grown with the addition of 5 mM $MgCl_2$ in the LB media compared to the cells grown in only LB media. Moreover, the graph also shows significant difference ($P < 0.0001$) with regards to an increase in the number of viable colonies (cfu/mL) when the respective cotransformed cells were grown with the addition of 5 mM $MgCl_2$ and 10 µM $CoCl_2$ in the LB media compared to the cells grown with 5 mM $MgCl_2$ in the LB media. The bar graph shown in panel (**B**) represents the similar results when Student's t test performed for assessing the significance of the decrease/increase ($P < 0.0001$) in the number of viable colonies (cfu/mL) for I-CthI-[IIB]-pET28b (+) and Cth-*rns*.pACYC184 cotransformed construct grown under the same conditions and media as the previous construct. Graphpad Prism 6.01 statistical analysis software was used to calculate the Student's t test and the respective bar graphs were drawn using the same software (Guha and Hausner, 2016).

4.4. Discussion

Group II introns are currently utilized as genome editing tools in the form of targetrons which can be applied for targeted insertional mutagenesis (Enyeart *et al*., 2014; Zhong *et al*., 2003). Previously we described a mtDNA encoded HEase that is encoded within a group I intron (mS1247) where the intron encoded ORF is disrupted by an ORFless group IIA1 intron (Guha and Hausner, 2014). This arrangement hinted at the possibility that the group II intron could be regulatory in nature with regards to the expression of the HEase. Herein we are applying group II introns as regulatory element that allow for the expression of a fungal mtDNA HEase within *E. coli*. Sequences representing either a group IIA1 or a group IIB type intron were inserted into the ORF for the HEase I-CthI at positions that allows for proper intron/exon (i.e. EBS/IBS) interactions so that splicing competent folds could be achieved.

It has been previously shown that an organellar group IIB intron can splice in *E. coli* (Kück *et al*., 1990), in this study we show that group IIA1 introns also have the potential to splice in *E. coli*. It is assumed that group II introns require intron and/or host encoded factors for efficient splicing (Michel *et al*., 2009; Lambowitz *et al*., 2011; Lambowitz and Belfort, 2015; Lamech *et al*., 2014) so this would suggest that within *E. coli* cells, factors are available that can be recruited for the removal of the organellar introns that were investigated in this study. In both cases HEase expressed from constructs where the HEase ORFs were disrupted by group II introns were active and cut their respective substrates at the expected cleavage sites.

For the group IIB intron the intron/exon junctions based on RT-PCR on RNA extracted from *E. coli* were as expected based on previous reports (Kück *et al*., 1990). However, for the group IIA1 intron the intron/exon junction shifted by 18 nucleotides adding 6 amino acid residues to the HEases. These alternate IBS/EBS interactions might be fortuitous but might

suggest that this group IIA1 intron splices differently in *E. coli* compared to its native environment. Therefore, with regards to designing HEases with an "intron based" regulatory element it is important to evaluate the intron/exon junction in the alternate host environment to ensure HEase functionality/specificity has not been compromised. With regards to I-CthI the altered splicing of the group IIA1 intron added six amino acids to a segment of the protein that apparently did not alter the HEases cleavage specificity or the stability of the protein.

HEases have applications in (a) synthetic biology such as iBrick (Liu *et al*., 2009) for the assemble of DNA molecules, (b) as genome editing tools by promoting homologous repair (gene replacements), (c) as a gene targeting tool by promoting mutation inducing non homologous end-joining repair, or (d) as rare cutting enzymes that are part of cloning vectors and cloning strategies (Hafez and Hausner, 2012; Stoddard, 2006). In some instances such as *in vivo* gene targeting temporal regulation of HEase activity might be desirable in order to minimize nonspecific activity of the enzyme. This study showed that modulating the activity of I-CthI in *E. coli* can be accomplished by inserting group II intron sequences into the HEase ORF as splicing of the intron can be stimulated by the addition of Mg^{+2} or antagonized by the addition of Co^{+2}. This strategy would have applications in bacterial systems which are more suited towards group II intron splicing unlike eukaryotic cells (Liu *et al*., 2009; Truong *et al*., 2015; Yao and Lambowitz, 2007). Group II intron sequences in general are readily available (Candales *et al*., 2012) and unlike previous attempts to control HEase activity via redox switches (see PI-SceI; Posey and Gimble, 2002) *in vivo* applications are possible.

The Mg^{+2} transport systems in *E.coli* have not yet been fully elucidated (Ishijima *et al*., 2015). In one study, magnesium has been shown to modulate the function of riboswitches by facilitating the ligand-riboswitch interactions e.g. the btuB riboswitch from *E. coli* (Choudhary

and Siegel, 2014). In our study, exogenous Mg^{+2} concentration was evaluated for manipulating intron splicing which allowed for attenuating the expression of a HEase. In the presence of certain concentrations of Mg^{+2} (5 mM or 10 mM) in the growth media the group II introns appeared to splice and thus functional HEases were produced. Magnesium appears to act as a cationic trigger which might enter the *E. coli* cells through the magnesium transport systems, raising the intracellular magnesium concentration hence facilitating intron splicing. In order to assess if intron splicing is occurring due to the import of Mg^{+2} in the bacterial cells, $CoCl_2$ was used to antagonize the Mg^{+2} effect. Earlier studies have shown that cobalt ion, at concentrations as low as 10 µM, inhibits the energy-dependent transport of Mg^{+2} into cells of *E.coli* (Nelson and Kennedy, 1971, 1972). *In vivo* endonuclease assays in the presence of $MgCl_2$ and 10 µM of $CoCl_2$ showed a reduction in the expression of HEase. This is probably due to Co^{+2} interfering with the entry of magnesium into the cells, leading to Mg^{+2} levels that are not amenable to intron splicing. The failure of the addition of 20 mM $MgCl_2$ to stimulate splicing might be an indication that excess Mg^{+2} can interfere with the proper folding of the group II introns (Donghi *et al*., 2013; Siegel, 2005).

Recently the RNA-guided CRISPR-associated endonuclease Cas9 has been developed into a genome editing tool (Barrangou *et al*., 2007; Cong *et al*., 2013; Mali *et al*., 2013; O′ Connell *et al*., 2014; Sternberg and Doudna, 2015) and it appears well suited for mammalian systems although off-target activity is a concern (Pattanayak *et al*., 2014; Kim and Kim, 2014). The Cas9 protein also appears to be less effective in *E. coli* (Pul *et al*., 2010; Westra *et al*., 2010). With regards to addressing the off-target activity several methods have been developed to control the nuclease activity of Cas9, such as generating versions of Cas9 that are split into two components and these have been engineered to combine by the addition of a chemical signal

such as rapamycin or by blue light irradiation (i.e. a photoactivatable form of Cas9) (Nihongaki *et al*., 2015; Zetsche *et al*., 2015). Another strategy has been to place an "intein" sequence within Cas9 and the intein has been engineered to splice from the host protein when a ligand (4-hydroxytamoxifen) is added to the media (Davis *et al*., 2015). This ligand-dependent intein is somewhat analogous to our "self-splicing" group II introns that can be promoted to splice at the RNA level when suitable levels of Mg^{+2} are present in the media. One can foresee the application of group II intron sequences as agents that allow for inducible genome editing in cell types that are more suited towards supporting the splicing of these elements and can uptake suitable amounts of Mg^{+2}. The ability to antagonize splicing with Co^{+2} provides a "switch like" mechanism where the production of HEase can be stopped or at least attenuated to limit the amount of endonuclease that is produced in a cell and thus potentially avoid nonspecific activities.

In the future with regards to group II intron based "switches" one could achieve even more tighter control by utilizing trans-splicing group II introns. Trans-splicing group II introns (or fragmented group II introns) have been noted in organellar genomes, however it is unknown if these types of introns can function in *E. coli* (Bonen, 2008; Merendino *et al*., 2006). However, it has been shown that the Ll.LtrB group II intron (including a version where the ORF was deleted) from the Gram-positive bacterium *Lactococcus lactis* can splice in *trans* when fragmented at various locations throughout its structure (Quiroga *et al*., 2011). Therefore, a HEase ORF could be split and encoded by two compatible plasmids, carrying different selectable markers and different promoters; with one construct bearing the amino terminal part of the HEase ORF plus the 5' segment of a group II intron sequence and the other construct carrying the 3' segment of group II intron sequence plus the carboxyl terminal part of the HEase ORF. Upon

expression, these two RNAs can assemble via the intron segments into a tertiary structure that promotes trans-splicing of the intron sequences and thus the exons get ligated together to produce a functional HEase transcript.

The current study is "a proof-of-concept" that shows that the expression of a gene can be controlled or at least attenuated by the activity of autocatalytic group II intron sequences. The exact nature of Mg^{+2} or Co^{+2} transport from the media into *E. coli* is not clear but based on our data we can conclude that manipulation of the concentration of positive cations such as Mg^{+2} and Co^{+2} can influence splicing of heterologous introns within *E. coli*. Group II introns could be applied to other heterologous or native proteins that are components of biochemical pathways to allow for temporal control of their expression and possibly promote a shift in metabolic processes. Therefore in the future group II introns could be a potential tool that can be applied not only to genome editing but also to metabolic engineering (Thakker *et al*., 2015; Li *et al*., 2015; Pyne *et al*., 2014).

I would like to thank my lab colleague, Alvan Wai (M.Sc. student) for *in silico* data analysis and taking effort in designing the group IIB intron-HEase construct.

Chapter 5

Bioprospecting for native homing endonucleases from fungal mitochondrial genomes

5.0. Abstract

Buried within the introns at positions mS569 and mS952 of the mitochondrial encoded small ribosomal subunit (*rns*) gene of the ascomycetous fungus *Ophiostoma minus* [strain WIN (M) 371], a group IC2 and a group IIB1 introns were identified respectively. Both introns have open reading frames embedded that encode double motif LAGLIDADG homing endonucleases (I-OmiI and I-OmiII respectively). Unlike I-OmiII protein, expression and purification of the I-OmiI protein was difficult, thus the endonuclease activity of this protein was tested via *in vivo* assays. The cytochrome oxidase b (*cyt-b*) gene of *Ophiostoma novo-ulmi* subspecies *americana* (CM001753.1) was also examined for the presence of introns and intron-encoded ORFs. A similar approach was undertaken for checking the endonuclease activity of a putative homing endonuclease encoded within the c490 intron of the *cyt-b* gene. Overall this study showed that there are native forms of functional homing endonucleases yet to be discovered among fungal mtDNA genomes. Moreover, *in vivo* endonuclease assays provide an alternative approach to screen for potential active homing endonucleases before one invests the time and effort into developing and optimizing protein overexpression and purification strategies for further biochemical studies.

The *in vivo* endonuclease activity conducted with I-OmiI homing endonuclease has been published. Hafez M, Guha TK, Hausner G. 2014. I-OmiI and I-OmiII: Two intron-encoded homing endonucleases within the *Ophiostoma minus rns* gene. Fungal Biol. **118**(8): 721-731.

Conceived and designed the experiments: MH, TKG, GH. Performed the *in vivo* endonuclease experiments for I-OmiI HEase: TKG. Analyzed the data: MH, TKG, GH. Contributed reagents/materials/analysis tools: GH. Wrote the paper: MH, TKG, GH.

5.1. Introduction

Mitochondrial (mt) genomes are highly variable in size and gene arrangements among the fungi, ranging in size from approximately 19 kb to 235 kb (Clark-Walker, 1992; Hausner, 2012; Losada *et al*., 2014). Comparative fungal mtDNA analysis has shown that this variability is in part due to gene content, intergenic spacers, and the presence of intervening sequences (IVS) such as group I and group II introns and intron-encoded open reading frames (ORFs) (Cummings *et al*., 1990; Saldanha *et al*., 1993; Hausner, 2003; Wu and Hao, 2014; Wu *et al*., 2015). The ORFs present within group I and group II introns typically encode homing endonucleases (HEases) and reverse transcriptases (RTs), respectively (Stoddard, 2006; Lambowitz and Zimmerly, 2011).

HEases are rare cutting DNA endonucleases that have been applied in genome editing and as agents for targeted mutagenesis (Arnould *et al*., 2007; Redondo *et al*., 2008; Stoddard, 2011, 2014; Hafez and Hausner, 2012). HEases also have applications in synthetic biology (Liu *et al*., 2014), and as components of cloning vectors and cloning strategies (Li *et al*., 2014). Group I and group II introns are also of significant interest, as autocatalytic RNAs (i.e., ribozymes) have applications in editing RNA molecules (Phylactou *et al*., 1998; Fiskaa and Birfisdottir, 2010) and as regulatory elements (Guha and Hausner, 2014).

Besides being a potential genome engineering tool, HEase contributes towards mtDNA rearrangements by promoting intron mobility and recombination events and there are reports that suggest that HEase activity can cause mtDNA defects resulting in respiratory abnormalities (Dujon and Belcour, 1989; Abu-Amero *et al*., 1995; Hamari *et al*., 2001; Baidyaroy *et al*., 2011; Hausner, 2012; Hafez *et al*., 2013). In addition, mtDNA introns have been associated with QoI fungicide resistance in plant pathogens (Grasso *et al*., 2006). Therefore, testing IEPs for activity

may allow for a better appreciation as to the impact these elements have towards mtDNA evolution and mitochondrial function.

Although group I and group II introns are frequently encountered among fungal mtDNAs, little is known about the functionality of the IEPs; also many intron ORFs appear to be defective (premature stop codons) or are assumed to be pseudogenes (Goddard and Burt, 1999; Gogarten and Hilario, 2006). Moreover, very few of them have been actually tested for their ability to cut DNA (Chevalier and Stoddard, 2001; Stoddard, 2005; Barzel *et al*., 2011; Prieto *et al*., 2012). Although modifying or reprogramming existing HEases scaffolds along with engineering synthetic HEases have been achieved to increase the target repertoire of these proteins (reviewed in Belfort and Bonocora, 2014), the bioengineering of these molecular 'scissors' may sometimes be labour-intensive and time-consuming (reviewed in Hafez and Hausner, 2012; Sander and Joung, 2014). Therefore, characterizing mtDNA genes and their introns can yield a valuable resource for those interested in finding new ribozymes and HEases with novel target sites (Sethuraman *et al*., 2009; Barzel *et al*., 2011; Hafez and Hausner, 2012; Hafez *et al*., 2013). So far most applications are based on a very limited number of well characterized native LHEases (I-SceI, I-CreI, I-DmoI, I-AniI, and I-OnuI; Jacoby *et al*., 2012; Prieto *et al*., 2012). The utility of native HEases in genome editing applications based on systematic characterization of sufficient numbers of LHEase scaffolds has been demonstrated by Takeuchi *et al*. (2011) on HEGs first characterized in blue-stain fungi such as *Ophiostoma novo-ulmi* subspecies *americana* (I-OnuI; Gibbs and Hausner, 2006) and *Leptographium truncatum* (I-LtrI; Sethuraman *et al*., 2009). In that study, the DNA-bound crystal structure of two representative enzymes from the LAGLIDADG family (I-OnuI and I-LtrI) were solved in order to assess the conservation of their protein folds and DNA recognition mechanisms. One of the

enzymes (I-OnuI) was subsequently engineered to cleave and disrupt the human monoamine oxidase B (*MAO-B*) gene.

Characterization of the *Ophiostoma minus* (strain WIN(M)371 = UAMH 9805; Genbank accession: HQ292071) *rns* gene showed the presence of a group IC2 and a group IIB1 intron at positions mS569 and mS952 respectively and both introns contain ORFs that encode double motif LHEases (Hafez and Hausner, 2011a). An initial survey followed by comparative sequence analysis of the cytochrome oxidase (*cyt-b*) gene for members of the Ophiostomatales also revealed the presence of introns and IEPs (unpublished data). Further analysis of the *cyt-b* gene of *Ophiostoma novo ulmi* subspecies *americana* (Genbank accession: CM001753.1) showed that a group IA intron is inserted at position c490 which also encodes a double motif LHEase.

It is not uncommon that recombinant proteins are difficult to express in heterologous expression systems (Samuelson, 2012; Gopal and Kumar, 2013), thereby leaving none or very low amounts of the desired protein for effective purification and other downstream applications. While a common problem is a misfolded protein in a foreign host leading to insolubilities and formation of inclusion bodies (Palmer and Wingfield, 2004), the other challenges may come from factors such as codon usage, translation rate and redox potential (Selleck and Tan, 2008; reviewed in Rosano and Ceccarelli, 2014). The inherent properties of the target protein may be significantly different (such as toxicity) and represent challenges for the expression host (Duong-Ly and Gabelli, 2014). For such 'non-cooperative' proteins or HEases, an indirect approach to test the endonuclease activity was undertaken in this study similar to the concept described by Seligman *et al.* (1997). In that study, a genetic assay for I-CreI activity in *E. coli* was designed by placing the I-CreI ORF on a pB-E plasmid under the control of *araC* promoter and an I-CreI homing substrate site on an F' 128 factor adjacent to a kanamycin resistance cassette; the idea

being, expression of an active I-CreI endonuclease would lead to the loss of the F' kanamycin marker, presumably due to homing site cleavage. Conversely, an inactive HEase would be unable to cleave the homing site, thereby the perpetuating bacterial cells would retain the kanamycin resistance marker.

This chapter describes the protein overexpression and purification trials and potential endonuclease activity of the native, 'non-cooperative' HEases namely, I-OmiI (derived from the mS569 intron position of the *rns* gene) and cytb.i3ORF (derived from the c490 intron position of the *cyt-b* gene). These HEases were studied by using *in vivo* endonuclease assays. Therefore the work presented in this chapter provides an alternative route to determine the potential activity for such 'non-cooperative' HEases.

5.2. Materials and Methods

The methods exclusively related to this chapter have been detailed in this section. For the common materials and methods used in this chapter (which are appropriately mentioned in the text), the readers are directed to Chapter 2 (General Materials and Methods).

The expression, purification and *in vitro* endonuclease assay for the I-OmiI HEase has been previously undertaken by a former Ph.D. student, Dr. Mohamed Hafez (see PhD. thesis "*Exploring the rns gene landscape in Ophiostomatoid fungi and related taxa: Molecular characterization of mobile genetic elements and biochemical characterization of intron-encoded homing endonucleases*", 2012, section 7.4.2 and 7.4.3). However, the expression and purification of this HEase has proven to be problematic during that study and further troubleshooting was needed. In this study, an *in vivo* endonuclease approach was undertaken to determine the activity of I-OmiI HEase.

5.2.1. Design of the *Escherichia coli* expression vectors

The genetic code for the cytb.i3ORF was optimized for expression in *E. coli* and the sequence was synthesized by Genscript. The synthesized ORF (889 bp) was inserted into the pET28 b (+) vector as a BamHI and NdeI fragment with an N-terminal 6 x His-tag; this construct was named cytb.i3ORF.pET28 b (+) which was transformed in *E.coli* BL21 (DE3) for protein expression, purification and biochemical studies. However, repeated trials with this construct did not result in effective protein purification or endonuclease activity, hence a second overexpression construct using the pMAL™ fusion expression vector (NEB) was generated following the manufacturer's protocol. Briefly, the codon optimized ORF segment from cytb.i3ORF.pET28 b (+) was PCR amplified using the cytbi3pMAL-F and cytbi3pMAL-R with

an attempt to introduce the restriction site for BamHI enzyme (italicized, see Table 2.1) at the 3′ end of the PCR fragment. A proofreading polymerase Pfusion (NEB; catalog # M0530S) was absolutely needed to create a blunt end PCR product. The reaction conditions were as listed: initial denaturation: 98 °C for 30 seconds, denaturation: 98 °C for 10 seconds, annealing: 55 °C for 30 seconds, extension: 72 °C for 60 seconds, final extension: 72 °C for 5 minutes. The PCR fragment was purified using the Gel/PCR DNA Fragments Extraction Kit and instructions from the manufacturer (FroggaBio) and 0.5 µg of this purified product was digested with 10 units of BamHI and ligated in pMAL-c5x vector (digested with XmnI and BamHI) using the Quick Ligation kit reagents and protocol (NEB, catalog # M2200S). The construct was named cytb.i3ORF.pMAL-c5x (also see Figure 5.1). Eventually, the fusion construct was used to transform chemically competent *E. coli* ER2523 (NEB Express; catalog # E4131S) following the manufacturer's protocol. In order to check whether the inserted DNA fragment was in-frame with the upstream *malE* gene, the fusion construct was sequenced using either the *malE* forward primer (malE-F) which initiated sequencing 78-81 bases upstream of the multiple cloning site (MCS) or using the pMAL reverse primer (malE-R) which initiated sequencing 75 bases downstream of the SbfI site.

5.2.2. Construction of substrate and non-substrate plasmids

One substrate plasmid and two non-substrate plasmids were constructed in order to assay the endonuclease activity of the HEase expressed from either cytb.i3ORF.pET28 b (+) or the cytb.i3ORF.pMALc5x construct. Briefly, a 193 bp segment of the *cyt-b* sequence (GenBank accession number: CM001753.1) which includes the c490 intron insertion site was synthesized by Genscript and cloned in pUC57 vector as an EcoRV fragment. This substrate plasmid was

named cytb sub.pUC57 and the overall size of the construct is 2.9 kb. Two non-substrate plasmids were also constructed and were used to challenge the specificity of the enzyme. A *C. thermophilum rns* segment containing the mS1247 twintron plus flanking exon sequences (GenBank accession number: JN007486.1) cloned in the pUC57 vector (see Chapter 3; Guha and Hausner, 2014) and a NdeI/BamHI fragment of the cytochrome oxidase (*cox*) gene from *Annulohypoxylon stygium* (GenBank accession number: NC_023117.1) cloned into the pUC57 vector served as the negative controls.

5.2.3. Fusion protein expression and purification

The protein expression and purification for the cytb.i3ORF HEase encoded from cytb.i3ORF.pET28 b (+) construct have been performed following the methods described in section 2.12 and 2.13 respectively, however, the protocols were not successful in recovering enough purified protein for further analysis. Since the cytb.i3ORF.pET28 b (+) construct was non-cooperating in terms of expression and purification, a second construct was generated using the pMAL^{TM} fusion expression vector.

The expression and the purification of the soluble fusion protein from the pMAL-c5x vector was followed according to the manufacturer's instructions (NEB). Briefly, 10 mL of an overnight culture of *E.coli* cells containing the fusion plasmid was inoculated in 1 L of LB media supplemented with 2 g of glucose and ampicillin (100 µg/mL). Glucose was absolutely necessary to repress the maltose genes (e.g. amylase) on the chromosomes which could degrade the amylose on the affinity column. IPTG to a final concentration of 0.2 mM was added to the media (O.D. at A_{600} reached ~ 0.5) to induce the expression of the fusion protein. The cells were further incubated at 37 °C for 3 hours. The cells were harvested by centrifugation at 4000 x g for 20

minutes and resuspended in 25 mL of column buffer (20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, 1 mM DTT) based on the expectation of about 5 g of cells/L i.e., 5 mL for every 1 g of wet weight of the cells. Cell suspension was homogenized twice using the French Press (1200 psi) and the resulting lysate was centrifuged at 20000 x g for 20 minutes at 4 °C. The clear lysate (10 mL) was then mixed with 3 mL bed volume of amylose resin (NEB, catalog # E8021S) and incubated at 4 °C for 20 minutes. The resulting slurry was then poured in a gravity column. The washing steps were listed as follows: Wash 1: 12 column volumes of column buffer or until no more elution of background protein occurred, Wash 2: column buffer supplemented with 10 mM maltose. Collected 15 fractions of 1 mL each and after confirming the presence of the fusion protein on a denaturing SDS gel electrophoresis (12.5%), the desired fractions were pooled and dialyzed against 2 L column buffer to remove the presence of maltose from the fusion protein sample and concentrated to a final volume of 2 mL (3.6 mg/mL) as described in section 2.13 (also see Figure 5.2).

5.2.4. Western blot analysis for detecting fusion protein expression

In order to detect the fusion protein expression at various temperatures, western blot analysis was performed by loading 1/5th (1.6 µg) of the amount of the crude lysate that would normally (8 µg) be run for a Coomassie stained gel. The protein from the gel was transferred to a nitrocellulose membrane (ThermoFisher Scientific; catalog # 88018) following the manufacturer's protocol. The membrane was rinsed thoroughly with TBST buffer (20 mM Tris-HCl, 150 mM NaCl, 0.1% Tween® 20) followed by incubation with 30 mL blocking buffer (TBST + 5% Non-fat Dry Milk) overnight at 4 °C with gentle shaking. The membrane was washed with TBST with gentle shaking at room temperature 3 times for 10 minutes. One µL of

anti-MBP antiserum (NEB; catalog # E8032S) to 10 mL blocking buffer (1/10000 dilution) was used to cover the membrane for 1 hour at room temperature and then washed thoroughly and a 1/10000 dilution of anti-rabbit IgG-peroxidase conjugate (Sigma; catalog # A6154) was used to further incubate the membrane for 1 hour. The membrane was washed with TBST and the manufacturer's directions were followed for detection (also see Figure 5.3).

5.2.5. *In vitro* endonuclease assay

The *in vitro* endonuclease activity was tested for the HEase expressed from cytb.i3ORF.pMALc5x construct as described in section 2.15. Also, following the same protocol in order to rule out the possibility that the Maltose Binding Protein (MBP) alone has influence on the cutting activity on the substrate DNA, 9 μg of the MBP was incubated under the same condition with 25 μg/mL of the substrate DNA (also see Figure 5.4).

5.2.6. *In vivo* endonuclease assay for cytb.i3ORF HEase

In order to confirm the *in vitro* endonuclease assay with regards to the potential of the cytb.i3ORF and I-OmiI to encode active endonucleases, two separate *in vivo* endonuclease assays were established. The overall method used in this assay is common for both of these proteins, therefore the method mentioned in this section has been restricted to the *in vivo* endonuclease assay involving the cytb.i3ORF HEase.

In this assay, two compatible plasmids were maintained in *E. coli* BL21 (DE3) based on the antibiotic selection markers [kanamycin (kan) and chloramphenicol (cam)]. The cytb.i3ORF.pET28 b (+) (kan) (ColE1 origin of replication) construct allowed for the expression of the cytb.i3ORF HEase and the second plasmid served as a substrate. Briefly, the substrate

plasmid was constructed by sequential digestion with XbaI and BamHI and ligated into the pACYC184 plasmid [ATCC 37033 (American Type Culture Collection American Type Culture Collection, Manassas, VA, USA); p15A origin of replication] that was also digested with the same enzymes. The substrate plasmid was named cytb sub.pACYC184 (cam) and if the expressed protein has endonuclease activity, it would cleave the target site within the substrate plasmid leading to the loss of the cam resistance marker. The mechanism by which chloramphenicol is detrimental to the cell viability has been described in Chapter 4 (section 4.2.6). To ensure that proteins expressed by the pET28 b (+) vector (without HEase ORF) were not involved in the endonuclease activity (i.e., negative control), 50 ng of the expression vector was co-transformed along with 50 ng of cytb sub.pACYC184 into 100 µl of chemically competent *E. coli* BL21 (DE3) cells. Another negative control was prepared by co-transforming 50 ng of the expression vector with 50 ng of Cth sub.pACYC184 and transformed cells were plated on LB-agar containing 100 mg/mL kan and 60 mg/mL cam. Plates were incubated at 37 °C for 12-16 hours until the colonies were clearly visible.

For the *in vivo* endonuclease assay three cultures were prepared: (1) cotransformed *E. coli* BL21 (DE3) cells with cytb.i3ORF.pET28 b (+) and cytb sub.pACYC184 or (2) with pET28 b (+) and cytb sub.pACYC184 or (3) cytb.i3ORF.pET28 b (+) and Cth sub.pACYC184 were grown overnight in three separate 5 mL LB media in the presence of the appropriate antibiotics. One percent glucose was added to each media to prevent leaky expression from the T7 promoter. A 0.5 mL aliquot from the 5 mL overnight cultures was used to inoculate 50 mL LB broth cultures supplemented with 100 mg/mL kan, 60 mg/mL cam and 1 % glucose. The cells were grown at 37 °C with vigorous shaking (200 rpm) and the cultures were induced with 0.2 mM IPTG when the O.D. at A_{600} reached 0.58. Separately, for each of the above LB cultures

subcultures generated and these were not induced (i.e., no IPTG was added) which served as additional controls. The cultures were further incubated for 3 hours at 28 °C with vigorous shaking (200 rpm) for expression of the HEase. After 3 hours, the cultures were diluted to 10^{-6} and 100 mL of the diluted cultures were plated on each of the following plates/conditions (done in triplicate): LB agar plates 'A' = without any antibiotics, 'B' = with both 100 mg/mL kan and 60 mg/mL cam, 'C' = with 60 mg/mL cam, and 'D' = with both 0.2 mM IPTG and 60 mg/mL cam. Plates were incubated at 37 °C for 12-16 hours until colonies developed and the colonies were counted. For statistical analysis, unpaired student's t test was performed to determine the significance of the results obtained. Graphpad Prism 6.01 statistical analysis software was used to calculate the Student's t test and the respective bar graphs were drawn using the same software (also see Figure 5.5).

For the details regarding construction of the substrate plasmid and the *in vivo* endonuclease assay performed for I-OmiI, the readers are directed to Chapter 7: Appendices (S7.1; also see Figure 5.6). However, the results of this assay have been provided in section 5.3.5.

5.3. Results

5.3.1. Fusion protein expression and purification reveals several truncated protein fractions for the cytb.i3ORF product

The cytb.i3ORF.pMAL-c5x construct was overexpressed in *E. coli* ER2523 with 0.2 mM IPTG concentration at 37 °C for 5 hours. Upon harvesting and lysing the cells, a small fraction of the full length fusion protein [cytb.i3ORF HEase (35 kDa) + MBP (42 kDa) = 77 kDa] migrated near 75 kDa when resolved by a 12.5% SDS-PAGE. The purification of the fusion protein (MBP-cytb.i3ORF) was achieved by affinity chromatography involving amylose resin. Initially, to wash out the background proteins, 12 volumes of column buffer was used and then column buffer supplemented with 10 mM maltose was applied to elute the fusion protein. The 12.5% SDS-PAGE was used to monitor the purification of the fusion protein. Although a faint band appeared near the desired molecular weight of the fusion protein (77 kDa), there were several intensified protein bands concentrated near the 55 kDa marker; i.e. the full length expression of the fusion protein was minimal (Figure 5.2). The desired fractions containing the fused protein mixture were pooled, dialyzed and concentrated to a final concentration of 3.6 mg/mL, which was sufficient to perform the *in vitro* endonuclease assays. In order to further analyze whether the appearance of these unexpected lower molecular weight intermediates was due to the result of a protein truncation event during the translation to a full length protein or some sort of proteolysis/degradation in the *E. coli* cell environment, a western blot analysis was performed.

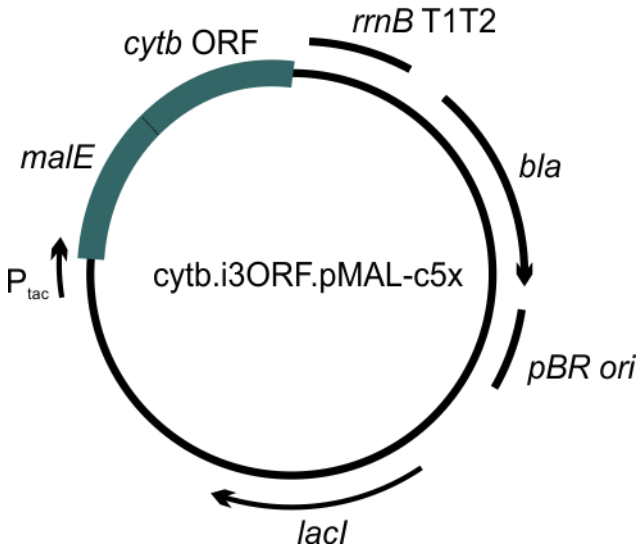


Figure 5.1. Diagram of the pMAL-c5x plasmid bearing the cytb.i3ORF in frame with the upstream *malE* gene encoding MBP. The fusion protein (MBP-cytb.i3ORF) expression is under the control of a tight regulator (P_{tac}) and also has a *lacI* gene which acts as a repressor and prevents any leaky expression in the absence of inductant, IPTG. The plasmid has an ampicillin resistance cassette (*bla*) and pBR_{ori} origin of replication. In addition, the plasmid is also provided with *E. coli* transcriptional termination *rrnB* T_1T_2 downstream of the multiple cloning site.

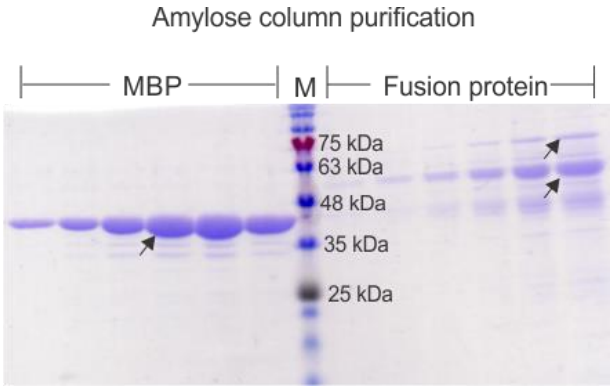


Figure 5.2. Amylose column purification of fusion protein (MBP-cytb.i3ORF) resolved by 12.5% SDS-PAGE. The left hand side of the gel shows the purified bands of MBP (shown by arrow) corresponding to the 42 kDa marker. The right side of the gel depicts the purification of the fusion protein. However, a faint band appeared near the desired molecular weight of the fusion protein, 77 kDa (shown by top arrow). There were several intensified protein bands concentrated near the 55 kDa marker (shown by bottom arrow). 'M' represents the protein ladder.

5.3.2. Western blot analysis confirms truncated/proteolytic cytb.i3ORF products

An immunoblot of fusion protein expression induced at temperatures (28 °C and 37 °C) probed with anti-MBP antibody revealed several darker intensified bands near 55 kDa marker indicating various versions of truncated intermediates or partial degradations. These intermediate bands were absent in the control MBP sample (pMAL-c5x vector), uninduced protein samples and also for the protein samples induced and grown in 16 °C. However, a distinct band of lower intensity corresponding to the 75 kDa marker was observed, indicating a lower expression of the full length fusion protein at 37 °C (Figure 5.3).

5.3.3. *In vitro* endonuclease assay for HEase encoded from the cytb.i3ORF.pMAL-c5x construct partially linearizes the substrate plasmid

In vitro endonuclease assays were performed by incubating the partially purified protein mixture (containing the fusion protein) with the circular substrate plasmid (cytb sub.pUC57). The endonuclease activity of the enzyme was tested at four different time points (0, 30, 60, 90 minutes). At 60 and 90 minutes, the circular substrate plasmid was shown to be partially linearized by the fusion protein MBP-cytb.i3ORF (Figure 5.4). The fusion protein was unable to cleave the non-substrates at the same incubation condition. Moreover, there was no cutting activity observed when the same concentration of MBP was incubated with the substrate plasmid, indicating that MBP alone had no endonuclease activity on the substrate being tested (Figure 5.4).

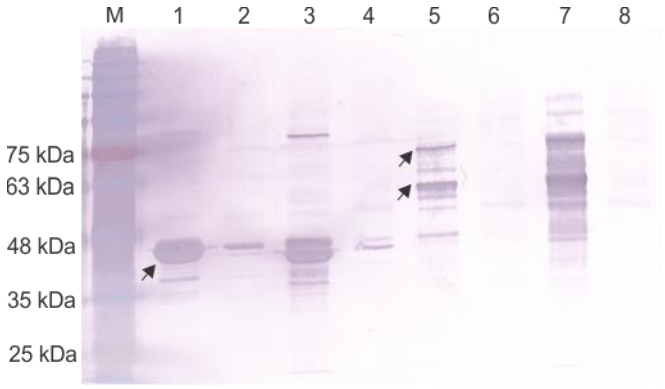


Figure 5.3. Western blot analysis of the fusion protein expression (MBP-cytb.i3ORF). Lane 1 and 2 represent the MBP (shown by arrow) expressed from the control pMAL-c5x vector under induced and uninduced conditions respectively. Lane 3, 5 and 7 represent the expression of the fusion protein at 16 °C, RT (22.8 °C) and 37 °C when 0.2 mM IPTG was added to the protein expression media (LB). The expression of the full length protein (lane 5) corresponding to the 75 kDa protein ladder is shown with an arrow. Moreover, the truncated version of the fusion protein corresponding to 55 kDa protein ladder fragment has also been shown with another arrow. In lane 7 (i.e., fusion protein expressed at 37 °C), most of the intensified bands (detected by antibody against MBP) were located between the molecular size of the MBP (42 kDa) and the full length fusion protein (77 kDa). Lane 4, 6 and 8 represent the expression profile of the fusion protein at 16 °C, RT (22.8 °C) and 37 °C in non-inducive condition where no such expression has been detected. The protein ladder is represented with ‘M’.

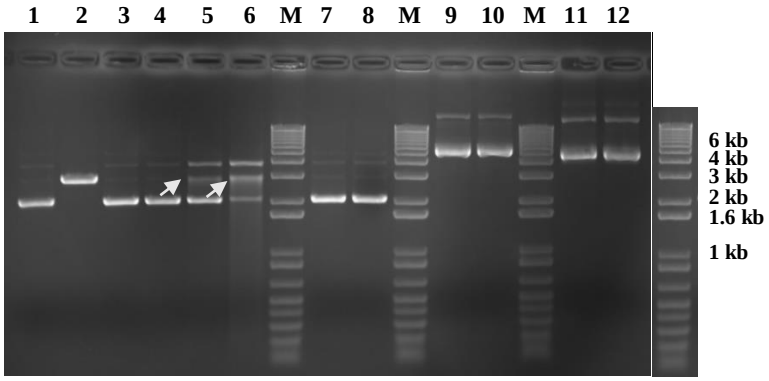


Figure 5.4. A 1% agarose gel showing the results of the *in vitro* endonuclease cleavage assay of the fusion protein MBP-cytb.i3ORF on both substrate plasmid and two non-substrate plasmids. Lane 1 and 2 represent the uncut circular substrate plasmid (cytb sub.pUC57) and the linearized version of the plasmid (cleaved with BamHI) respectively. Lanes 3 through 6 represent the incubation time in minutes (0, 30, 60, 90) of the circular substrate plasmid with the fusion protein. The fusion protein partially linearizing the substrate at 60 and 90 minutes (shown by arrows). Lane 7 and 8 represent the uncut circular substrate plasmid (cytb sub.pUC57) and the same plasmid incubated with 9 µg of MBP alone for 90 minutes to rule out the possibility that this protein is responsible for linearizing the substrate. Lane 9 and 10 represent the uncut circular non-substrate plasmid (Cth-rns sub.pUC57) and the same plasmid incubated with the fusion protein respectively. Lane 11 and 12 represent another uncut circular non-substrate plasmid (cox sub.pUC57) and the same plasmid incubated with the fusion protein respectively. For both the non-substrates, no cutting activity was observed when incubated for 90 minutes at 37 °C. Lane M represents the 1 kb plus™ DNA ladder. The DNA ladder has been selectively labeled in a separate image.

5.3.4. *In vivo* endonuclease assay shows cytb.i3ORF is an active HEase

With a partial success from the results obtained from the *in vitro* endonuclease assay, an indirect approach to test for the endonuclease activity of the HEase was performed. *In vivo* endonuclease assays were performed to evaluate the expression and functionality of the cytb.i3ORF HEase. Three technical and two biological replicates (i.e. six independent values) were performed for each of the assay plates. The mean value for the bacterial colony forming units (cfu/mL) are provided along with their respective standard deviations (σ) in Table S7.3. (see Supplemental Table S7.3 in Chapter 7: Appendices).

The plate assay results from the negative control pET28 b (+) vector cotransformed with cytb sub.pACYC184 substrate were listed in the left column of the table. The results showed that even in the absence or in the presence of 0.2 mM IPTG, when the cells were plated on cam plates, the bacterial colony count was 1.8×10^{10} cfu/mL, $\sigma = 1.8 \times 10^{9}$ and 2.2×10^{10} cfu/mL, $\sigma = 1.3 \times 10^{9}$ respectively. A bacterial lawn was also observed when no antibiotic was applied. These data suggested that the proteins encoded from the empty pET28 b (+) vector were not detrimental to the substrate plasmid carrying the cam resistance marker.

The plate assay results from the *E. coli* cells co-transformed with cytb.i3ORF.pET28 b (+) and cytb sub.pACYC184 are listed in the middle column. The results showed that in the absence of IPTG induction, when the cells were plated on cam plates, the bacterial colony count was 1.6×10^{10} cfu/mL, $\sigma = 1.1 \times 10^{9}$, however when the induced cells were plated on cam plates, there was a 16 fold decrease in the number of viable colonies calculated as 1.0×10^{9} cfu/mL, $\sigma = 0.9 \times 10^{9}$ indicating that the induction of this DNA cutting enzyme must have targeted the substrate plasmid leading to the loss of cam marker. Student's t test performed on the cfu/mL

indicated significant difference (decrease; P value < 0.0001) when compared to the cells grown in the presence of IPTG induction (Figure 5.5).

The right column lists the plate assay results from the *E. coli* cells cotransformed with cytb.i3ORF.pET28 b (+) and Cth sub.pACYC184. This set of assays was performed to establish a negative control for determining the specificity and functionality of this HEase, or to exclude the possibility of native enzymes cutting the vectors. The results showed that even in the absence or in the presence of 0.2 mM IPTG, when the cells were plated on cam plates, the bacterial colony count was 1.9×10^{10} cfu/mL, $\sigma = 1.5 \times 10^{9}$ and 1.9×10^{10} cfu/mL, $\sigma = 2.2 \times 10^{9}$ respectively. These results indicate that even with or without the expression of the functional HEase, the protein could not cleave a non-substrate plasmid thereby sustaining the cam marker.

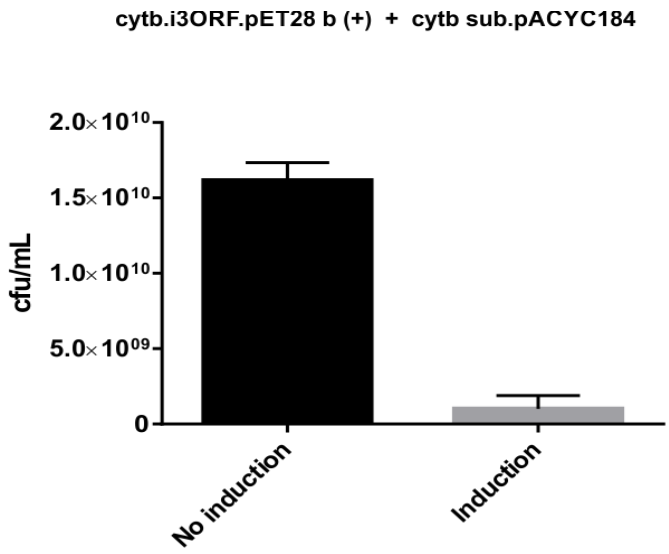


Figure 5.5. The bar graph showing the result (in cfu/mL) of the *in vivo* endonuclease assay when the induced cotransformed bacterial cells were plated on cam plates. The Student's t test was performed for assessing the significant decrease ($P < 0.0001$) in the number of viable colonies (cfu/mL) for cytb.i3ORF-pET28 b (+) and cytb sub.pACYC184 cotransformed construct when grown in the absence and presence of 0.2 mM IPTG in the LB media. Graphpad Prism 6.01 statistical analysis software was used to calculate the Student's t test and the respective bar graphs were drawn using the same software.

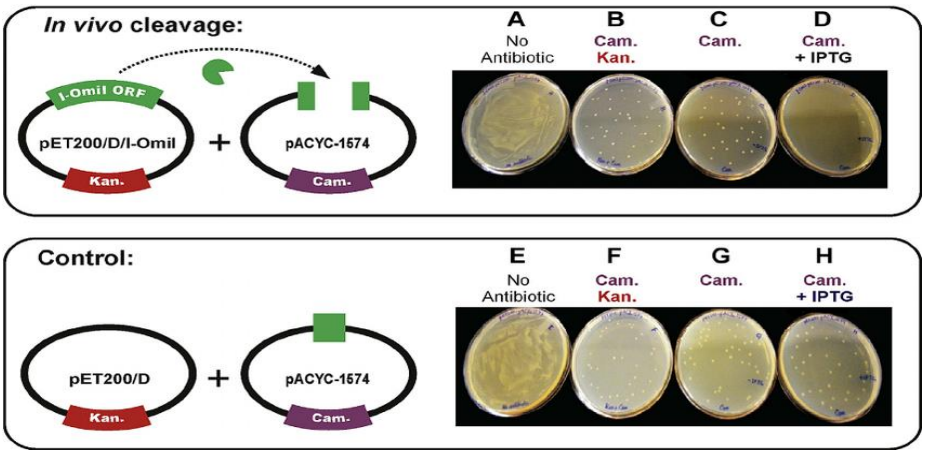


Figure 5.6. The left side shows the constructs used for co-transformation into *E. coli* BL21 (DE3) cells. On the right are indicated the media conditions along with one representative LB agar plate to demonstrate the potential of I-OmiI to cleave the substrate plasmid (for details of the method, see Chapter 7: Appendix). Cells co-transformed with pET200/D/I-OmiI and pACYC/1574 were plated on four sets of LB-agar plates (A through D) with (A) containing no antibiotics, (B) contains both kan and cam antibiotics, (C) contains only cam, and (D) contains cam plus IPTG. As expected on plate (A) (media without any selection) a lawn of cells developed. Plates (B) and (C) contain numerous (>50) colonies due to selection pressure by the antibiotics and plate (D) contains no colonies. The latter is probably due to the induced expression of I-OmiI resulting in the cleavage of the substrate plasmid and the loss of the cam resistance marker. Plates (E) through (H) are a set of control plates (BL21 cells cotransformed with pET200/D and pACYC/1574). Even after induction with IPTG, LB agar plates (H) containing cam showed large number of colonies. This indicates that only I-OmiI was able to cleave the substrate plasmid and not any other endogenous proteins encoded by the plasmids or by the *E. coli* genome (see S7.1 for the method).
Hafez M, Guha TK, Hausner G. 2014. I-OmiI and I-OmiII: Two intron-encoded homing endonucleases within the Ophiostoma minus rns gene. Fungal Biol. **118**(8): 721-731. (Elsevier Publications. Image reproduced with permission. License #3851591051898).

5.3.5. *In vivo* endonuclease assay shows I-OmiI is an active HEase

The intron at position mS569 of the mitochondrial encoded *rns* gene of the ascomycetous fungus *Ophiostoma minus* [strain WIN (M) 371], a group IC2 was identified which encodes a double motif LAGLIDADG HEase (I-OmiI). The expression and purification of the I-OmiI protein was difficult, thus the endonuclease activity of this protein was tested via *in vivo* assays. In terms of expression, I-OmiI expressed within *E. coli* BL21 (DE3) but were found mostly in the inclusion bodies. Therefore expression within this host cell remained a significant barrier to the production of larger amounts of soluble proteins. To better evaluate the potential of I-OmiI to cut the substrate plasmid, an *in vivo* endonuclease system was utilized (Figure 5.6). For the *in vivo* endonuclease assay cells with the pET200/D/I-OmiI and the substrate (pACYC/1574) plasmids, a bacterial lawn was observed on LB agar plates lacking antibiotics (plate A) and numerous colonies were observed on LB agar plates that contain both kan and cam (plate B) and on LB agar plates that contained cam but no IPTG (plate C). However, no colonies were found growing on LB agar plates containing cam and IPTG (plate D). This suggests that I-OmiI has cut the target plasmid leading to the loss of the cam resistance gene. All LB agar plates that served as negative controls (plates E through H) showed a large number of bacterial colonies so the expression of the plasmids did not affect cell viability.

5.4. Discussion

Genome engineering with site-specific nucleases is a rapidly evolving discipline, in which the HEases not only succeeded as therapeutic agents (Arnould *et al*., 2007; Grizot *et al*., 2009; Takeuchi *et al*., 2011) but also proven effective in curbing pest population (Deredec *et al*., 2011), crop-bioengineering (Guo *et al*., 2010; Watanabe *et al.*, 2016) and other genome editing applications (reviewed in Hafez and Hausner, 2012). However, this high specificity of HEases is a "double-edged sword". On one hand, it makes them a powerful tool for precise gene targeting (Marcaida *et al*., 2010; Stoddard, 2011; Hafez and Hausner, 2012). On the other hand, this same high specificity limits the number of DNA sequences that can be efficiently cleaved (Epinat *et al*., 2003; Villate *et al*., 2012). Therefore, much effort has been put into the engineering of these HEases so that one can target almost any gene of choice (Arnould *et al*., 2006; Takeuchi *et al*., 2011; Arnould *et al*., 2011; Prieto *et al*., 2012). Nevertheless, the protein-engineering and production of site-specific HEases is a labour-intensive and expensive process (Barzel *et al*., 2011; Sander and Joung, 2014).

In a study conducted by Takeuchi *et al*. (2011), the idea of scanning the reservoir of natural diversity within the LHEase family has been described as an attractive alternative to extensive protein engineering. In another study by Barzel *et al*. (2011), the tendency of LHEases to tolerate base substitutions in their DNA target sites that correspond to degenerate or "wobble" positions in their host genes has been documented. Therefore, this indicates that homologous conserved genes in humans and animal model systems might be targeted with the same HEase inspite of slight sequence variation among them (Barzel *et al*., 2011). With regards to finding more HEases and thus increasing the repertoire of potential target sites one can screen microorganisms as self-splicing introns/inteins and their encoded HEases tend to be found in

conserved genes such as ribosomal genes (Sethuraman *et al*., 2009; Hafez *et al*., 2013) or protein coding genes like the *cyt-b* and *cox1* (Yin *et al*., 2012; Ferandon *et al*. 2010). PCR based surveys would allow for rapid detection of potential insertions with putative HEases within these conserved genes (Hafez *et al*., 2013, 2014). Even though sequence analysis from metagenomics and microbial sequence databases have hinted at the presence of large number of LHEases (Barzel *et al*., 2011; Takeuchi *et al*., 2011), only a few native HEases so far have been biochemically characterized and applied (Marcaida *et al*., 2010; Takeuchi *et al*., 2011; Prieto *et al*., 2012) thereby limiting the current genome editing applications for these endonucleases. Through *in vivo* endonuclease assays, the current study has demonstrated that both c490 intron from the *cyt-b* gene of *O. novo-ulmi* subspecies *americana* and mS569 intron from the *rns* gene of *O. ulmi* express active HEases, indicating that the blue-stain fungi are a resource for native HEGs, which could be used in various biotechnological applications.

Determining HEases activity and target sites through mechanistic studies of DNA cleavage by HEases demands molecular cloning, expression and purification of the endonuclease proteins, which can sometimes be difficult. Also, reports have shown that overexpression of HEases in common expression systems can lead to cell lysis thereby limiting the desired protein for further application (Jurica and Stoddard, 1999). In terms of expression and purification, both cytb.i3ORF and I-OmiI expressed within *E. coli* but they were found mostly in the inclusion bodies. Therefore expression within this host cell remained a significant barrier to the production of larger amounts of soluble proteins. However, cell free synthesis of meganucleases for structural and functional studies have been described (Villate *et al*., 2012) and these types of systems may have to be explored in the future in order to investigate their utility with regards to the HEases examined in this study.

A statistical analysis conducted on the composition of 81 proteins that do and do not form inclusion bodies in *E. coli* concluded that six parameters are correlated with inclusion body formation: charge average, turn-forming residue fraction, cysteine fraction, proline fraction, hydrophilicity and total number of residues (Wilkinson and Harrison, 1991; reviewed in Makrides, 1996). It may be possible that a combination of the above parameters might have played a detrimental role in terms of protein solubility of both cytb.i3ORF and I-OmiI. While a mild detergent, 1-10% sarkosyl was used to extract the desired proteins from the bacterial cell debri, refolding the aggregated protein and the uncertainty of whether the refolded protein retained its biological activity, remained a challenging task for the studied HEases. Moreover, nickel/cobalt affinity column or any other conventional protein purification techniques such as heparin column or size-exclusion column chromatography failed to purify both the proteins, cytb.i3ORF and I-OmiI HEase.

It has been previously suggested that in *E.coli*, fusion of a target protein to maltose binding protein (MBP) permits significantly enhanced solubility of various non-soluble proteins, ultimately leading to a one-step purification using amylose resin (di Guan *et al*., 1988; Riggs, 2000; Kellerman and Ferenci, 1982). Therefore, the pMAL-c5x vector (NEB) was used for expressing and purifying an MBP fusion protein involving the cytb.i3ORF protein. However, the western blot results showed limiting amounts of the full length protein at 75 kDa instead several unproductive, truncated versions of the protein were generated. This may be due to limitations in the *E. coli* translational machinery to effectively scan long mRNA molecules encoding the fusion ORF. Usually, larger protein molecules are difficult to express effectively in *E. coli* (reviewed in Rosano and Ceccarelli, 2014). This larger fusion protein may also lead to incorrect folding thereby leading to proteolysis, as the cytoplasm of *E. coli* contains a greater number of proteases

than does the periplasm (Swamy and Goldberg, 1982; reviewed in Makrides, 1996). Therefore, in such situations, it would be interesting in the future to investigate the activity of proteins directed to the periplasm as they are less likely to be degraded.

Since the availability of purified HEases in this study has become a bottleneck for biochemical characterization and substrate site determination, we harnessed an alternative approach similar to the *in vivo* systems described previously (Seligman *et al*., 1997; Gruen *et al*., 2002; Chen and Zhao, 2005; Doyon *et al*., 2006). This assay helped to determine the activity of HEases where the overexpression and purification steps were not required. Now that these HEases seem active as evident from the *in vivo* endonuclease assay results, other expression systems might have to be explored in the future. Therefore, the *in vivo* system as developed in this study offers an opportunity to evaluate native HEases that cannot be readily overexpressed in sufficient amounts or recovered in active forms from *E. coli*. Therefore, this method is a screen for HEase activity before more effort and time are investigated in potential candidate IEPs. In addition, this system also allows one to evaluate the functionality of modified/engineered HEases without the need of extensive protein purification and *in vitro* analysis. Moreover, this technique may have practical applications such as isolating mutant forms of native HEases that are able to potentially recognize and cleave variant or novel homing sites. Overall bioprospecting for HEases in intron rich fungal mitochondrial genomes might be good strategy as these are rich sources for potential mobile introns and IEPs.

Chapter 6

Conclusions and Future directions

6.0. The platform for this research

Comparative sequence analysis of the *rns* gene residing within the mtDNA among the species of Ascomycota by Hafez *et al*. (2013) was extremely useful, hence this study provided a solid platform for the research undertaken in this thesis. Hafez and coworkers showed that the *rns* gene was a reservoir for mobile introns and HEGs, indicating that bioprospecting for native HEases could be an attractive alternative to engineering HEases in order to increase the target site repertoire (Hafez and Hausner, 2012; Hafez *et al*., 2013, 2014). Moreover, their study also revealed that the *rns* gene from a thermophilic fungus *Chaetomium thermophilum* DSM 1495 is interrupted by a twintron (nested intron) at position mS1247. This nested intron is composed of an external group I intron encoding a double motif LAGLIDADG open reading frame (ORF) that is interrupted by an ORF-less internal group II intron (Hafez *et al*., 2013). This composition presents a unique possibility whereby splicing of the internal group II intron would allow the ORF to be reconstituted, thus allowing for the expression of the encoded HEase.

The following sections will briefly discuss the major findings which will address the objectives of this thesis. The future studies will also be highlighted which may widen the portal for further accomplishments.

6.1. Major findings

6.1.1. The mS1247 twintron (nested intron) encodes an active I-CthI HEase

In this study, the mS1247 twintron (nested intron) from *C. thermophilum* has been further investigated. My work showed that the mS1247 twintron (nested intron) encoded an active HEase I-CthI, when the internal group II intron was removed from the ORF. This endonuclease was able to cleave both plasmid and linear substrates bearing the *rns* target ‘homing’ site.

Further, the cleavage site mapping analysis showed that the enzyme cuts 8 nucleotide downstream of the twintron (nested intron) insertion site. By performing an *in vitro* splicing assay, we have demonstrated that the group II intron only splices under non-physiological conditions (high salt) that allows for the interrupted HEase ORF to be reconstituted during RNA processing. Furthermore, comparison of the internal group II intron exon binding sequences (EBS) and corresponding intron binding sequences (IBS) of the mS1247 internal group II intron with sequences found in the ORFs located in non-twintron versions of the mS1247 intron, we speculated the possibility of ectopic integration for the origin of this nested intron arrangement in *Chaetomium thermophilum*.

Typically for HEase to be characterized and potentially reprogrammed, one needs to first establish its native target site followed by studying the structure of the protein by obtaining crystals where the HEase will be co-crystalized with its target DNA substrate (Takeuchi *et al*., 2012). In this respect, the crystal structure of I-CthI with its substrate will provide a contact map showing the exact amino acid/DNA sequence interactions which in turn provide valuable information on the strategies that could be used to modify the binding and cleavage activities.

Both, our collaborator, Dr. Barry Stoddard's research group (Fred Hutchinson Cancer Research Centre, Seattle, USA) and in-house crystallographic trials with help from Dr. Brian Mark's laboratory personnels attempted to crystallize the I-CthI protein with very limited success. A possible reason for such limitation is attributed to the highly soluble nature of the protein which apparently failed to precipitate and crystallize in various crystal screens. In the future, it may be possible to revisit the I-CthI protein crystallization using an alternative protein crystallization technique which is Cross-Influence Procedure (CIP) where a set of additives (metallic salts) can be included in the "separate chambers" during the hanging/sitting drop

method that will influence the vapour pressure of the water molecules in the reservoir leading to the nucleation and the quality of crystal growth (Nemčovičová and Smatanová, 2012). The use of the Opti-Salts Suite (Qiagen), which is comprised of premixed salt additive solutions at different pH available in deep-well blocks may be another choice for crystal screening trials.

Several other potential approaches can be taken in order to gain atomic insights. Even if after evaluating many crystallization experiments, no successful conditions are obtained, variations in the I-CthI protein sequence can be generated. This may allow for physical properties of the protein to be modified thus increasing the probability of obtaining crystals. For example, reduction of protein surface charge (Derewenda and Vekilov, 2006; Walter *et al*., 2006), or the removal of flexible loops or the expression and purification of only the essential subdomains of the protein can be taken into consideration. Further, it would be interesting to make an attempt to test homologous proteins (if any) from other organisms and analyze their behaviour towards different crystallization techniques. This homology dependent (indirect) method may help to solve the structure of conserved domains and shed light on the atomic insights to some extent.

6.1.2. Modulating the splicing activity of internal group II introns regulates the expression of the I-CthI HEase in *E. coli* (A proof-of-concept study)

The twintron ORF investigated in this thesis further offers a system wherein an endonuclease could be engineered with an "on switch" where splicing of the internal group II intron could be a regulatory step (or rate limiting step) that allows for the maturation of the HEase transcript, eventually yielding an active HEase. In this aspect, sequences representing either a group IIA1 or a group IIB type intron were inserted into the I-CthI ORF at positions

without compromising the proper intron/exon interactions so that splicing competent folds could be achieved. *In vivo* splicing assays showed that splicing of either group IIA or group IIB intron could be accomplished by the addition of 5 mM or 10 mM $MgCl_2$ in the bacterial growth media. Furthermore, the results from both *in vitro* protein translation and *in vivo* protein expression (*Escherichia coli*) studies supported the above observation. For *in vitro* analysis, the protein production was only observed when the RNA was extracted from cells grown in the culture media supplemented with either 5 mM or 10 mM $MgCl_2$; presumable the correct mRNA for the HEases was generated as the group II introns excised. The functionality of this protein was checked by performing *in vitro* endonuclease assays and cleavage site mapping. Finally, employing *in vivo* endonuclease plate assays involving a pair of HEase construct (HEase ORF interrupted by either group IIA or IIB intron) and a substrate construct carrying different antibiotic markers and origins of replication, we were able to show that exogenous $MgCl_2$ stimulated the expression of a functional HEase. The HEase cleaved the target site on the substrate plasmid leading to the loss of the antibiotic marker; but the addition of cobalt chloride ($CoCl_2$) to growth media antagonized the expression of HEase activity, thereby in these cells the substrate plasmid was not cleaved, thus retaining the antibiotic marker.

Controlling the production of active HEases may be of value in studies where specific target genes have to be modified at a particular stage of development. In the future, one could utilize trans-splicing group II introns while engineering a HEase with group II intron based "switches" in order to achieve even tighter control. However it is uncertain at this point how such trans-splicing introns would operate with regards to functionality in *E. coli* (Bonen, 2008; Merendino *et al*., 2006). A "split-ORF" concept could be utilized in combination with trans-splicing introns. Here the HEase ORF could be split and encoded by two compatible plasmids

carrying different selectable markers and promoters. The amino terminal part of the HEase ORF plus the 5' segment of a group II intron sequence will constitute one construct while the second construct should harbour the 3' segment of group II intron sequence plus the carboxyl terminal part of the HEase ORF. Upon conditions conducive for expression, these two RNAs can promote trans-splicing of the intron sequences and thus ligation of the exons will produce a continuous HEase transcript. This strategy would have applications in bacterial systems which are more suited to group II intron splicing unlike eukaryotic cells due to the limiting free intracellular Mg^{+2} concentrations (Liu *et al*., 2009; Truong *et al*., 2015; Yao and Lambowitz, 2007). However, Truong *et al*. (2013) has shown that enhanced retrohoming and group II intron splicing could be possible in lower Mg^{+2} concentrations by selecting mutations in the distal stem of domain V of the group II intron RNA ribozyme core suggesting a potential application of HEases with group II intron regulators in gene targeting in eukaryotes and mammalian cells.

HEases have applications as rare cutting enzymes that are part of cloning vectors and cloning strategies and also as genome editing tools (Stoddard, 2006; Hafez and Hausner, 2012). In some instances such as *in vivo* gene targeting, temporal regulation of HEase activity might be desirable in order to minimize nonspecific activity of the enzyme. The strategy of inserting an "intein" sequence within Cas9 endonuclease, where the intein has been designed to splice from the host protein only in the presence of a specific ligand being added to the media was a commendable approach (Davis *et al*., 2015). This ligand-dependent intein is somewhat analogous to our group II introns that can be promoted to splice at the RNA level when suitable levels of Mg^{+2} are present in the media. This study showed that modulating the activity of I-CthI in *E. coli* can be accomplished by inserting group II intron sequences into the HEase ORF as splicing of the intron can be stimulated by the addition of Mg^{+2} or antagonized by the addition of

Co^{+2}. Therefore, group II intron sequences as agents that allow for inducible genome editing in cell types may be exploited in biotechnological applications where temporal regulation of expression for DNA cutting enzymes are required. Moreover, group II introns could be applied to other heterologous or native proteins that are components of biochemical pathways to allow for temporal control of their expression. Therefore, this could be a useful component in metabolic engineering (Thakker *et al*., 2015; Li *et al*., 2015; Pyne *et al*., 2014).

6.1.3. Bioprospecting for native HEases, cytb.i3ORF and I-OmiI encoded from introns in fungal mitochondrial genes

The genomes of bacteria, Archaea, phages as well as organellar genomes of many eukaryotes are large natural reservoir of HEases (Barzel *et al*., 2011). Characterization of the *O. minus rns* gene showed the presence of a group IC2 and a group IIB1 intron at positions mS569 and mS952 respectively and both introns contain ORFs that encode double motif LHEases (Hafez and Hausner, 2011a). Similarly, when the *cyt-b* gene of ophiostomatoid species were analyzed, this gene has shown to harbour several introns and IEPs (unpublished data). Further analysis of the *cyt-b* gene of *Ophiostoma novo-ulmi* subspecies *americana*, reveals a group IA intron inserted at position c490 which also encodes a double motif LHEase, this intron ORF was designated as cytb.i3ORF. Since both cytb.i3ORF and I-OmiI HEases were difficult to express and purify through nickel affinity and other protein purification strategies, an alternate route, *in vivo* endonuclease assays was applied to examine if these HEases could be active. The results demonstrated that both these native proteins from fungal mtDNA genomes are active endonucleases.

There are data bases such as REBASE (Roberts *et al*., 2010, 2015), LAHEDES (the

LAGLIDADG homing endonuclease database and engineering server; Taylor *et al*., 2012) that are being assembled and updated, which contains lists of restriction modification enzymes and DNA endonucleases including HEases and their target sites. In the future, these target sites can be screened against the sequences representing the genes of interest (such a genes associated with monogenic diseases) that contain segments that are identical or highly similar to HEase target sites which have been explored in this and similar studies. Moreover, using various target sequences as queries in the NCBI data base, one can scan for sequences as a strategy towards targeting sequences in pathogenic organisms, vectors of pathogens or sequences in human genes involved in monogenic diseases. Therefore, one can aspire and prioritize towards building a catalog of native yet active HEases as these DNA endonucleases due to their high degree of target site specificities have shown potential applications in genome editing and genome modification.

In general, the long term survival of mobile introns and their encoded HEGs by horizontal gene transfer is often attributed to the conservation of target site sequences within the fungal species (or genomes) (Sethuraman *et al*., 2009; Hafez *et al*., 2013). In this respect, the conservative nature of the *rns* and the *cyt-b* genes (unpublished) present with the mtDNA in Ascomycota fungi appear to be the targets for many mobile elements, therefore represent rich reservoirs for native HEases (Hafez *et al*., 2013). Mobile introns are viewed as neutral elements as they avoid damaging the host genome (Hausner, 2012). Due to the absence of the selection pressure on these neutral elements, the ORFs encoding the HEases start to accumulate mutations, subsequently degenerate leading to complete deletion, and thus regeneration of possible homing sites will allow the homing cycle to be repeated (Goddard and Burt, 1999). Therefore, the

presence of an active HEG encoding an active protein (HEase) within a mobile intron could be viewed as an indicator for a more recent horizontal gene transfer event.

Besides contributing towards the rearrangement of the fungal mtDNA by promoting intron mobility and recombination events, HEase activity and improper intron splicing activity can cause fungal mtDNA defects. For example, the splicing deficiency of a mtDNA *rns* group II intron in *C. parasitica* was linked to growth abnormalities and hypovirulence (Baidyaroy *et al*., 2011). In addition to the bioprospecting for native HEases required for genome engineering, testing HEases for activity may allow for a better appreciation as to the impact these elements have towards mtDNA evolution and mitochondrial function.

Chapter 7

Appendices

S7.1. *In vivo* endonuclease assay for I-OmiI HEase

In this assay, two compatible plasmids were maintained in *E. coli* BL21 (DE3) based on antibiotic selection [kanamycin (kan) and chloramphenicol (cam)]. The pET200/D/I-OmiI - kan; ColE1 origin of replication) construct allowed for the expression of I-OmiI and a second plasmid served as the substrate plasmid. For the latter, the *rns* gene (no introns) of WIN(M)1574 was amplified with primers mtsR1 and mtsR2 (see Table 2.1). The resulting PCR product was treated with BamHI and HindIII and ligated into the pACYC1574 plasmid [ATCC 37033 (American Type Culture Collection American Type Culture Collection, Manassas, VA, USA); p15A origin of replication] that was also digested with BamHI and HindIII. The substrate plasmid was named pACYC/1574 - cam and if the expressed protein has endonuclease activity it would cleave a target site within the substrate plasmid leading to the loss of the cam resistance marker.

To ensure that proteins expressed by the pET200 vector (without HEase ORF) were not involved in the endonuclease activity, 50 ng of the empty vector was cotransformed along with 50 ng of pACYC/1574 into 100 µL of chemically competent *E. coli* BL21 (DE3) cells. The transformed cells were plated on LB-agar containing 100 µg/mL kan and 60 µg/mL cam. Plates were incubated at 37 °C for 12-16 hours until the colonies were clearly visible. This assay served as one of the negative controls in the *in vivo* homing endonuclease assay described below.

For the *in vivo* endonuclease assay cotransformed *E. coli* BL21 (DE3) cells (i.e., [pET200/D/I-OmiI and pACYC/1574] or [pET200 and pACYC/1574]) were grown overnight in two separate 5 mL LB media in the presence of the appropriate antibiotics. One percent glucose was added to media containing the HEase-cotransformed construct to prevent the leaky expression from the T7 promoter. A 0.5 mL aliquot from the 5 mL overnight cultures was used to inoculate 50 mL LB broth cultures supplemented with 100 µg/mL kan, 60 µg/mL cam and 1%

glucose. The cells were grown at 37 °C with vigorous shaking (200 rpm) and the cultures were induced with 0.5 mM IPTG when the O.D. at A_{600} reached ~ 0.58. To serve as additional controls, a 50 mL LB culture was not induced (i.e., no IPTG was added). The cultures were further incubated for 3 hours at 28 °C with vigorous shaking (200 rpm) for expression of the I-OmiI HEase. After 3 hours, the cultures were serial diluted to 10^{-6} and 100 µL of the diluted cultures were plated on each of the following plates (done in triplicate): LB agar plates 'A'- without any antibiotics, 'B'- with both 100 µg/mL kan and 60 µg/mL cam, 'C'- with 60 µg/mL cam and 'D'- with both 0.5mM IPTG and 60 µg/mL cam. For the control experiment, plates E through H follow the same order as mentioned above. Plates were incubated at 37 °C for 12-16 hours until colonies developed (see Figure 5.6).

Plate assay (two biological and three technical replicates)	0 mM $MgCl_2$ in LB media (LB) I-CthI-[IIB]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]	5 mM $MgCl_2$ in LB media ($LB+Mg^{+2}$) I-CthI-[IIB]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	4.2×10^{10} cfu/mL $\sigma = 1.5 \times 10^9$	4.3×10^{10} cfu/mL $\sigma = 1.1 \times 10^9$
Plate 'C' No induction (cam)	3.5×10^{10} cfu/mL $\sigma = 2.8 \times 10^9$	3.1×10^{10} cfu/mL $\sigma = 1.3 \times 10^9$
Plate 'D' 0.5 mM IPTG (cam)	3.8×10^{10} cfu/mL $\sigma = 1.8 \times 10^9$	2.0×10^9 cfu/mL $\sigma = 0.8 \times 10^9$

Table S7.1. *In vivo* activity of I-CthI expressed from I-CthI-[IIB]-pET28b (+). *In vivo* endonuclease assay showing the HEase activity as demonstrated in cells that were cotransformed with I-CthI-[IIB]-pET28b (+) and Cth-*rns*.pACYC184 [BL21]; results are reported in cfu/mL. The plate assay results of the above construct under different conditions, one is without added $MgCl_2$ and the other is with addition of 5 mM $MgCl_2$. Standard deviations are also indicated for each of the above observations.

Plate assay (two biological and three technical replicates)	10 µM $CoCl_2$ in LB media I-CthI-[IIB]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]	10 µM $CoCl_2$ + 5 mM $MgCl_2$ in LB media I-CthI-[IIB]-pET28b (+) + Cth-*rns*.pACYC184 [BL21]
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	2.9×10^{10} cfu/mL $\sigma = 2.2 \times 10^{9}$	3.2×10^{10} cfu/mL $\sigma = 2.1 \times 10^{9}$
Plate 'C' No induction (cam)	3.6×10^{10} cfu/mL $\sigma = 1.7 \times 10^{9}$	3.6×10^{10} cfu/mL $\sigma = 1.2 \times 10^{9}$
Plate 'D' 0.5 mM IPTG (cam)	3.2×10^{10} cfu/mL $\sigma = 2.4 \times 10^{9}$	3.4×10^{10} cfu/mL $\sigma = 1.3 \times 10^{9}$

Table S7.2. *In vivo* activity of I-CthI-[IIB] in the presence of $CoCl_2$. Cobalt chloride antagonism on the possible uptake of magnesium in *E.coli* cells during the *in vivo* HEase endonuclease assay in cells cotransformed with I-CthI-[IIB]-pET28b (+) and Cth-*rns*.pACYC184 [BL21]; results reported in cfu/mL. This table depicts the plate assay results of the above construct under different conditions with the addition of either exogenous $CoCl_2$ (10 µM) or 10 µM $CoCl_2$ and 5 mM $MgCl_2$ in the LB media. Standard deviations are also indicated for each of the above results.

Plate assay (two biological and three technical replicates)	**pET28 b (+)/cytb sub.pACYC184 [BL21]**	**cytb.i3ORF.pET28 b (+) / cytb sub.pACYC184 [BL21]**	**cytb.i3ORF.pET28 b (+) / Cth sub.pACYC184 [BL21]**
Plate 'A' No antibiotic	Bacterial lawn observed	Bacterial lawn observed	Bacterial lawn observed
Plate 'B' (kan + cam)	2.3×10^{10} cfu/mL $\sigma = 2.0 \times 10^{9}$	2.1×10^{10} cfu/mL $\sigma = 1.1 \times 10^{9}$	2.0×10^{10} cfu/mL $\sigma = 1.3 \times 10^{9}$
Plate 'C' No induction (cam)	1.8×10^{10} cfu/mL $\sigma = 1.8 \times 10^{9}$	1.6×10^{10} cfu/mL $\sigma = 1.1 \times 10^{9}$	1.9×10^{10} cfu/mL $\sigma = 1.5 \times 10^{9}$
Plate 'D' 0.2 mM IPTG (cam)	2.2×10^{10} cfu/mL $\sigma = 1.3 \times 10^{9}$	1.0×10^{9} cfu/mL $\sigma = 0.9 \times 10^{9}$	1.9×10^{10} cfu/mL $\sigma = 2.2 \times 10^{9}$

Table S7.3. *In vivo* endonuclease activity of pET28 b (+) / cytb sub.pACYC184 [BL21] construct, cytb.i3ORF.pET28 b (+) / cytb sub.pACYC184 [BL21] construct and cytb.i3ORF.pET28 b (+) / Cth sub.pACYC184 [BL21] presented in cfu/mL. The standard deviations are also also indicated. For details, see section 5.3.4.

S7.2. Insertion of ribozyme based switches into a homing endonuclease genes

S7.2.0. Abstract

Fungal mitochondrial genomes act as "reservoirs" for homing endonucleases. These enzymes with their DNA site specific cleavage activities are attractive tools for genome editing, targeted mutagenesis and gene therapy applications. Herein we present strategies where homing endonuclease open reading frames (HEases ORFs) are interrupted with group II intron sequences. The goal is to achieve *in vivo* expression of HEases that can be regulated by manipulating the splicing efficiency of the HEase ORF embedded group II introns. That addition of exogenous magnesium chloride ($MgCl_2$) appears to stimulate splicing of non-native group II introns in *Escherichia coli* and the addition of cobalt chloride ($CoCl_2$) to the growth media antagonizes the expression of HEase activity (i.e. splicing). Group II introns are potentially autocatalytic self-splicing elements and thus can be used as molecular switches that allow for temporal regulated HEase expression. This should be useful in precision genome engineering, mutagenesis, and minimizing off target activities.

Guha TK, Hausner G. 2016. Insertion of ribozyme based switches into homing endonuclease genes. Methods in Molecular Biology - *In Vitro* Mutagenesis: Methods and Protocols; ed. Reeves A. Springer Verlag. 1488; doi. 10.1007/978-1-4939-6472-7 (in press).

Conceived and designed the experiments: TKG, GH. Performed the experiments: TKG. Analyzed the data: TKG, GH. Contributed reagents/materials/analysis tools: GH. Wrote the book chapter: TKG, GH.

S7.2.1. Introduction

Homing endonucleases (HEases) are site-specific DNA cleaving enzymes that are encoded by homing endonuclease genes (HEGs) which are frequently found embedded within mobile elements (reviewed in Stoddard, 2006) but sometimes HEGs can be freestanding (Gimble, 2000). HEases promote their own mobility and the mobility of the elements that host them by introducing site-specific double-stranded breaks in cognate alleles that lack HEGs or intron/intein insertions thereby stimulating the double-stranded DNA repair process which involves homologous recombination (Stoddard, 2006; Hausner, 2012). The LAGLIDADG family of HEGs (LHEases) are frequently encoded within fungal mitochondrial group I introns (Hausner, 2012) and these enzymes recognize long asymmetrical 12-40 bp of DNA sequences as their target sites. Due to their target site specificity HEases have applications in (a) DNA sequence assemble or synthetic biology (Liu *et al*., 2014), (b) as genome editing tools by promoting gene replacements via homologous repair (Stoddard *et al*., 2008; Marcaida *et al*., 2010; Takeuchi *et al*., 2011; Stoddard, 2011; Hafez and Hausner, 2012; Prieto *et al*., 2012), as a gene targeting tool by promoting mutations generated by non-homologous end-joining repair (Takeuchi *et al*., 2011; Stoddard, 2011; Hafez and Hausner, 2012; Prieto *et al*., 2012) or as rare cutting enzymes that are part of cloning vectors (Hafez and Hausner, 2012). Sometimes procedures involving *in vivo* gene targeting the temporal regulation of HEase activity might be essential in order to minimize off target activities of the enzyme (Posey and Gimble, 2002).

In this chapter we describe an on/off "switch" system that provides an opportunity for the temporal control of HEase activity in *Escherichia coli*. Splicing of group II introns requires the intron RNA to fold into a splicing competent tertiary structure that requires interactions between intron and flanking exon sequences. So called intron binding sequences (IBS), located upstream of the intron insertion site, are needed for splicing as they interact with the corresponding exon

binding sequences (EBS1 and EBS2) present within the intron (Olga and Nora, 2007; Michel *et al*., 2009). Group II intron derived ribozymes are metalloenzymes (Donghi *et al*., 2013; Sigel, 2005) and they require positive cations like magnesium (Mg^{+2}) for catalysis (Lambowitz and Belfort, 2015). Strategies will be presented with regards to inserting group II intron sequences into expression constructs and for manipulating the *in vivo* splicing of these introns by stimulated splicing with the addition of Mg^{+2} or antagonizing splicing by the addition of cobaltous ion (Co^{+2}) in the form of cobalt chloride. It should be noted that this strategy of using ribozyme based switches could be applied to other protein based genome editing tools such as TALENS, Zn-finger endonucleases and the cas9 (CRISPR) based systems.

S7.2.2. Materials

2.1. Related to nucleic acids (Plasmid prep, transformation, RT-PCR etc.)

1. All buffers use DNAs/RNAse free sterile water.
2. Commercially synthesized HEase ORF (with group II intron) cloned in expression vector.
3. Commercially available *E. coli* competent cells (e.g. NEB5α-derivative of DH5α from New England Biolab; BL21 (DE3) from Thermo Fisher Scientific).
4. Temperature controlled water bath (42 °C/ 65 °C).
5. Media: Super Optimal broth with Catabolite repression (SOC); composed of 2% Tryptone, 0.5% Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM MgSO4, 20 mM glucose.
6. Shaker incubator / Rotary shaker incubator (37 °C).
7. Pre-warmed LB-agar plates.
8. Antibiotic stock solutions (e.g. 100 mg/mL Kanamycin, 60 mg/mL Ampicillin)
9. 70% Ethanol.
10. 95% Ethanol.
11. Wizard® Plus Minipreps DNA purification kit (Promega, Madison).
12. PCR reaction mixture (total volume 50 µL) ingredients (µL/reaction): 10 x Taq DNA polymerase buffer (5); 50 mM $MgCl_2$ (0.5); 2.5 mM dNTP (4); 40 pmol each forward and reverse primer (0.5 + 0.5); H_2O (38.25); DNA template (1 µL ~ 10 to 100 ng); and Taq DNA polymerase (0.25; ~ 2.5 units).
13. DNA storage buffer: 1 x Tris-EDTA (TE) buffer (10 mM Tris-HCl, pH 7.6, 1 mM $Na_2EDTA{\cdot}2H2O$).
14. Commercially available desired restriction enzymes and respective buffers.
15. RNA purification kit (GENEzol TriRNA Pure Kit, Geneaid, FroggaBio).

16. Thermoscript Reverse Transcriptase Kit (Thermo Fisher Scientific).

17. BigDye®Terminator sequencing system (Thermo Fisher Scientific).

18. Endonuclease reaction buffer: Reaction buffer #3 (Thermo Fisher Scientific): 50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 100 mM NaCl supplemented with 1 mM DTT.

19. Agarose gel loading buffer (6 x): 3 mL glycerol (30%), 25 mg bromophenol blue (0.25%) dH_2O to 10 mL.

20. Tris-borate EDTA buffer: 1 x TBE buffer (89 mM Tris-borate, 10 mM EDTA, pH 8.0).

21. Agarose Gel: Ultra-pure agarose (Thermo Fisher Scientific).

22. Micro centrifuge

2.2. Related to protein work

1. Luria-Bertani Broth (LB) media: For 1 L of LB mix the following reagents in a 2 L glass container and stir thoroughly; 10 g Tryptone, 5 g Yeast extract, 5 g NaCl, 1 L MilliQ water, add 200 µL of 5 N NaOH and autoclave.
2. Terrific broth (TB) media (optional): Measure ~ 900 mL of distilled H_2O, 16 g Tryptone, 10 g Yeast Extract, 5 g NaCl, adjust pH to 7.0 with 5 N NaOH, adjust to 1 L and autoclave.
3. Qiagen Nickel-NTA Superflow resin and column.
4. SDS PAGE: 30% Acrylamide/Bis solution (37.5:1), 10% Ammonium persulfate, TEMED (BioRad), 1 M Tris-HCl pH 8.8, 0.5 M Tris-HCl pH 6.8, 10% (w/v) of Sodium dodecylsulfate (SDS) stock solution in H_2O.
5. Cell Lysis (CL) buffer: 50 mM Tris-HCl, pH 8.0, 0.3 M NaCl.
6. Wash Buffer 1 (WB1): CL + 25 mM imidazole.
7. Wash Buffer 2 (WB2): CL + 50 mM imidazole.
8. Wash Buffer 3 (WB3): CL + 100 mM imidazole.

9. Elution Buffer 1 (EB1): CL + 250 mM imidazole.
10. Elution Buffer 2 (EB2): CL + 500 mM imidazole.
11. Dialysis Buffer: 50mM Tris-HCl pH 8.0, 150 mM NaCl and 1 mM DTT.
12. Protein storage buffer: 50 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM dithiothreotol (DTT), 30% (w/v) glycerol.
13. 2 x protein loading dye (65.8 mM Tris-HCl, pH 6.8, 26.3% (w/v) glycerol, 2.1% SDS, 0.01% Bromophenol blue).
14. Amicon concentrator, Ultrafiltration membranes (desired Molecular weight cut off), and Amicon Ultra-4 Centrifugal filters (select for desired Molecular weight cut off).
15. Highspeed centrifuge and rotors (SLA1500 and SS34 rotors)

S7.2.3. Methods

3.1. Design of the *Escherichia coli* expression vector for HEases

First one has to select a HEase sequence that is known to be functional and the sequence has to be codon optimized for being expressed in *E. coli*. HEases are commercially available and can be engineered to intended target sequences, but one can start with "native HEases" and see if some by chance cut within a gene of interest. Suitable webserver resources to aid in codon optimization and potentially evaluate the expression of the HEase sequence in *E. coli* (or other hosts) are http://genomes.urv.es/OPTIMIZER/ (Puigbo *et al*., 2007) and http://mbs.cbrc.jp/ESPRESSO/TopPage.html (Hirose and Noguchi, 2013) respectively (see section 3.2). The choice of group II introns is obviously critical. Group II introns have a wide distribution and are found in all three domains of life (Lambowitz and Belfort, 2015) they have been primarily classified based on structural details (RNA folding, Toor *et al*., 2001; Lambowitz

and Zimmerly, 2004; Lambowitz and Zimmerly, 2011), their intron encoded proteins (if present), and depending on how these intron fold it tends to have implication on exon sequences involved in generating splicing competent folds (Olga and Nora, 2007; Michel *et al*., 2009). The secondary structure of Group II intron RNA can be viewed as a central wheel from which 6 "fingers" i.e. domains (I through VI) emerge. Domain I contains the exon binding sequences (EBS) that ultimately interact with elements within the flanking exon sequences (referred to as intron binding sequences - IBS). So it is important to investigate the choice of intron and be aware of the splicing requirements for the intron. Based on the current literatures, the following group II introns have been well characterized and may offer good starting points: *Chaetomium thermophilum* mtDNA mS1247 nested group II intron (Hafez *et al*., 2013; Guha and Hausner, 2014); the mtDNA rI1 of *Scenedesmus obliquus* (Kück *et al*., 1990; Hollander and Kück, 1990); and the bacterial Ll.LtrB, Ecl5, Rmint1 and B.hl1-B introns (Lambowitz and Belfort, 2015). All of these introns have been well characterized with regards to their requirements for exon recognition and splicing conditions in various hosts (Lambowitz and Belfort, 2015). It is also best to choose introns that lack ORFs or remove ORFs if present and select introns that have rather "simple" exon recognition requirements. For example group IIB introns require three intron/exon interactions (i.e. EBS1, 2 and 3 plus corresponding IBS 1, 2 and 3) whereas it has been show that the rI1 a group IIB intron actually will splice efficiently in *E. coli* as long as the IBS1 sequence is provided in the upstream exon (Kück *et al*., 1990; Hollander and Kück, 1999). The less interactions needed by the intron means less manipulation of the HEase ORF sequence is required.

1. Select a known functional HEase for the insertion of a ribozyme based switch (see Note 1).
2. For selecting a suitable group II intron that could serve as a "switch" examine either group IIA

intron or group IIB introns from the NCBI Genbank (http://www.ncbi.nlm.nih.gov/genbank/), or consult a group II data base (http://webapps2.ucalgary.ca/~groupii/) (Candales *et al*., 2012) and/or the Comparative RNA web site (http://www.rna.icmb.utexas.edu/ ; Cannone *et al*., 2002) (see Note 2).

3. Determine the Intron binding sites (IBS) usually upstream (6-12 nucleotides) of the intron insertion site however depending on the type of intron IBS components can be downstream of the introns native insertion site. (Lambowitz and Belfort, 2015; also see Note 3).
4. Prior to the group II intron sequence being inserted in the HEase ORF, it is necessary to match the required intron based EBS sequences with the exons (ORF) potential IBS sequences (see Note 4). This may require some manipulation of the HEases sequence and determine where the intron is inserted.

3.2. Codon-optimization and gene synthesis

1. A codon-optimized version of the HEG sequence should be synthesized to account for differences between the fungal mitochondrial and bacterial genetic code and codon-biases (see Note 5). DO NOT modify the selected IBS sequence(s) and the internal intron sequences.
2. Clone the ORF sequence with the embedded group II intron sequence in an expression plasmid with an inducible T7 promoter (e.g., pET28 b (+)) for overexpression. We will refer this HEase ORF interrupted by group II intron construct as “ORF-switch” in the text.
3. Sequence the “ORF-switch” plasmid in order to confirm the orientation and to ensure that the ORF is in frame with the vector that provides the start codon and the N-terminal 6 x-His-tag.

3.3. Design of the HEase substrate to access functionality of the HEase ORF

1. Construct a substrate plasmid by inserting a DNA segment that contains the target site for the HEase, such as an allele that does not contain the HEase (and/or associated intron) sequence.
2. Also generate a control plasmid by inserting a DNA fragment that lacks the HEase target site such as the allele with the HEG (and/or intron) insertion (see Note 6).
3. Synthesize and clone the substrate sequence in any suitable plasmid (e.g., pUC57 vector).
4. Sequence the substrate plasmid to ensure that the insert is in place (see Note 7).
5. Transform the plasmids (substrate and control) into *E. coli* DH5α separately (see section 3.4) and then purify the constructs from ~5 mL LB overnight cultures with any suitable plasmid purification kit.

3.4. Chemical Transformation protocol

For transforming the "ORF-switch" construct, *E. coli* BL21 cells are recommended as they are efficient for the overexpression of heterologous proteins and for maintaining the substrate or non-substrate control constructs, *E. coli* DH5α cells can be considered. The chemical transformation method will be detailed below as a common procedure. Readers must take into account which constructs are being transformed into what cell line (see Note 8).

1. Add 1 µL of the plasmids into vials containing 100 µL of chemically competent *E. coli* cells and mix gently. Avoid pipetting up and down.
2. Incubate the vials on ice for 5 to 30 minutes (see Note 9).
3. Heat-shock the cells for 1 minute at exactly 42 °C without shaking.
4. Transfer the vials onto ice and keep for 2 minutes.
5. Add 300 µL of pre-warmed SOC medium at room temperature to the vials.

6. Tightly cap the tubes and shake horizontally (200 rpm) at 37 °C for 1 hour.
7. Spread 100-150 µL of the mixture on a warm LB agar plate containing the appropriate antibiotic(s) and incubate at 37 °C till the colonies are clearly visible (usually 16-24 hours).

3.5. Analyzing clones of interest

1. From the above LB agar plate, examine colonies and take cells of interest and inoculate 5 mL LB cultures that are kept at 37 °C with agitation for 14-18 hours.
2. Three mL of the LB culture are collected for extracting plasmid DNAs. Plasmid DNAs are recovered by various methods (Green and Sambrook, 2012); however, one can also perform colony PCR (from step 2 above) screening (Dafa'alla *et al*., 2000) to confirm colonies that maintain the plasmid of interest.
3. Perform restriction enzyme digestion to confirm the presence of the correct construct/plasmid. Ideally one should use a restriction enzyme or a combination of enzymes that cut once in the vector and once in the insert.
4. Resolve and visualize restriction digests by agarose gel electrophoresis (Green and Sambrook, 2012).

3.6. Gel electrophoresis

1. Preparation of a 1% Agarose gel: Add 1 g ultra-pure agarose (Life technologies) to 100 mL (volume depends on size of gel tray, adjust according to manufactures recommendation) of 1 x TBE buffer then mix and melt agarose in microwave oven. Once the agarose has completely dissolved, allow to cool to about 55-60 °C and pour into an assembled gel casting tray with positioned comb. Let the gel to solidify at room temperature and carefully remove the comb

and place the gel into an electrophoresis box containing 1 x TBE buffer.

2. Mix each DNA sample with the agarose gel loading buffer and load samples into the wells of the gel. Electrophorese at 80-120 volts until the tracking dye migrates to the positive electrode end of the gel. Resolved DNA fragments are sized with a DNA ladder (such as 1 kb plus™ DNA ladder by Invitrogen/Life technologies).
3. Stain nucleic acids by soaking gel in 1 x TBE buffer supplemented with 0.5 µg/mL ethidium bromide (EtBr) and expose the stained gel with ultraviolet light.

3.7. Preparing the cells (transformants) for long-term storage

1. Once a colony with a construct of interest has been identified, mix 0.85 mL of the culture with 0.15 mL of 50% sterile glycerol and transfer to a cryovial and store at -80 °C. For simplicity, for the HEase expression construct with the group II intron, we will refer this glycerol stock as "ORF-switch" stock. Other constructs such the substrate plasmids etc. can be preserved in the same manner.
2. As an additional backup always store an aliquot of purified plasmid DNA at -20 °C.

3.8. *In vivo* RNA splicing assay

Reverse Transcriptase PCR (RT-PCR) needs to be employed to examine *in vivo* splicing activity of the HEase ORF group II intron. In particular to determining the concentration of exogenous $MgCl_2$ that has to be added to the growth media in order to induce splicing of the group II intron. The plasmid derived HEG transcript has to be evaluated to verify that splicing has occurred and to ensure that splicing in *E. coli* maintains the expected intron/exon junctions. This is important otherwise frameshift mutations could be introduced.

1. Inoculate the “ORF-switch” stock in 10 mL of LB media supplemented with appropriate antibiotic (e.g., 100 µg/mL of kanamycin for constructs if cloned in pET28 b (+)) and 0.25% w/v glucose. Also inoculate the control (no HEase ORF) vector (e.g., pET28 b (+) in BL21) in another LB media with the same concentration of antibiotic and glucose.
2. Incubate the cultures overnight in a rotary incubator at 37 °C.
3. Prepare several 50 mL LB culture flasks and supplement with 1 mM, 5 mM, 10mM, 20 mM upto 100 mM of magnesium chloride ($MgCl_2$).
4. Inoculate the 50 mL LB culture flasks with 500 µL of the overnight cultures (see section 3.8.2). For a negative control, inoculate a 50 mL LB culture flask with any added $MgCl_2$ with 500 µL of the overnight culture.
5. Grow the 50 mL cultures at 37 °C with agitation till the O.D. at A_{600} reaches 0.65. (see Note 10).
6. Pellet the bacterial cells from 10 mL aliquots from the above cultures (including the negative controls) by centrifuging for 3 minutes at 4000 x g with SS34 rotor (Sorval, Thermo Fisher Scientific).
7. Lyse the pelleted cells and extract the RNA using any bacterial RNA extraction kit following the manufacturer’s protocol. Make sure you set aside at least 1 µg of RNA for *in vitro* translation (see section 3.9).
8. Treat the extracted RNA samples with 2 units of DNaseI and incubate at 37 °C for 15 minutes. Stop the reaction by adding 1 µL EDTA (50 mM) followed by 10 minute incubation at 65 °C.
9. Take out 2 µL from each of the reaction mixtures and perform a standard PCR reaction by using the HEase ORF specific primers in order to confirm the complete elimination of any residual DNA from the extracted samples.

10. Run 1% agarose gel, this PCR reaction should not yield any amplification products to indicate the removal of all DNA from the RNA sample (see Note 11).
11. Perform RT-PCR to make cDNA from the transcript of interest contained within the extracted RNA samples using a standard RT-PCR kit following the manufacturer's protocol.
12. The cDNA obtained in step 11 can now be used as template for performing standard PCR using HEase ORF specific primers. This will now determine the splicing potential of the group II intron and show possible splicing intermediates. A successful group II splicing event should yield a single PCR product corresponding to the difference between the distance of the forward and reverse primers minus the nucleotide length of the inserted group II intron used for the study (see Note 12).
13. Gel excise the PCR amplicon corresponding to the desired length as mentioned above using Gel/PCR DNA extraction kit following the manufacturer's protocol.
14. Sequence the gel extracted fragment utilizing the primers required for obtaining the amplicon in order to determine whether correct splicing occurred or not. (i.e., investigate the intron/exon splicing junction) (see Note 13).
15. Note the concentration of $MgCl_2$ added in the LB culture flask(s) that yielded the correct splicing product.

3.9. *In vitro* HEase expression

1. To assess whether the HEase can be expressed in an "*E. coli*" environment, perform *in vitro* translation with the RNA extracted from the *E. coli* bacterial cells grown in LB media which was supplemented with the pre-determined $MgCl_2$ concentration that induced proper splicing of the group II intron (see section 3.8.7).

2. For *in vitro* translation, one can use a commercial *in vitro* protein synthesis kit (e.g., PURExpress *In Vitro* Protein Synthesis Kit, New England Biolab, MA, USA) following the manufacturer's protocol (see Note 14).
3. After a minimum incubation of at least 3 hours at 37 °C, mix 2.5 µL of the reaction mixture with 2.5 µL of the 2 x protein loading dye, resolve the proteins in SDS-PAGE and analyze for the presence of the desired HEase protein by comparing the resolved proteins and scanning for those with to the expected molecular weight based on the protein ladder.

3.10. *In vivo* HEase overexpression-Small scale overexpression trials

It is assumed that the overexpression conditions for the functional HEase are known (see section 3.11). However, readers can reassess the overexpression conditions as follows:

1. Inoculate small flasks (50 mL of LB media containing appropriate antibiotic supplemented with 0.25% w/v glucose) with 500 µL of overnight culture of *E.coli* (which was transformed with the "ORF-switch" construct).
2. Supplement the LB media with the pre-determined concentration of $MgCl_2$. Inoculate another small flask with just *E.coli* BL21 containing only the control plasmid (plasmid containing no insert).
3. Grow the cultures (with agitation) at 37 °C till $O.D._{600}$ reaches 0.65 and then induce with 0.2 mM IPTG (low) and 1 mM IPTG (high) to the respective flasks and shift flasks to various temperatures. Several trials may be required to optimize the concentration of IPTG (range: 0.1 mM-1 mM) and temperature (range: 15 °C-37 °C) for proper induction (i.e. stable protein expression).
4. Incubate the flasks at various temperatures for 6 hours or overnight.

5. Pellet cells via centrifugation at 4000 x g for 10 minutes at 4 °C with high speed centrifuge.
6. Discard supernatant and resuspend pellets in 2 mL of cell lysis buffer.
7. Sonicate in short pulses for 15 seconds thoroughly to lyse the cells. Keep vials on ice during the entire period.
8. Centrifuge at 16000 x g for 15 minutes at 4 °C with a high speed centrifuge and collect the crude protein extract in microcentrifuge tubes. Keep on ice.
9. Determine the concentration of the crude protein mixture by A_{260} / A_{280} ratio using a spectrophotometer.
10. Analyze the samples by SDS-PAGE using about 8 µg of each of the protein extracts plus the same amount of protein from the control sample(s).
11. Check the SDS-PAGE protein gel for overexpression of the protein of interest by scanning for a band in the appropriate expected size range that is absent in the control lane. One can also perform a western blot with any commercially available anti-His antibody to further confirm the presence of the His-tag on the overexpressed protein which is required for purification in the later steps. Once specific parameters have been determined for the overexpression one can proceed to the large scale overexpression of the HEase.

3.11. Large scale overexpression of the HEase

1. Inoculate 10 mL LB media (supplemented with appropriate antibiotic and 0.25% w/v glucose) with a small amount ~ 10 µl of the "ORF-switch" glycerol stock and incubate overnight at 37 °C in a rotatory incubator.
2. Inoculate 1 L of LB medium (supplemented with 100 µg/mL of kanamycin and 0.25% w/v glucose plus the optimal amount of $MgCl_2$, see section 3.8) with 5 mL of the overnight

culture prepared in step 1 above.

3. Grow the culture at 37 °C with agitation and induce with IPTG (see section 3.10) when the $O.D._{600}$ reaches ~ 0.65 and grow further at the pre-determined conditions for over expression.
4. Harvest the cells by centrifugation at 4000 x g for 10 minutes and freeze pellet at -80 °C.

3.12. Purification of the HEase

1. Thaw the pellet in a warm water bath and resuspend in 10 mL of CL buffer per 1 g wet weight of cells. Stir the suspension for 30 minutes at 4 °C in order to make it homogeneous.
2. Lyse the cells using a French press two times (as needed) and centrifuge lysate at 16000 x g for 30 minutes at 4 °C to pellet cellular debris.
3. Add the clear lysate to 3 mL of Ni-NTA resin (Qiagen, Toronto) and incubate at 4 °C with shaking for 30 to 60 minutes.
4. Load the crude-extract onto a Ni-NTA super flow column (Qiagen, Toronto).
5. Carry out the following series of washings with wash 1: 30 mL the WB1; wash 2: 30 mL of WB2 buffer; and wash 3: 30 mL of WB3 buffer. Collect and save 1 mL of each wash.
6. Elute the protein in Elution buffer EB1, if necessary EB2. Collect the eluting samples in 1.5 mL microfuge tubes as 700 µL fractions.
7. Remove excess imidazole by dialysing in the dialysis buffer using a slide-a-lyzer dialysis cassette (Millipore, Billerica, USA) with a desired molecular weight (MW) cut-off.
8. Check the concentration of the protein using the absorbance (A_{280}) function of a spectrophotometer and analyse the fractions by performing SDS-PAGE (see Note 15).
9. Pool the desired fractions to a final volume of 9 mL in a protein storage buffer and concentrate using Amicon Ultracel centrifugal filters (Millipore, Billerica, MA) with a pre-determined

molecular weight cut-off and centrifuge at 4000 g at 4 °C until the sample is concentrated in a final volume of 500 µL. Keep the protein in small aliquots (20 µL) at -80 °C. Check the concentration of the protein before freezing. Do not freeze-thaw the purified HEase.

3.13. *In vitro* endonuclease cleavage assay

1. Combine: 15 µL of substrate plasmid (25 µg/mL), 5 µL *in vitro* endonuclease reaction buffer supplemented with 1 mM DTT, 5 µL of HEase protein (~50 µg/mL) and 25 µL H_2O. In addition the linearized substrate plasmid can be tested as a substrate for the endonuclease activity.
2. Set up a parallel reaction as in step 3 but with the control plasmid which contains an insert that comprises the HEase/intron containing allele; a negative control that should not be cleaved by the HEase.
3. Incubate the cleavage reactions at 37 °C and 10 µL aliquots are taken at the following time intervals 0, 30 and 60 minutes; stop the reactions by adding 2 µL of 200 mM EDTA (pH 8.0) and 1 µL of proteinase K (1 mg/mL) to each 10 µl aliquots followed by incubation for 30 minutes at 37 °C.
4. Resolve the cleavage reaction products on a 1% agarose gel; in addition samples representing an untreated version of the substrate; ideally a restriction enzyme linearized version, and the control (negative control) plasmid(s) should be resolved on this gel along with a suitable molecular weight marker.

3.14. Cleavage site mapping

1. Treat substrate plasmid with HEase under optimal conditions (as outlined in section 3.13).

2. Resolve the cleaved substrate plasmid on a 1% agarose gel and excise the DNA fragment from the gel with any suitable PCR product gel clean-up/extraction system.
3. Treat the linearized substrate plasmid with T4 DNA polymerase under conditions that generate blunt ends (Bae *et al*., 2009); reaction mixture contains 40 µL of HEase treated linearized plasmid (25 µg/mL), 2 µL T4 DNA polymerase (5u/µL), 20 µL 5 x T4 DNA polymerase buffer, 20 µL dNTP mixture (0.5 mM) and the total volume is adjusted to 100 µL with sterile distilled water.
4. Incubate the reaction mixture at room temperature (~24 °C) for 20 minutes and place on ice for 5 minutes and terminate the reaction by incubating for 10 minutes at 70 °C.
5. Purify the T4 DNA polymerase treated linearized DNA (now blunt ended) and add 2 µL of T4 DNA Ligase (1u/µL) in the presence of 10 µL 5 x Ligase buffer in a total volume of 40 µL. Incubate the ligation reaction at room temperature for 2 hours to generate the desired religated plasmid.
6. Dilute the ligation reaction 5-fold and use 10 µL of this dilution to transform chemical competent *E. coli* DH5α cells.
7. Transformed *E. coli* cells are grown overnight at 37 °C in 5 mL of LB media (supplemented with appropriate antibiotics).
8. Purify the plasmid from the transformed overnight cultures with a suitable plasmid purification kit (such as Wizard® Plus Minipreps DNA purification kit, Promega) and sequence the recovered plasmid using the BigDye® Terminator Cycle Sequencing Kit (Applied Biosystems) following the manufacturer's instructions.
9. Compare the chromatogram for the obtained sequence with the sequence for the original uncleaved substrate plasmid or sequence the uncleaved substrate plasmid in parallel with the

HEase cleaved/T4 DNA polymerase treated substrate plasmid using the same primers for both types of constructs.

10. Nucleotides missing in the sequence of the HEase/T4 DNA polymerase treated substrate plasmid when compared to the original untreated substrate sequence define the nucleotides removed by T4 DNA polymerase. This approach works for LAGLIDADG type HEases that typically generate 4 nucleotide 3′ overhangs at their cleavage sites, these staggered cuts are blunt ended by the T4 DNA polymerase (Stoddard, 2006; Gimble, 2000; Bae *et al*., 2009).

3.15. $MgCl_2$ as the trigger for the ribozyme switch needed for the *in vivo* HEase expression

We have noticed that manipulating the exogenous [Mg^{+2}] (i.e., in the media) stimulates group II intron splicing and thus the removal of the intron acts like a switch that can control the expression of the HEase. In order to evaluate the appropriate amount of Mg^{+2} (suggested range from 1 mM to 10 mM) to be added to the media an *in vivo* endonuclease assay has to be established. This assay is based on two-plasmid *in vivo* endonuclease assay (see Chapter 4, Figure 4.1 and Figure 4.2) has to be established where two compatible plasmids, a HEase "donor" plasmid ("ORF-switch" plasmid as the ORF contains a group II intron sequence) and a HEase "substrate" plasmid and both need to be maintained in *E. coli* BL21 (DE3). See section 3.8 on evaluating the splicing potential of the group II intron. The plasmids have different (compatible) origins of replication and can be selected for based on antibiotic selection [kanamycin (kan) and chloramphenicol (cam) respectively]. For example the pET28 b (+) vector (ColE1 origin of replication and kanR) can be used for the overexpression of the HEase and a second plasmid pACYC184 (ATCC 37033 - American Type Culture Collection, Manassas, VA, USA; p15A origin of replication and camR) can be used to provide the target site for the HEase.

The “donor” plasmid hosts the HEase ORF with the group II intron at an appropriate location to facilitate suitable EBS/IBS interactions and the “substrate” plasmid has a sequence inserted that offers the HEase a cleavage target site. Successful expression and production of the HEase will lead to the loss of the substrate plasmid and the kanR marker. The loss of cell viability is an indicator of intron splicing which will lead to the production of a functional HEase. To ascertain the group II intron as an on/off switch for *in vivo* HEase expression the media can be supplemented 10 µM of $CoCl_2$; Co^{+2} appears to negate the stimulator effect of exogenous Mg^{+2} on intron splicing, possible Co^{+2} interferes with the update of Mg^{+2} into *E. coli* cells (Nelson and Kennedy, 1971, 1972). In summary, the addition of Mg^{+2} stimulates the expression of the HEase and the addition of Co^{+2} is inhibitory for HEase expression.

1. Cotransform *E. coli* BL21 with two plasmids - one containing the “ORF-switch” (i.e., the HEase ORF plus group II intron) plasmid and the other being the substrate plasmid. Make sure the plasmids are compatible (different origin of replications) and also have different antibiotic selection markers (e.g. the “ORF-switch” plasmid has the kanamycin cassette while the substrate construct has the chloramphenicol cassette).
2. Repeat step 1 to cotransform control vector and substrate plasmid. (For chemical transformation see section 3.4). Store the positive cotransformed clones as 50% glycerol stocks.
3. Grow precultures overnight at 37 °C derived from the glycerol stocks (see above) in culture tubes containing 5 mL LB media plus appropriate antibiotics. Add 1% glucose to the media to prevent leaky expression (if T7 promoter containing vectors are used).
4. Inoculate 50 mL LB broth (containing appropriate antibiotics, 1% glucose) with 500 µL from the 5 mL precultures. Add the pre-determined concentrations of $MgCl_2$ to the media. Label the

flask as LB + Mg^{+2}. For additional 50 mL LB flasks inoculate with the same amount of preculture and keep all supplements constant but do not add $MgCl^{+2}$. Label this culture flask as LB – Mg^{+2} (negative control).

5. In order to antagonize the stimulatory effect of $MgCl_2$ on splicing of group II introns, in one LB culture flask add 10 µM of $CoCl_2$ along with the desired concentration of $MgCl_2$. Label this flask as LB + Mg^{+2} + Co^{+2}.
6. Grow the cells at 37 °C with vigorous shaking (210 rpm) till the O.D.600 reaches 0.65.
7. Induce protein overexpression with the pre-determined concentration of IPTG in LB + Mg^{+2}, LB – Mg^{+2} culture and LB + Mg^{+2} + Co^{+2} culture flasks.
8. Incubate the flasks with vigorous shaking (210 rpm) at the pre-determined (optimal) temperature for at least 4 to 6 hours.
9. Perform a serial dilution for each of the above cultures and plate the diluted cells (10^{-6}) on pre warmed (37 °C) LB agar plates containing only the antibiotic that was selected for by the substrate plasmid (e.g., if the substrate plasmid contained chloramphenicol, plate the diluted cells on the LB agar chloramphenicol plates). Perform at least two biological and three technical replicates for each of the above cultures. Incubate the plates at 37 °C until the colonies are clearly visible and count the number of colonies in order to get mean cfu/mL values and standard deviations. This will establish suitable parameters for setting up conditions for the temporal expression of a HEase that could cut an intended target during a specific growth phase of the bacterium.

S7.2.4. Notes

1. In order to find potential HEGs see article by Hafez *et al*. (2013). Also see the LAGLIDADG

homing endonuclease database (Taylor *et al*., 2012).

2. Use group IIC intron with caution as these tend to have three IBS/EBS interactions for establishing splicing competent folds this can complicate the design of the construct. We use group IIA and IIB introns as they tend to have fewer IBS/EBS interactions.
3. Some group II introns require one IBS (IBS1), some require two IBS sequences (IBS1 and IBS2) while some group II intron categories require three sets of sequences to satisfy all IBS/EBS interactions for proper splicing (Lambowitz and Belfort, 2015).
4. Example: First one has to determine with great degree of certainty what the group II introns EBS sequences are, from here one can proceed and scan the HEase ORF sequence for a location that provides compatible (complementary) IBS sequences, keeping in mind that with regards to RNA U can interact with A or G. If a group II intron has been selected and it requires an IBS1 sequences that is aacagg, one would scan the nucleotide sequences of the HEase ORF and try to locate a match for this sequence. If the sequence is found in the middle (or near middle) of the ORF sequence, that should be an ideal location for inserting a group II intron. Keep in mind there might be additional IBS sequences required (IBS2 etc.). It is important to find a suitable location in the HEase ORF sequence that would maintain all the required IBS/EBS interactions with minimal modification to the HEase coding sequence.
5. Several online programs assist in codon optimization e.g., http://www.encorbio.com/protocols/Codon.htm, http://genomes.urves/OPTIMIZER/. Several commercial outfits will perform codon optimization and gene synthesis such as GenScript (http://www.genescript.com/), GeneArt (Thermo Fisher Scientific), Gene Oracle (Sigma-Aldrich), etc.
6. The latter plasmid should not be cleaved by the HEase as the cleavage site is disrupted by the

HEase/intron sequence. One could also obtain substrates and controls by using PCR products of alleles that lack the HEase/intron insertion and alleles that contain the HEase/intron; however some HEases appear to prefer plasmid DNAs as substrates (i.e., supercoiled templates).

7. Evaluating for potential inserts within the pUC57 vector, use M13 Forward primer (M13F) and M13 Reverse primer (M13R). One must use the respective vector specific primers to sequence the insert to confirm that the correct sequence is present.
8. *E. coli* BL21 is specifically designed for the over-expression of genes regulated by the T7 promoter. However, DO NOT use this strain for the propagation and maintenance of plasmids as this strain has leaky T7 RNA polymerase expression, which might lead instability and eventual loss of the plasmid.
9. Sometimes incubation for 1 hour in ice leads to better transformation compared to 5 minutes.
10. Check the $O.D._{600}$ of the cultures in order to see whether high concentrations of $MgCl_2$ are detrimental to the bacterial cell growth or not.
11. Presence of PCR products indicates residual DNA contamination.
12. It is always possible to observe RT-PCR product that were generated due to the presence of unspliced transcripts or some other splicing intermediates.
13. Alternative splicing is a possibility and such an event might happen if alternative EBS/IBS interactions can be established; this would also shift the intron/exon junction and thus could alter the coding sequence.
14. Although the PURExpress kit is designed for coupled transcription/translation from an expression construct containing T7 promoter, direct translation is also possible provided purified RNA (1 µg-5 µg) with a proper ribosome binding site (RBS) is incubated within the

in vitro translation reaction mixture.

15. Imidazole concentration in the wash buffer should be adjusted dependent on the affinity of the protein to the nickel resin.

References

Abu-Amero SN, Charter NW, Buck KW, Brasier, CM. 1995. Nucleotide-sequence analysis indicate that a DNA plasmid in a diseased isolate of *Ophiostoma novo-ulmi* is derived by recombination between two long repeat sequences in the mitochondrial large subunit ribosomal RNA gene. Curr. Genet. **28**: 54-59.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389-3402.

Amlacher S, Sarges P, Flemming D, van Noort V, Kunze R, Devos DP, Arumugam M, Bork P, Hurt E. 2011. Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile. Cell. **146**: 277-289.

Anraku Y, Hirata R, Wada Y, Ohya Y. 1992. Molecular genetics of the yeast vacuolar H (+)-ATPase. J. Exp. Biol. **172**: 67-81.

Arnould S, Chames P, Perez C, Lacroix E, Duclert A, Epinat JC, Stricher F, Petit AS, Patin A Guillier S. 2006. Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. J. Mol. Biol. **355**: 443-458.

Arnould S, Delenda C, Grizot S, Desseaux C, Pâques F, Silva GH, Smith J. 2011. The I-CreI meganuclease and its engineered derivatives: applications from cell modification to gene therapy. Protein Eng. Des. Sel. **24**: 27-31.

Arnould S, Perez C, Cabaniols JP, Smith J, Gouble A, Grizot S, Epinat JC, Duclert A, Duchateau P, Pâques F. 2007. Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. J. Mol. Biol. **371**(1): 49-65.
Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, Baker D. 2006. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature. **441**: 656-659.

Ashworth J, Taylor GK, Havranek JJ, Quadri SA, Stoddard BL and Baker D. 2010. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. Nucleic Acids Res. **38**(16): 5601-5608.

Aubert M, Ryu BY, Banks L, Rawlings DJ, Scharenberg AM, Jerome KR. 2011. Successful Targeting and Disruption of an Integrated Reporter Lentivirus Using the Engineered Homing Endonuclease Y2 I-AniI. PLoS One. **6**(2): e16825.

Back E, Van Meir E, Müller F, Schaller D, Neuhaus H, Aeby P, Tobler H. 1984. Intervening sequences in the ribosomal RNA genes of *Ascaris lumbricoides*: DNA sequences at junctions and genomic organization. EMBO J. **3**: 2523-2529.

Bae H, Kim KP, Song JM, Kim JH, Yang JS, Kwon ST. 2009. Characterization of intein homing endonuclease encoded in the DNA polymerase gene of Thermococcus marinus. FEMS Microbiol. Lett. **297**: 180-188.

Baidyaroy D, Hausner G, Hafez M, Michel F, Fulbright D, Bertrand H. 2011. Detection of a 973 bp insertion within the mtDNA rns gene in a mitochondrial hypovirulent strain of *Cryphonectria parasitica* isolated from nature. Fungal Genet. Biol. **48**: 775-783.

Baltes NJ, Gil-Humanes J, Cermak T, Atkins PA, Voytas DF. 2014. DNA Replicons for Plant Genome Engineering. Plant Cell. **26**(1): 151-163.

Barrangou R. 2013. CRISPR-Cas systems and RNA-guided interference. Wiley Interdiscip. Rev. RNA. **4**(3): 267-278.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science. **315**: 1709-1712.

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.

Baxter S, Lambert AR, Kuhar R, Jarjour J, Kulshina N, Parmeggiani F, Danaher P, Gano J, Baker D, Stoddard BL, Scharenberg AM. Engineering domain fusion chimeras from I-OnuI family LAGLIDADG homing endonucleases. Nucleic Acids Res. **40**: 7985-8000.

Beaudet D, Nadimi M, Iffis B, Hijri M. 2013. Rapid mitochondrial genome evolution through invasion of mobile elements in two closely related species of arbuscular mycorrhizal fungi. PLoS One **8**: e60768.

Belcour L, Rossignol M, Koll F, Sellem CH, 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**: 308-317.

Belfort M. 2003. Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. Genes Dev. **17**: 2860-2863.

Belfort M, Bonocora RP. 2014. Homing endonucleases: from genetic anomalies to programmable genomic clippers. Methods Mol. Biol.**1123**: 1-26.

Belfort M, Derbyshire V, Cousineau B, Lambowitz A. 2002. Mobile introns: pathways and proteins. In: Mobile DNA II. Craig N, Craigie R, Gellert M, Lambowitz A (eds.). ASM Press, New York. pp. 761-783.

Belfort M, Perlman PS. Mechanisms of intron mobility.1995. J. Biol. Chem. **270**: 30237-30240.

Belfort M, Reaban ME, Coetzee T and Dalgaard JZ. 1995. Prokaryotic introns and inteins: a panoply of form and function. J. Bacteriol. **177**(14): 3897-3903.

Belfort M and Roberts RJ. 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Res. **25**: 3379-3388.

Bell-Pedersen D, Quirk S, Clyman J and Belfort M. 1990. Intron mobility in phage T4 is dependent upon a distinctive class of endonucleases and independent of DNA sequences encoding the intron core: mechanistic and evolutionary implications. Nucleic Acids Res. **18**(13): 3763-3770.

Bell-Pedersen D, Quirk SM, Aubrey M, Belfort M. 1989. A site-specific endonuclease and co-conversion of flanking exons associated with the mobile td intron of phage T4. Gene. **82**(1):119-126.

BenJamaa ML, Lieutier F, Yart A, Jerraya A, Khouja ML. 2007. The virulence of phytopathogenic fungi associated with the bark beetles *Tomicus piniperda* and *Orthotomicus erosus* in Tunisia. Forest Pathol. **37**: 51-63.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. USA. **74**: 3171-3175.

Bhaya D, Davison M, Barrangou R. 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu. Rev. Genet. **45**: 273-297.

Biderre C, Méténier G, Vivarès CP. 1998. A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidia *Encephalitozoon cuniculi.* Mol. Biochem. Parasitol. **94**: 283-286.

Boch J, Scholze H, Schornack S, Landgraf A, Simone Hahn, Sabine Kay, Lahaye T, Nickstadt A, Bonas U. 2009. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science. **326**(5959): 1509-1512.

Bogdanove AJ, Voytas DF. 2011. TAL effectors: customizable proteins for DNA targeting. Science. **333**: 1843-1846.

Boissel S, Jarjour J, Astrakhan A, Adey A, Gouble A, Duchateau P, Shendure J, Stoddard BL, Certo MT, Baker D, Scharenberg AM. 2013. megaTALs: a rare-cleaving nuclease architecture for therapeutic genome engineering. Nucleic Acids Res. **42**(4): 2591-2601.

Bolduc JM, Spiegel PC, Chatterjee P, Brady KL, Downing ME, Caprara MG, Waring, RB and Stoddard BL. 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.

Bonen L. 2008. Cis- and trans-splicing of group II introns in plant mitochondria. Mitochondrion. **8**: 26-34.

Bonen L, Vogel J. 2001. The ins and outs of group II introns. Trends Genet. **17**: 322-331.

Bonocora RP, Shub DA. 2009. A likely pathway for formation of mobile group I introns. Curr. Biol. **19**: 223-228.

Bos JL, Heyting C, Borst P. 1978. An insert in the single gene for the large ribosomal RNA in yeast mitochondrial DNA. Nature. **275**: 336-338.

Bradley RW, Buck M, Wang B. 2015.Tools and Principles for Microbial Gene Circuit Engineering. J. Mol. Biol. doi.10.1016/j.jmb.2015.10.004.

Bryk M, Quirk SM, Mueller JE, Loizos N, Lawrence C, Belfort M. 1993. The td intron endonuclease I-TevI makes extensive sequence-tolerant contacts across the minor groove of its DNA target. EMBO J. **12**(5): 2141-2149.

Bryk M, Belisle M, Mueller JE, Belfort M. 1995. Selection of a remote cleavage site by I-TevI, the td intron-encoded endonuclease. J. Mol. Biol. **247**(2): 197-210.

Budman J, Chu G. 2005. Processing of DNA for nonhomologous end-joining by cell-free extract. EMBO J. **24**(4): 849-860.

Busk PK, Lange L. 2013. Cellulolytic potential of thermophilic species from four fungal orders. AMB Express. **3**: 47.

Calos MP. 2016. The CRISPR Way to Think about Duchenne's. N. Engl. J. Med. **374**(17): 1684-1686.

Calvin K, Li H. 2008. RNA splicing endonuclease structure and function. Cell Mol. Life. Sci. **65**: 1176-1185.

Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. 2012. Database for bacterial group II introns. Nucleic Acids Res (Database issue). **40**: D187-190.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR. 2002. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs. BMC Bioinformatics. **3**: 2.

Capecchi MR. 1980. High efficiency transformation by direct microinjection of DNA into cultured mammalian cells. Cell. **22**: 479-488.

Caprara MG, Waring RB. 2005. Group I introns and their maturases: uninvited, but welcome guests. In: Homing endonucleases and inteins. Belfort M, Derbyshire V, Stoddard BL, Wood DL (eds). Springer Press, New York. pp. 103-119.

Carter JM, Friedrich NC, Kleinstiver B, Edgell DR. 2007. Strand-specific contacts and divalent metal ion regulate double-strand break formation by the GIY-YIG homing endonuclease I-BmoI. J. Mol. Biol. **374**: 306-321.

Cathomen T, Joung JK. 2008. Zinc-finger nucleases: the next generation emerges. Mol. Ther. **16**(7): 1200-1207.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. Trends Genet. **7**: 145-148.

Cech TR. 1990. Self-splicing of group-I introns. Annu. Rev. Biochem. **55**: 599-629.

Certo MT, Gwiazda KS, Kuhar R, Sather B, Curinga G, Mandt T, Brault M, Lambert AR, Baxter SK, Jacoby K, Ryu BY, Kiem HP, Gouble A, Paques F, Rawlings DJ, Scharenberg AM. 2012. Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. Nat. Methods. **9**: 973-975.

Chan YS, Naujoks DA, Huen DS, Russell S. 2011. Insect population control by homing endonuclease-based gene drive: an evaluation in Drosophila melanogaster. Genetics. **188**: 33-44.

Chan YS, Takeuchi R, Jarjour J, Huen DS, Stoddard BL, Russell S. 2013. The design and in vivo evaluation of engineered I-OnuI-based enzymes for HEG gene drive. PLoS ONE. **8**: e74254.

Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat. Rev. Microbiol. **11**: 525-538.

Chapdelaine P, Pichavant C, Rousseau J, Paques F, Tremblay JP. 2010. Meganucleases can restore the reading frame of a mutated dystrophin. Gene Ther. **17**: 846-858.

Charter NW, Buck KW, Brasier CM. 1996. Multiple insertions and deletions determine the size differences between the mitochondrial DNAs of the EAN and NAN races of *Ophiostoma novo-ulmi*. Mycol. Res. **100**: 368-378.

Chatterjee P, Brady KL, Solem A, Ho Y, Caprara MG. 2003. Functionally distinct nucleic acid binding sites for a group I intron encoded RNA maturase/DNA homing endonuclease. J. Mol Biol. **329**: 239-251.

Chen Z, Zhao H. 2005. A highly sensitive selection method for directed evolution of homing endonucleases. Nucleic Acids Res. **33**: e154.

Chen S, Oikonomou G, Chiu CN, Niles BJ, Liu J, Lee DA, Antoshechkin I, Prober DA. 2013. A large-scale in vivo analysis reveals that TALENs are significantly more mutagenic than ZFNs generated using context-dependent assembly. Nucleic acids Res. **41**: 2769-2778.

Chevalier BS, Monnat RJ Jr, Stoddard BL. 2001. The homing endonuclease I-CreI uses three metals, one of which is shared between the two active sites. Nat. Struct. Biol. **8**(4): 312-316.

Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, Stoddard BL. 2002. Design, activity and structure of a highly specific artificial endonuclease. Mol. Cell. **10**(4): 895-905.

Chevalier BS, Turmel M, Lemieux C, Monnat RJ Jr, Stoddard BL. 2003. Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. J. Mol. Biol. **329**(2): 253-269.

Chevalier BS, Sussman D, Otis C, Noël AJ, Turmel M, Lemieux C, Stephens K, Monnat RJ Jr, Stoddard BL. 2004. Metal-dependent DNA cleavage mechanism of the I-CreI LAGLIDADG homing endonuclease. Biochemistry. **43**(44): 14015-14026.

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

Choudhary PK, Sigel RK. 2014. Mg^{2+}-induced conformational changes in the btuB riboswitch from E. coli. RNA. **20**: 36-45.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. **12**: 1-8.

Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ and Voytas DF. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics. **1862**: 757-761.

Choulika A, Perrin A, Dujon B, Nicolas JF. 1995. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol. Cell Biol. **15**(4): 1968-1973.

Clark-Walker GD. 1992. Evolution of mitochondrial genomes in fungi. Int. Rev. Cytol. **141**: 89-127.

Clyman J, Belfort M. 1992. Trans and cis requirements for intron mobility in a prokaryotic system. Genes Dev. **6**(7): 1269-1279.

Colleaux L, d'Auriol L, Betermier M, Cottarel G, Jacquier A, Galibert F, Dujon B. 1986. Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into *E. coli* as a specific double strand endonuclease. Cell. **44**(4): 521-533.

Colleaux L, D'Auriol L, Galibert F, Dujon B. 1988. Recognition and cleavage site of the intron-encoded omega transposase. Proc. Natl. Acad. Sci. USA. **85**(16): 6022-6026.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. 2013. Multiplex genome engineering using CRISPR/Cas systems. Science. **339**: 819-823.

Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. Nature. **404**: 1018-1021.

Cox DBT, Platt RJ, Zhang F. 2015. Therapeutic genome editing: prospects and challenges. Nat. Medicine. **21**: 121-131.

Cui X, Davis G. 2007. Mobile group II intron targeting: applications in prokaryotes and perspectives in eukaryotes. Front. Biosci.**12**: 4972-4985.

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.

Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. Cell. **84**(1): 9-12.

Daboussi, F, Zaslavskiy M, Poirot L. Loperfido M, Gouble A, Guyot V, Leduc S, Galetto R, Grizot S, Oficjalska D, Perez C, Delacôte F, Dupuy A, Chion-Sotinel I, Le Clerre D, Lebuhotel C, Danos O, Lemaire F, Oussedik K, Cédrone F, Epinat JC, Smith J, Yáñez-Muñoz RJ, Dickson G, Popplewell L, Koo T, VandenDriessche T, Chuah MK, Duclert A, Duchateau P, Pâques F. 2012. Chromosomal context and epigenetic mechanisms control the efficacy of genome editing by rare-cutting designer endonucleases. Nucleic Acids Res. **40**: 6367-6379.

Dafa'alla TH, Hobom G, Zahner H. 2000. Direct Colony Identification by PCR-Miniprep. Mol Biol. Today. **1**: 65-66.

Dalgard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS. 1997. Statistical modeling and analysis of the LAGLIDADG family of site specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the H-N-H family. Nucleic Acids Res. **25**: 4626-4638.

Dassa B, London N, Stoddard BL, Schueler-Furman O, Pietrokovski S. 2009. Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. Nucleic Acids Res. **37**(8): 2560-2573.

Davé UP, Akagi K, Tripathi R, Cleveland SM, Thompson MA, Yi M, Stephens R, Downing JR, Jenkins NA, Copeland NG. 2009. Murine leukemias with retroviral insertions at Lmo2 are predictive of the leukemias induced in SCID-X1 patients following retroviral gene therapy. PLoS Genet. **5**(5): e1000491.

Davis KM, Pattanayak V, Thompson DB, Zuris JA, Liu DR. 2015. Small molecule-triggered Cas9 protein with improved genome-editing specificity. Nat. Chem. Biol. **11**: 316-318.

Dawkins R. 1976. The Selfish Gene. Oxford University Press.

Delacôte F, Perez C, Guyot V, Duhamel M, Rochon C, Ollivier N, Macmaster R, Silva GH, Pâques F, Daboussi F, Duchateau P. 2013. High frequency targeted mutagenesis using engineered endonucleases and DNA-end processing enzymes. PLoS One. **8**(1): e53217.

Deredec A, Burt A, Godfray HC. 2008. The population genetics of using homing endonuclease genes in vector and pest management. Genetics. **179**: 2013-2026.

Deredec A, Godfray HC, Burt A. 2011. Requirements for effective malaria control with homing endonuclease genes. Proc. Natl. Acad. Sci. USA. **108**(43): 874-880.

Derewenda ZS, Vekilov PG. 2006. Entropy and surface engineering in protein crystallization. Acta Crystallogr D Biol Crystallogr. **62**(Pt 1): 116-124.

Derbyshire V, Kowalski JC, Dansereau JT, Hauer CR, Belfort, M. 1997. Two-domain structure of the td intron-encoded endonuclease I-TevI correlates with the two-domain configuration of the homing site. J. Mol. Biol. **265**(5): 494-506.

di Guan C, Li P, Riggs PD, Inouye H. 1988. Vectors that facilitate the expression and purification of foreign peptides in Escherichia coli by fusion to maltose-binding protein. Gene. **67**(1): 21-30.

Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. Proc. Natl. Acad. Sci. USA. **98**: 13207-13212.

Donghi D, Pechlaner M, Finazzo C, Knobloch B, Sigel RKO. 2013. The structural stabilization of the κ three-way junction by Mg (II) represents the first step in the folding of a group II intron. Nucleic Acids Res. **41**: 2489-2504.

Doyon JB, Pattanayak V, Meyer CB, Liu DR. 2006. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. J. Am. Chem. Soc. **128**: 2477-2484.

Drager RG, Hallick RB. 1993. A complex twintron is excised as four individual introns. Nucleic Acids Res. **21**: 2389-2394.

Drouin M, Lucas P, Otis C, Lemieux C, Turmel M. 2000. Biochemical characterization of I-CmoeI reveals that this H-N-H homing endonuclease shares functional similarities with H-N-H colicins. Nucleic Acids Res. **28**(22): 4566-4572.

Duan X, Gimble FS, Quiocho FA. 1997. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. Cell. **89**(4):555-564.

Dujon B. 1980. Sequence of the intron and flanking exons of the mitochondrial 21S rRNA gene of yeast strains having different alleles at the omega and rib-1 loci. Cell. **20**(1): 185-197.

Dujon B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations-a review. Gene. **82**(1): 91-114.

Dujon B, Belcour L. 1989. Mitochondrial DNA instabilities and rearrangements in yeasts and fungi. In: Berg DE, Howe MM (eds.). Mobile DNA. ASM, Washington DC. pp. 861-878.

Dujon B, Colleaux L, Jacquier A, Michel F, Monteilhet C. 1986. Mitochondrial introns as mobile genetic elements: the role of intron-encoded proteins. Basic Life Sci. **40**: 5-27.

Dujon B, Slonimski PP, Weill L. 1974. Mitochondrial genetics IX: A model for recombination and segregation of mitochondrial genomes in *Saccharomyces cerevisiae*. Genetics. **78**(1): 415-437.

Dunin-Horkawicz S, Feder M, Bujnick JM. 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. BMC Genomics. **7**: 98.

Duong-Ly KC, Gabelli SB. 2014. Explanatory chapter: troubleshooting recombinant protein expression: general. Methods Enzymol. **541**: 209-229.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature. **284**: 601-603.

Eddy SR, Gold L. 1991. The phage T4 *nrdB* intron: a deletion mutant of a version found in the wild. Genes Dev. **5**(6): 1032-1041.

Edgell DR, Belfort M, Shub DA. 2000. Barriers to intron promiscuity in bacteria. J. Bacteriol. **182**: 5281-5289.

Edgell DR, Derbyshire V, Van Roey P, LaBonne S, Stanger MJ, Li Z, Boyd TM, Shub DA, Belfort M. 2004. Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. Nat. Struct. Mol. Biol. **11**(10): 936-944.

Edgell DR, Derbyshire V, Van Roey P, LaBonne S, Stanger MJ, Li Z, Boyd TM, Shub DA, Belfort M. 2004. Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. Nat. Struct. Mol. Biol. **11**: 936-944.

Edgell DR, Chalamcharla VR, Marlene Belfort. 2011. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biology. **9**: 22.

Eisenschmidt K, Lanio T, Simoncsits A, Jeltsch A, Pingoud V, Wende W, Pingoud A. 2005. Nucleic Acids Res. **33**: 7039-7047.

Einvik C, Elde M, Johansen S. 1998a. Group I twintrons: genetic elements in myxomycete and schizopyrenid amoeboflagellate ribosomal DNAs. J. Biotechnol. **64**: 63-74.

Einvik C, Nielsen H, Westhof E, Michel F, Johansen S. 1998b. Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. RNA. **4**: 530-541.

Elbadawy HM, Gailledrat M, Desseaux C, Salvalaio G, Di Iorio E, Ferrari B, Bertolin M, Barbaro V, Parekh M, Gayon R, Munegato D, Franchin E, Calistri A, Palù G, Parolin C, Ponzin D, Ferrari S. 2014. Gene transfer of integration defective anti-HSV-1 meganuclease to human corneas *ex vivo*. Gene Ther. **21**(3):272-281.

Elde M, Haugen P, Willassen NP, Johansen S. 1999. I-NjaI, a nuclear intron-encoded homing endonuclease from Naegleria, generates a pentanucleotide 3' cleavage-overhang within a 19 base-pair partially symmetric DNA recognition site. Eur. J. Biochem. **259**(1-2):281-288.

Elde M, Willassen NP, Johansen S. 2000. Functional characterization of isoschizomeric His-Cys box homing endonucleases from Naegleria. Eur. J. Biochem. **267**(24): 7257-7266.

Ellis BL, Hirsch ML, Barker JC, Connelly JP, Steininger RJ III, Porteus MH. 2013. A survey of *ex vivo/in vitro* transduction efficiency of mammalian primary cells and cell lines with nine natural adeno-associated virus (AAV1-9) and one engineered adeno-associated virus serotype. Virol. J. **10**: 74.

Ellison EL, Vogt VM. 1993. Interaction of the intron-encoded mobility endonuclease I-PpoI with its target site. Mol. Cell Biol. **13**(12): 7531-7539.

Enyeart PJ, Mohr G, Ellington AD, Lambowitz AM. 2014. Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. Mob DNA. **5**: 21.

Fauser F, Rotha N, Pachera M, Ilga G, Sánchez-Fernández R, Biesgen C, Puncta H. 2012. In planta gene targeting. Proc. Natl. Acad. Sci. USA. **109**(19): 7535-7540.

Faye G, Dennebouy N, Kujawa C, Jacq C. 1979. Inserted sequence in the mitochondrial 23S ribosomal RNA gene of the yeast *Saccharomyces cerevisiae*. Mol. Gen. Genet. **168**(1): 101-109.

Férandon C, Moukha S, Callac P, Benedetto JP, Castroviejo M, Barroso G. 2010. The Agaricus bisporus cox1 gene: the longest mitochondrial gene and the largest reservoir of mitochondrial group I introns. PLoS One. **5**: e14048.

Fedorova O. 2012. Kinetic characterization of group II intron folding and splicing. Methods Mol. Biol. **848**: 91-111.

Fedorova O, Zingler N. 2007. Group II introns: structure, folding and splicing mechanism. Biol. Chem. **388**: 665-678.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. **39**: 783-791.

Férandon C, Moukha S, Callac P, Benedetto JP, Castroviejo M, Barroso G. 2011. The *Agaricus bisporus cox1* gene: the longest mitochondrial gene and the largest reservoir of mitochondrial group I introns. PLoS One **5**: e14048.

Fiskaa T, Birgisdottir AB. 2010. RNA reprogramming and repair based on trans-splicing group I ribozymes. N. Biotechnol. **27**(3): 194-203.

Flannagan RS, Linn T, Valvano MA. 2008. A system for the construction of targeted unmarked gene deletions in the genus Burkholderia. Environ. Microbiol. **10**(6): 1652-1660.

Flick KE, Jurica MS, Monnat RJ Jr, Stoddard BL. 1998. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. Nature. **394**: 96-101.

Flipphi M, Fekete E, Ag N, Scazzocchio C, Karaffa L. 2013. Spliceosome twin introns in fungal nuclear transcripts. Fungal Genet. Biol. **57**: 48-57.

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.

Formey D, Molès M, Haouy A, Savelli B, Bouchez O, Bécard G, Roux C. 2012. Comparative analysis of mitochondrial genomes of *Rhizophagus irregularis* - syn. *Glomus irregulare* - reveals a polymorphism induced by variability generating elements. New Phytol. **196**: 1217-1227.

Friedhoff P, Franke I, Meiss G, Wende W, Krause KL, Pingoud A. 1999. A similar active site for non-specific and specific endonucleases. Nat. Struct. Biol. **6**: 112-113.

Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat. Biotechnol. **32**(3): 279-284.

Fujisawa T, Narikawa R, Okamoto S, Ehira S, Yoshimura H, Suzuki I, Masuda T, Mochimaru M, Takaichi S, Awai K, Sekine M, Horikawa H, Yashiro I, Omata S, Takarada H, Katano Y, Kosugi H, Tanikawa S, Ohmori K, Sato N, Ikeuchi M, Fujita N, Ohmori M. 2010. Genomic structure of an economically important cyanobacterium, *Arthrospira (Spirulina) platensis* NIES-39. DNA Res. **17**: 85-103.

Gaj T, Gersbach CA, Barbas CF 3rd. 2013. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. **31**(7): 397-405.

Gao H, Smith J, Yang M, Jones S, Djukanovic V, Nicholson MG, West A, Bidney D, Falco SC, Jantz D, Lyznik LA. 2010. Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J. **61**: 176-187.

Gasiunas G, Sasnauskas G, Tamulaitis G, Urbanke C, Razaniene D, Siksnys V. 2008. Tetrameric restriction enzymes: expansion to the GIY-YIG nuclease family. Nucleic Acids Res. **36**(3): 938-949.

Gibb EA, Hausner G. 2005. Optional mitochondrial introns and evidence for a homing endonuclease gene in the mtDNA rnl gene in *Ophiostoma ulmi sensu lato*. Myc. Res. **109**: 1112-1126.

Gilbert W. 1978. Why genes in pieces? Nature. **271**(5645): 501.

Gillham NW. 1994. Organelle Genes and Genomes, Oxford University Press, New York.

Gimble FS. 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. FEMS Microbiol. Lett. **185**(2): 99-107.

Gimble FS. 2005. Engineering homing endonucleases for genomic applications. In: Belfort, M., Derbyshire, V., Stoddard, B.L., Wood D.L. (eds.), Homing endonucleases and inteins. Springer Verlag, New York, NY. pp. 177-192.

Gimble FS. 2007. Engineering homing endonucleases to modify complex genomes. Gene Therapy and Regulation. **3**: 33-50.

Gimble FS, Moure CM and Posey KL. 2003. Assessing the plasticity of DNA target site recognition of the PI-SceI homing endonuclease using a bacterial two-hybrid selection system. Mol. Biol. **334**(5): 993-1008.

Giraldo-Perez P, Goddard MR. 2013. A parasitic selfish gene that affects host promiscuity. Proc. Biol. Sci. **280**(1770): 20131875.

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

Gogarten JP, Hilario E. 2006. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol. Biol. **6**: 94.

Goodrich-Blair H, Scarlato V, Gott JM, Xu MQ, Shub DA. 1990. A self-splicing group I intron in the DNA polymerase gene of *Bacillus subtilis* bacteriophage SPO1. Cell. **63**(2): 417-424.

Goodrich-Blair H, Shub DA. 1994. The DNA polymerase genes of several HMU-bacteriophages have similar group I introns with highly divergent open reading frames. Nucleic Acids Res. **22**(18): 3715-3721.

Goodrich-Blair H, Shub DA. 1996. Beyond homing: competition between intron endonucleases confers a selective advantage on flanking genetic markers. Cell. **84**(2): 211-221.

Gopal GJ, Kumar A. 2013. Strategies for the production of recombinant protein in *Escherichia coli*. Protein J. **32**(6): 419-425.

Gorbalenya AE. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. Protein Sci. **3**(7):1117-1120.

Gordon PM, Piccirilli JA. 2001. Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. Nat. Struct. Biol. **8**: 893-898.

Gordon PM, Fong R, Piccirilli JA. 2007. A second divalent metal ion in the group II intron reaction center. Chem. Biol. **14**: 607-612.

Gordon JW, Ruddle FH. 1981. Integration and stable germ line transmission of genes injected into mouse pronuclei. Science. **214**:1244-1246.

Gorton C, Kim SH, Henricot B, Webber J, Breuil C. 2004. Phylogenetic analysis of the blue stain fungus Ophiostoma minus based on partial ITS rDNA and b-tubulin gene sequences. Mycol. Res. **108**: 759-765.

Gorton C, Webber JF. 2000. Re-evaluation of the status of the blue stain fungus and bark beetle associate Ophiostoma minus. Mycologia. **92**: 1071-1079.

Gouble A, Smith J, Bruneau S, Perez C, Guyot V, Cabaniols JP, Leduc S, Fiette L, Avé P, Micheau B, Duchateau P, Pâques F. 2006. Efficient in toto targeted recombination in mouse liver by meganuclease-induced double-strand break. J. Gene Med. **8**(5): 616-622.

Grasso V, Palermo S, Sierotzki H, Garibaldi A, Gisi U. 2006. Cytochrome b gene structure and consequences for resistance to Qo inhibitor fungicides in plant pathogens. Pest Manag. Sci. **62**: 465-472.

Green MR, Sambrook R. 2012. Molecular Cloning, A laboratory manual. 4th edition. Cold Spring Harbor Laboratory Press.

Greenfield NJ. 2006. Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. **1**: 2876-2890.

Grizot S, Smith J, Daboussi F, Prieto J, Redondo P, Merino N, Villate M, Thomas, S, Lemaire L, Montoya G, Blanco FJ, Pâques F, Duchateau P. 2009. Efficient targeting of a SCID gene by an engineered single chain homing endonuclease. Nucleic Acids Res. **37**: 5405-5419.

Grizot S, Epinat JC, Thomas S, Duclert A, Rolland S, Pâques F, Duchateau P. 2010. Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. Nucleic Acids Res. **38**(6): 2006-2018.

Gruen M, Chang K, Serbanescu I, Liu DR. 2002. An in vivo selection system for homing endonuclease activity. Nucleic Acids Res. **30**(7): e29.

Guha TK, Hausner G. 2014. A homing endonuclease with a switch: characterization of a twintron encoded homing endonuclease. Fungal Genet. Biol. **65**: 57-68.

Guha TK, Hausner G. 2016. Using Group II Introns for Attenuating the *In Vitro* and *In Vivo* Expression of a Homing Endonuclease. PLoS ONE. **11**(2): e0150097.

Gao H, Smith J, Yang M, Jones S, Djukanovic V, Nicholson MG, West A, Bidney D, Carl Falco S, Jantz D, Lyznik LA. 2010. Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J. **61**: 176-187.

Guo J, Gaj T, Barbas CF 3rd. 2010. Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. J. Mol. Biol. **400**: 96-107.

Guo H, Karberg M, Long M, Jones, JP, Sullenger B, Lambowitz AM. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. Science, **289**(5478): 452-457.

Hafez M, Guha TK, Hausner G. 2014. I-OmiI and I-OmiII: Two intron-encoded homing endonucleases within the *Ophiostoma minus rns* gene. Fungal Biol. **118**(8): 721-731.

Hafez M, Guha TK, Shen C, Sethuraman J, Hausner G. 2014. PCR-based bioprospecting for homing endonucleases in fungal mitochondria rRNA genes. In: Edgell DR (ed), Methods Mol. Biol. **1123**: 37-53.

Hafez M, Hausner G. 2011a. The highly variable mitochondrial small subunit ribosomal RNA gene of *Ophiostoma minus*. Fungal Biol. **115**: 1122-1137.

Hafez M, Hausner G. 2011b. Characterization of the O.ul-mS952 intron: a potential molecular marker to distinguish between Ophiostoma ulmi and *Ophiostoma novo-ulmi* subsp. *americana*. World Acad. Sci. Eng. Technol. **59**: 1767-1775.

Hafez M, Hausner G. 2012. Homing endonucleases: DNA scissors on a mission. *Genome.* **55**(8): 553-569.

Hausner G, Hafez M, Edgell, DR. 2014. Bacterial group I introns: mobile RNA catalysts. *Mob DNA*. **5**: 8.

Hafez M, Majer A, Sethuraman J, Rudski SM, 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.

Hamari Z, Juhasz A, Gacser A, Kucsera J, Pfeiffer I, Kevei F. 2001. Intron mobility results in rearrangement in mitochondrial DNAs of heterokaryon incompatible *Aspergillus japonicus* strains after protoplast fusion. Fungal Genet. Biol. **33**: 83-95.

Haugen P, Bhattacharya D. 2004. The spread of LAGLIDADG homing endonuclease genes in rDNA. Nucleic Acids Res. **32**: 2049-2057.

Haugen P, Bhattacharya D, Palmer JD, Turner S, Lewis LA, Pryer KM. 2007. Cyanobacterial ribosomal RNA genes with multiple, endonuclease-encoding group I introns. BMC Evol. Biol. **7**: 159.

Haugen P, Reeb V, Lutzoni F, Bhattacharya D. 2004. The evolution of homing endonuclease genes and group I introns in nuclear rDNA. Mol. Biol. Evol. **21**: 129-140.

Haugen P, Simon DM, Bhattacharya D. 2005. The natural history of group I introns. Trends Genet. **21**: 111-119.

Hausner G. 2012. Introns, mobile elements and plasmids. In: Bullerwell, C.E. (Ed), Organelle Genetics: Evolution of Organelle Genomes and Gene Expression. Springer Verlag, Berlin, pp. 329-358.

Hausner G, Hafez M, Edgell DR. 2014. Bacterial group I introns: mobile RNA catalysts. Mobile DNA. **5**: 8.

Hausner G, Iranpour M, Kim JJ, Breuil C, Davis CN, Gibb EA, Reid J, Loewen PC, Hopkin AA. 2005. Fungi vectored by the introduced bark beetle *Tomicus piniperda* in Ontario, Canada and comments on the taxonomy of *Leptographium lundbergii*, *L. terebrantis*, *L. truncatum* and *L. wingfieldii*. Can. J. Bot. **83**: 1222-1237.

Hayakawa J, Ishizuka M. 2009. A group I self-splicing intron in the flagellin gene of the thermophilic bacterium *Geobacillus stearothermophilus*. Biosci. Biotechnol. Biochem. **73**: 2758-2761.

Hayakawa J, Ishizuka M. 2012. Temperature-dependent self-splicing group I introns in the flagellin genes of the thermophilic *Bacillus* species. Biosci. Biotechnol. Biochem. **76**: 410-413.

Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL. 1997. The structure of I-Crel, a group I intron-encoded homing endonuclease. Nat. Struct. Biol. **4**(6):468-476.

Heinemann I, Randau L, Tomko RJ Jr., Söll D. 2010. 3'-5' $tRNA^{His}$ Guanylyltransferase in Bacteria. FEBS Lett. **584**: 3567-3572.

Hensgens LA, Bonen L, de Haan M, van der Horst G, Grivell LA. 1983. Two intron sequences in yeast mitochondrial *cox1* gene: homology among URF-containing introns and strain-dependent variation in flanking exons. Cell. **32**(2): 379-389.

Herskowitz I, Rine J, Strathern J. 1992. Mating-type determination and mating-type interconversion in Saccharomyces cerevisiae. In The Molecular and Cellular Biology of the Yeast *Saccharomyces*: Gene Expression, Jones EW, Pringle JR, Broach JR (eds.) Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. pp. 583-656.

Hirose S, Noguchi T. 2013. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. Proteomics. **13**:1444-1456.

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.

Ho Y, Waring RB. 1999. The maturase encoded by a group I intron from *Aspergillus nidulans* stabilizes RNA tertiary structure and promotes rapid splicing. J. Mol. Biol. **292**: 987-1001.

Hoefel T, Faust G, Reinecke L, Rudinger N, Weuster-Botz D. 2012. Comparative reaction engineering studies for succinic acid production from sucrose by metabolically engineered *Escherichia coli* in fed-batch-operated stirred tank bioreactors. Biotechnol J. **7**: 1277-1287.

Hollander V, Kück U. 1999. Group II intron splicing in *Escherichia coli*: phenotypes of *cis*-acting mutations resemble splicing defects observed in organelle RNA processing. Nucleic Acids Res. **27**: 2339-2344.

Huchon D, Szitenberg A, Shefer S, Ilan M, Feldstein T. 2015. Mitochondrial group I and group II introns in the sponge orders Agelasida and Axinellida. BMC Evol. Biol. **15**: 278.

Ichiyanagi K, Ishino Y, Ariyoshi M, Komori K, Morikawa K. 2000. Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. J. Mol. Biol. **300**(4): 889-901.

Irimia M, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. Cold Spring Harbour Perspect Biol. **6**(6): a016071.

Ishijima S, Uda M, Hirata T, Shibata M, Kitagawa N, Sagami I. 2015. Magnesium uptake of *Arabidopsis* transporters, AtMRS2-10 and AtMRS2-11, expressed in *Escherichia coli* mutants: Complementation and growth inhibition by aluminum. Biochim. Biophys. Acta. **1848**: 1376-1382.

Jackson SA, Cannone JJ, Lee JC, Gutell RR, Woodson SA. 2002. Distribution of rRNA introns in the three-dimensional structure of the ribosome. J. Mol. Biol. **323**: 35-52.

Jacoby K, Metzger M, Shen BW, Certo MT, Jarjour J, Stoddard BL, Scharenberg AM. 2012. Expanding LAGLIDADG endonuclease scaffold diversity by rapidly surveying evolutionary sequence space. Nucleic Acids Res. **40**(11): 4954-4964.

Jacquier A, Dujon B. 1985. An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. Cell. **41**(2): 383-394.

Janice J, Jąkalski M, Makałowski M. 2013. Surprisingly high number of Twintrons in vertebrates. Biol. Direct. **8**: 4.

Jarjour J, West-Foyle H, Certo MT, Hubert CG, Doyle L, Getz MM, Stoddard BL, Scharenberg AM. 2009. High-resolution profiling of homing endonuclease binding and catalytic specificity using yeast surface display. Nucleic Acids Res. **37**(20): 6871-6880.

Jasin M. 1996. Genetic manipulation of genomes with rare-cutting endonucleases. Trends Genet. **12**(6): 224-228.

Jin Y, Binkowski G, Simon LD, Norris D. Ho endonuclease cleaves MAT DNA in vitro by an inefficient stoichiometric reaction mechanism. 1997. *J Biol Chem*. **272**(11): 7352-7359.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Scienc*e*. **337**(6096): 816-821.

Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA. 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science. **343**(6176): 1247997.

Johansen S, Embley TM, Willassen NP. 1993. A family of nuclear homing endonucleases. Nucleic Acids Res. **21**(18): 4405.

Johansen S, Elde M, Vader A, Haugen P, Haugli K, Haugli F. 1997. *In vivo* mobility of a group I twintron in nuclear ribosomal DNA of the myxomycete *Didymium iridis*. Mol. Microbiol. **24**(4): 737-745.

Jurica MS, Monnat RJ Jr, Stoddard BL. 1998. DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. Mol. Cell. **2**(4):469-476.

Jurica MS, Stoddard BL. 1999. Homing endonucleases: structure, function and evolution. Cell Mol. Life Sci. **55**(10): 1304-1326.

Kadyrov FA, Kriukov VM, Shliapnikov MG, Baev AA. 1994. SegE-a new site-specific endodeoxyribonuclease from bacteriophage T4. Dokl. Akad. Nauk. **339**(3): 404-406.

Kala S, Cumby N, Sadowski PD, Hyder BZ, Kanelis V, Davidson AR, Maxwell KL. 2014. HNH proteins are a widespread component of phage DNA packaging machines. Proc. Natl. Acad. Sci. USA. **111**(16): 6022-6027.

Kane PM, Yamashiro CT, Wolczyk DF, Neff N, Goebl M, Stevens TH. 1990. Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science. **250**(4981): 651-657.

Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM. 2001. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat. Biotech. **19**: 1162-1167.

Kellerman OK, Ferenci T. 1982. Maltose-binding protein from *E. coli*. Methods Enzymol. **90**: 459-463.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. **10**: 845-858.

Kelley LA, Sternberg MJE. 2009. Protein structure prediction on the web: A case study using the Phyre server. Nat. Protocols. **4**: 363-371.

Kelly S, Jess T, Price N. 2005. How to study proteins by circular dichroism. Biochim. et. Biophys. Acta. **1751**: 119-139.

Kennedy AB, Vowles JV, d'Espaux L, Smolke CD. 2014. Protein-responsive ribozyme switches in eukaryotic cells. Nucleic Acids Res. **42**(19): 12306-12321.

Khan H, Archibald JM. 2008. Lateral transfer of introns in the cryptophyte plastid genome. Nucleic Acids Res. **36**: 3043-3053.

Kim Y-G, Cha J, Chandrasegaran S. 1996. Hybrid restriction enzymes: zinc finger fusions to FokI cleavage domain. Proc. Natl. Acad. Sci. USA. **93**: 1156-1160.

Kim H, Kim JS. 2014. A guide to genome engineering with programmable nucleases. Nat. Rev. Genet. **15**: 321-334.

Klauser B, Atanasov J, Siewert LK, Hartig JS. 2015. Ribozyme-based aminoglycoside switches of gene expression engineered by genetic selection in S. cerevisiae. ACS Synth. Biol. **4**: 516-525.

Kleanthous C, Kühlmann UC, Pommer AJ, Ferguson N, Radford SE, Moore GR, James R, Hemmings AM. 1999. Structural and mechanistic basis of immunity toward endonuclease colicins. Nat. Struct. Biol. **6**(3):243-252.

Klein TA, Windbichler N, Deredec A, Burt A, Benedict MQ. 2012. Infertility resulting from transgenic I-PpoI male Anopheles gambiae in large cage trials. Pathog. Glob. Health. **106**(1): 20-31.

Kleinstiver BP, Wolfs JM, Edgell DR. 2013. The monomeric GIY-YIG homing endonuclease I-BmoI uses a molecular anchor and a flexible tether to sequentially nick DNA. Nucleic Acids Res. **41**(10): 5413-5427.

Ko TP, Liao CC, Ku WY, Chak KF, Yuan HS. 1999. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. Structure. **7**(1): 91-102.

Ko M, Choi H, Park C. 2002. Group I self-splicing intron in the recA gene of *Bacillus anthracis*. J Bacteriol. **184**: 3917-3922.

Kong S, Liu X, Fu L, Yu X, An C. 2012. I-PfoP3I: a novel nicking HNH homing endonuclease encoded in the group I intron of the DNA polymerase gene in *Phormidium foveolarum* phage Pf-WMP3. PLoS One. **7**(8): e43738.

Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient virus world and evolution of cells. Biol. Direct. **1**:29.

Kornberg A, Baker TA. 1992. DNA Replication (2d ed.). WH Freeman and Company.

Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, Baxter SM, Derbyshire V. 1999. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. Nucleic Acids Res. **27**(10): 2115-2125.

Kowalski JC, Derbyshire V. 2002. Characterization of homing endonucleases. Methods. **28**(3): 365-373.

Kück U, Godehardt I, Schmidt U. 1990. A self-splicing group II intron in the mitochondrial large subunit rRNA (LSUrRNA) gene of the eukaryotic alga *Scenedesmus obliquus*. Nucleic Acids Res. **18**: 2691-2697.

Kühlmann UC, Moore GR, James R, Kleanthous C, Hemmings AM. 1999. Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? FEBS Lett. **463**(1-2): 1-2.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. Annu. Rev. Biochem. **62**: 587-622.

Lambowitz AM, Belfort M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. Microbiol. Spectr. **3**(1): MDNA3-0050-2014.

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. **3**(8): a003616.

Lambowitz AM, Caprara MG, Zimmerly S and Perlman PS. 1999. Group I and group II ribozymes as RNPs: Clues to the past and guides to the future. In: Gesteland RF, Cech TR, Atkins JF (eds.). The RNA World, Cold Spring Harbour Laboratory Press, New York, pp. 451-485.

Lamech LT, Mallam AL, Lambowitz AM. 2014. Evolution of RNA-protein interactions: non-specific binding led to RNA splicing activity of fungal mitochondrial tyrosyl-tRNA synthetases. PLoS Biol. **12**: e1002028.

Landthaler M, Lau NC, Shub DA. 2004. Group I intron homing in *Bacillus* phages SPO1 and SP82: a gene conversion event initiated by a nicking homing endonuclease. J. Bacteriol. **186**(13): 4307-4314.

Landthaler M, Shub DA. 2003. The nicking homing endonuclease I-BasI is encoded by a group I intron in the DNA polymerase gene of the Bacillus thuringiensis phage Bastille. Nucleic Acids Res. **31**(12): 3071-3077.

Lang BF, Laforest MJ, Burger G. 2007. Mitochondrial introns: a critical view. Trends Genet. **23**: 119-125.

Li CF, Costa M, Bassi G, Lai YK, Michel F. 2011. Recurrent insertion of 5'-terminal nucleotides and loss of the branchpoint motif in lineages of group II introns inserted in mitochondrial preribosomal RNAs. RNA.**17**: 1321-1335.

Li MV, Shukla D, Rhodes BH, Lall A, Shu J, Moriarity BS, Largaespada DA. 2014. HomeRun Vector Assembly System: A Flexible and Standardized Cloning System for Assembly of Multi-Modular DNA Constructs. PLoS ONE. **9**(6): e100948.

Li Y, Lin Z, Huang C, Zhang Y, Wang Z, Tang YJ, Chen T, Zhao X. 2015. Metabolic engineering of *Escherichia coli* using CRISPR-Cas9 meditated genome editing. Metab. Eng. **31**: 13-21.

Li H, Pellenz S, Ulge U, Stoddard BL, Monnat RJ Jr. 2009. Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. Nucleic Acids Res. **37**(5): 1650-1662.

Li T, Huang S, Jiang WZ, Wright D, Spalding MH, Weeks DP, Yang B. 2011. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res. **39**: 359-372.

Lieber MR. 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. Annu Rev Biochem. **79**: 181-211.

Lippow SM, Aha PM, Parker MH, Blake WJ, Baynes BM, Lipovsek D. 2009. Creation of a type IIS restriction endonuclease with a long recognition sequence. Nucleic Acids Res. **37**: 3061-3073.

Liu JK, Chen WH, Ren SX, Zhao GP, Wang J. 2014. iBrick: a new standard for iterative assembly of biological parts with homing endonucleases. PLoS One. **9**: e110852.

Loizos N, Silva GH, Belfort M. 1996. Intron-encoded endonuclease I-TevII binds across the minor groove and induces two distinct conformational changes in its DNA substrate. J. Mol. Biol. **255**(3): 412-424.

Loizos N, Tillier ER, Belfort M. 1994. Evolution of mobile group I introns: recognition of intron sequences by an intron-encoded endonuclease. Proc. Natl. Acad. Sci. USA. **91**: 11983-11987.

Longo A, Leonard CW, Bassi GS, Berndt D, Krahn JM, Tanaka-Hall TM, Weeks KM. 2005. Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. Nat. Struct. Mol. Biol. **12**: 779-787.

Lucas P, Otis C, Mercier JP, Turmel M, Lemieux C. 2001. Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. Nucleic Acids Res. **29**: 960-969.

Lykke-Andersen J, Aagaard C, Semionenkov M and Garrett RA. 1997. Archaeal introns: splicing, intercellular mobility and evolution. Trends Biochem Sci. **9**: 326-331.

Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. **12**(6): 701-710.

Lyznik LA, Djukanovic V, Yang M, Jones S. 2012. Double strand break-induced targeted mutagenesis in plants. Methods Mol. Biol. **847**: 399-416.

Maeder ML, Linder SJ, Reyon D, Angstman JF, Fu Y, Sander JD, Joung JK. 2013. Robust, synergistic regulation of human gene expression using TALE activators. Nat. Methods.**10**: 243-245.

Maier UG, Rensing SA, Igloi GL, Maerz M. 1995. Twintrons are not unique to the *Euglena* chloroplast genome: structure and evolution of a plastome cpn60 gene from a cryptomonad. Mol. Gen. Genet. **246**(1): 128-131.

Makrides SC. 1996. Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol Rev. **60**(3): 512-538.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. Science. **339**: 823-826.

Malik HS, Henikoff S. 2000. Dual recognition-incision enzymes might be involved in mismatch repair and meiosis. Trends Biochem. Sci. **25**(9): 414-418.

Marcaida MJ, Muñoz IG, Blanco FJ, Prieto J, Montoya G. 2010. Homing endonucleases: from basics to therapeutic applications. Cell. Mol. Life Sci. **67**: 727-748.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. Nature. **440**(7080): 41-45.

Marton I, Honig A, Omid A, De Costa N, Marhevka E, Cohen B, Zuker A, Vainstein A. 2013. From Agrobacterium to viral vectors: genome modification of plant cells by rare cutting restriction enzymes. Int. J. Dev. Biol. **57**: 639-650.

Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM. 2008. Group II intron-based gene targeting reactions in eukaryotes. PLoS One. **3**: e3121.

Matsuura M, Noah JW, Lambowitz AM. 2001. Mechanism of maturase-promoted group II intron splicing. EMBO J. **20**: 7259-7270.

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes Dev. **11**(21): 2910-2924.

McConnell Smith A, Takeuchi R, Pellenz S, Davis L, Maizels N, Monnat RJ Jr, Stoddard BL. 2009. Generation of a nicking enzyme that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. Proc. Natl. Acad. Sci. USA. **106**(13): 5099-5104.

McNeil BA, Semper C, Zimmerly S. 2016. Group II introns: versatile ribozymes and retroelements. WIREs RNA. doi: 10.1002/wrna.1339.

McVey M, Lee SE. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. **24**(11): 529-538.

Mehta P, Katta K, Krishnaswamy S. 2004. HNH family subclassification leads to identification of commonality in the His-Me endonuclease superfamily. Protein Sci. **13**(1): 295-300.

Meng Q, Zhang Y, Liu XQ. 2007. Rare group I intron with insertion sequence element in a bacterial ribonucleotide reductase gene. J. Bacteriol. **189**: 2150-2154.

Merendino L, Perron K, Rahire M, Howald I, Rochaix JD, Goldschmidt-Clermont M. 2006. A novel multifunctional factor involved in trans-splicing of chloroplast introns in *Chlamydomonas*. Nucleic Acids Res. **34**: 262-274.

Michel F, Costa M, Westhof E. 2009. The ribozyme core of group II introns: a structure in want of partners. Trends Biochem. Sci. **34**: 189-199.

Michel F, Ferat JL.1995. Structure and activities of group II introns. Annu. Rev. Biochem. **64**: 435-461.

Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Mol. Biol. **216**(3): 585-610.

Mohr G, Smith D, Belfort M, Lambowitz AM. 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes Dev. **14**(5): 559-573.

Molina-Sánchez MD, Toro N. 2015. Inactivation of group II intron RmInt1 in the *Sinorhizobium meliloti* genome. Sci Rep. **5**: 12036.

Monteilhet C, Dziadkowiec D, Szczepanek T, Lazowska J. 2000. Purification and characterization of the DNA cleavage and recognition site of I-ScaI mitochondrial group I intron encoded endonuclease produced in *Escherichia coli*. Nucleic Acids Res. **28**(5):1245-1251.

Moran JV, Zimmerly S, Eskes R, Kennell, JC, Lambowitz AM, Butow RA, Perlman PS. 1995. Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. Mol. Cell. Biol. **15**: 2828-2838.

Moreira S, Breton S, Burger G. 2012. Unscrambling genetic information at the RNA level. Wiley Interdiscip. Rev. RNA **3**: 213-228.

Mota EM, Collins RA. 1988. Independent evolution of structural and coding regions in a *Neurospora* mitochondrial intron. Nature. **332**: 654-656.

Moure CM, Gimble FS, Quiocho FA. 2002. Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. Nat. Struct. Biol. **9**(10): 764-770.

Moure CM, Gimble FS, Quiocho FA. 2003. The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. J. Mol. Biol. **334**(4):685-695.

Mueller JE, Clyman J, Huang YJ, Parker MM, Belfort M. 1996. Intron mobility in phage T4 occurs in the context of recombination-dependent DNA replication by way of multiple pathways. Genes Dev. **10**: 351-364.

Mueller JE, Smith D, Bryk M, Belfort M. 1995. Intron-encoded endonuclease I-TevI binds as a monomer to effect sequential cleavage via conformational changes in the td homing site. EMBO J. **14**(22): 5724-5735.

Mueller JE, Smith D, Belfort M. 1996. Exon coconversion biases accompanying intron homing: battle of the nucleases. Genes Dev. **10**: 2158-2166.

Mueller MW, Allmaier M, Eskes R, Schweyen RJ. 1993. Transposition of group II intron aI1 in yeast and invasion of mitochondrial genes at new locations. Nature. **366**: 174-176.

Muir RS, Flores H, Zinder ND, Model P, Soberon X, Heitman J. 1997. Temperature-sensitive mutants of the EcoRI endonuclease. J. Mol. Biol. **274**: 722-737.

Mullineux ST, Costa M, Bassi GS, 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.

Muscarella DE, Ellison EL, Ruoff BM, Vogt VM. 1990. Characterization of I-PpoI, an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of *Physarum polycephalum*. Mol. Cell Biol. **10**(7): 3386-3396.

Nelson DL, Kennedy EP. 1971. Magnesium transport in *Escherichia coli*: Inhibition by cobaltous ion. J. Biol. Chem. **246**: 3042-3049.

Nelson DL, Kennedy EP. 1972. Transport of magnesium by a repressible and a nonrepressible system in *Escherichia coli*. Proc. Natl. Acad. Sci. USA. **69**: 1091-1093.

Nemčovičová I, Smatanová IK. 2012. Alternative Crystallization Technique: Cross Influence Procedure (CIP). In: Elena Borisenko (ed). Chapter11: In the Crystallization and Materials Science of Modern Artificial and Natural Crystals. 249-276. InTech.

Nicholas KB, Nicholas HB, Deerfield DW 2nd . 1997. GeneDoc: analysis and visualization of genetic variation. EMB News. **4**: 14.

Nihongaki Y, Kawano F, Nakajima T, Sato M. 2015. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. Nat. Biotechnol. **33**: 755-760.

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.

Noah JW, Lambowitz AM. 2003. Effects of maturase binding and Mg^{2+} concentration on group II intron RNA folding investigated by UV cross-linking. Biochemistry. **42**: 12466-12480.

Nord D, Sjöberg BM. 2008. Unconventional GIY-YIG homing endonuclease encoded in group I introns in closely related strains of the *Bacillus cereus* group. Nucleic Acids Res. **36**(1): 300-310.

O'Connell MR, Oakes BL, Sternberg SH, East-Seletsky A, Kaplan M, Doudna JA. 2014. Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature. **516**: 263-266.

Ochiai H, Sakamoto N, Suzuki K, Akasaka K, Yamamoto T. 2008. The Ars insulator facilitates I-SceI meganuclease-mediated transgenesis in the sea urchin embryo. Dev Dyn. **237**(9):2475-2482.

Ogino H, McConnell WB, Grainger RM. 2006. Highly efficient transgenesis in *Xenopus tropicalis* using I-SceI meganuclease. Mech. Dev. **123**(2): 103-113.

Olga F, Nora Z. 2007. Group II introns: structure, folding and splicing mechanism (Review). Biol. Chem. **388**: 665-678.

Öhman-Hedén M, Ahgren-Stålhandske A, Hahne S, Sjöberg BM. 1993. Translation across the 5′-splice site interferes with autocatalytic splicing. Mol Microbiol. **7**: 975-982.

Orgel LE, Crick FH. 1980. Selfish DNA: The ultimate parasite. Nature. **284**: 604-607.

Page RDM. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comp. Appl. Biosci. **12**: 357-358.

Palmer JD, Logsdon JM Jr. 1991. The recent origins of introns. Curr. Opin. Genet. Dev. **1**: 470-477.

Palmer I, Wingfield PT. 2004. Preparation and extraction of insoluble (inclusion-body) proteins from *Escherichia coli*. Curr. Protoc. Protein Sci. **6**: 6.3.

Paquin B, O'Kelly CJ, Lang BF. 1995. Intron-encoded open reading frame of the GIY-YIG subclass in a plastid gene. Curr. Genet. **28**(1): 97-99.

Paquin B, Laforest MJ, Forget L, Roewer I, Wang Z, Longcore J and Lang BF. 1997. The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. Curr. Genet. **31**: 380-395.

Pattanayak V, Guilinger JP, Liu DR. 2014. Determining the specificities of TALENs, Cas9, and other genome-editing enzymes. Methods Enzymol. **546**: 47-78.

Pattanayak V, Ramirez CL, Joung JK, Liu DR. 2011. Revealing off-target cleavage specificities of zinc-finger nucleases by *in vitro* selection. Nat Methods. **8**: 765-770.

Perler FB, Olsen GJ, Adam E. 1997. Compilation and analysis of intein sequences. Nucleic Acids Res. **25**: 1087-1094.

Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J, Belfort M. 1994. Protein splicing elements: inteins and exteins, a definition of terms and recommended nomenclature. Nucleic Acids Res. **22**(7): 1125-1127.

Pernstich C, Halford SE. 2012. Illuminating the reaction pathway of the FokI restriction endonuclease by fluorescence resonance energy transfer. Nucleic Acids Res. **40**:1203-1213.

Pfeifer A, Martin B, Kämper J, Basse CW. 2012. The mitochondrial LSU rRNA group II intron of *Ustilago maydis* encodes an active homing endonuclease likely involved in intron mobility. PLoS One. **7**: e49551.

Phylactou LA, Darrah C, Wood MJ. 1998. Ribozyme-mediated trans-splicing of a trinucleotide repeat. Nat. Genet. **18**: 378-381.

Pombert JF, James ER, Janouškovec J, Keeling PJ. 2012. Evidence for transitional stages in the evolution of euglenid group II introns and twintrons in the *Monomorphina aenigmatica* plastid genome. PLoS One. **7**: e53433.

Posey KL, Gimble FS. 2002. Insertion of a reversible redox switch into a rare-cutting DNA endonuclease. Biochemistry. **41**(7): 2184-2190.

Prieto J, Molina R, Montoya G. 2012. Molecular scissors for *in situ* cellular repair. Crit. Rev. Biochem. Mol. Biol. doi:10.3109/10409238.2011.652358.

Puigbò P, Guzmá E, Romeu A, Garcia-Vallvé S. 2007. OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. **35**: W126-W131.

Pul U, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R. 2010. Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. Mol. Microbiol. **75**: 1495–1512.

Pyle AM, Lambowitz AM. 2006. Group II introns: ribozymes that splice RNA and invade DNA. In: The RNA world (3rd ed) pp. 469-505.

Pyne ME, Moo-Young M, Chung DA, Chou CP. 2014. Expansion of the genetic toolkit for metabolic engineering of *Clostridium pasteurianum*: chromosomal gene disruption of the endogenous CpaAI restriction enzyme. Biotechnol. Biofuels. **7**: 163.

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell. **152**(5): 1173-1183.

Qin PZ, Pyle AM. 1998. The architectural organization and mechanistic function of group II intron structural elements. Curr Opin Struct Biol. **8**(3): 301-308.

Quiroga C, Kronstad L, Ritlop C, Filion A, Cousineau B. 2011. Contribution of base-pairing interactions between group II intron fragments during trans-splicing *in vivo*. RNA. **17**: 2212-2221.

Ralph D, McClelland M. 1993. Intervening sequence with conserved open reading frame in eubacterial 23S rRNA genes. Proc. Natl. Acad. Sci. USA. **90**: 6864-6868.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. **8**: 2281-2308.

Redondo P, Prieto J, Muñoz IG, Alibés A, Stricher F, Serrano L, Cabaniols JP, Daboussi F, Arnould S, Perez C, Duchateau P, Pâques F, Blanco FJ, Montoya G. 2008. Molecular basis of Xeroderma pigmentosum group C DNA recognition by engineered meganucleases. Nature. **456**: 107-111.

Rest JS, Mindell DP. 2003. Retroids in archaea: Phylogeny and lateral origins. Mol Biol. Evol. **20**: 1134-1142.

Reuter M, Bell G, Greig D. 2007. Increased outbreeding in yeast in response to dispersal by an insect vector. Curr Biol. 2007. **17**(3): 81-83.

Riggs P. 2000. Expression and purification of recombinant proteins by fusion to maltose-binding protein. Mol. Biotechnol. **15**(1): 51-63.

Rivière J, Hauer J, Poirot L, Brochet J, Souque P, Mollier K, Gouble A, Charneau P, Fischer A, Pâques F, de Villartay J-P, Cavazzana M. 2014. Variable correction of Artemis deficiency by I-Sce1-meganuclease-assisted homologous recombination in murine hematopoietic stem cells. Gene Therapy. **21**: 529-532.

Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. Cytogenet. Genome Res. **110**: 589-597.

Roberts RJ, Macelis D. 1997. REBASE - restriction enzymes and methylases. Nucleic Acids Res. **25**(1): 248-262.

Roberts RJ, Vincze T, Posfai J, Macelis D. 2003. REBASE: restriction enzymes and methyltransferases. Nucleic Acids Res. **31**(1): 418-420.

Roberts RJ, Vincze T, Posfai J, Macelis D. 2010. REBASE - a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. **38**(Database issue): D234-D236.

Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE - a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. **43**(Database issue): D298-D299.

Roman J, Woodson SA. 1995. Reverse splicing of the Tetrahymena IVS: Evidence for multiple reaction sites in the 23S rRNA. RNA. **1**: 478-490.

Ronquist F, Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinfo. **19**: 1572-1574.

Rosano GL, Ceccarelli EA. 2014. Recombinant protein expression in microbial systems. Front Microbiol. **5**: 341.

Rudski SM, Hausner G. 2012. The mtDNA rps3 locus has been invaded by a group I intron in some species of *Grosmannia.* Mycoscience. **53**: 471-475.

Rudi K, Fossheim T, Jakobsen KS. 2002. Nested evolution of a tRNA (Leu) (UAA) group I intron by both horizontal intron transfer and recombination of the entire tRNA locus. J. Bacteriol. **184**: 666-671.

Saguez C, Lecellier G, Koll F. 2000. Intronic GIY-YIG endonuclease gene in the mitochondrial genome of Podospora curvicolla: evidence for mobility. Nucleic Acids Res. **28**(6): 1299-1306.

Saldanha R, Mohr G, Belfort M, Lambowitz AM. 1993. Group I and group II introns. FASEB J. **7**: 15-24.

Salman V, Amann R, Shub DA, Schulz-Vogt HN. 2012. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. Proc. Natl. Acad. Sci. USA. **109**: 4203-4208.

Salvo JL, Rodeghier B, Rubin A, Troischt T. 1998. Optional introns in mitochondrial DNA of *Podospora anserina* are the primary source of observed size polymorphisms. Fungal Genet Biol. **23**: 162-168.

Samuelson JC. 2011. Recent developments in difficult protein expression: a guide to *E. coli* strains, promoters and relevant host mutations. Methods Mol. Biol. **705**: 195-209.

Sander JD, Joung JK. 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. **32**(4): 347-535.

Saunders S, Cooke B, McColl K, Shine R, Peacock T. 2010. Modern approaches for the biological control of vertebrate pests: an Australian perspective. Biol. Control. **52**(3): 288-295.

Scalley-Kim M, McConnell-Smith A, Stoddard BL. 2007. Coevolution of homing endonuclease specificity and its host target sequence. J. Mol. Biol. **372**: 1305-1319.

Scamborova P, Wong A, Steitz JA. 2004. An intronic enhancer regulates splicing of the twintron of *Drosophila melanogaster* prospero pre-mRNA by two different spliceosomes. Mol. Cell. Biol. **24**: 1855-1869.

Schleifman EB, Chin JY, Glazer PM. 2008. Triplex-mediated gene modification. Methods Mol. Biol. **435**: 175-190.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. Nucleic Acids Res. **33**(Web Server issue): W382-388.

Selleck W, Tan S. 2008. Recombinant protein complex expression in *E. coli*. Curr. Protoc. Protein Sci. **5**: 5.21.

Sellem CH, Belcour L. 1997. Intron open reading frames as mobile elements and evolution of a group I intron. Mol. Biol. Evol. **14**: 518-526.

Seligman LM, Chisholm KM, Chevalier BS, Chadsey MS, Edwards ST, Savage JH, Veillet AL. 2002. Mutations altering the cleavage specificity of a homing endonuclease. Nucleic Acids Res. **30**(17): 3870-3879.

Seligman LM, Stephens KM, Savage JH, Monnat RJ Jr. 1997. Genetic analysis of the *Chlamydomonas reinhardtii* I-CreI mobile intron homing system in *Escherichia coli*. Genetics. **147**: 1653-1664.

Sethuraman J, Majer A, Friedrich NC, Edgell DR, Hausner G. 2009. Genes-within-genes: Multiple LAGLIDADG homing endonucleases target the ribosomal protein S3 gene encoded within a rnl group I intron of *Ophiostoma* and related taxa. Mol. Biol. Evol. **26**: 2299-2315.

Sethuraman J, Rudski SM, Wosnitza KM, Hafez M, Guppy B, Hausner G, 2013. Evolutionary dynamics of introns and their open reading frames in the U7 region of the mitochondrial rnl gene in species of Ceratocystis. Fungal Biol. **117**: 791-806.

Sharma M, Ellis RL, Hinton DM. Identification of a family of bacteriophage T4 genes encoding proteins similar to those present in group I introns of fungi and phage. Proc. Natl. Acad. Sci. USA. **89**(14): 6658-6662.

Sharp PA. 1994. Split genes and RNA splicing. Cell. **77**: 805-815.

Shearman C, Godon JJ, Gasson M. 1996. Splicing of a group II intron in a functional transfer gene of *Lactococcus lactis*. Mol. Microbiol. **21**(1): 45-53.

Shen BW, Landthaler M, Shub DA, Stoddard BL. 2004. DNA binding and cleavage by the HNH homing endonuclease I-HmuI. J. Mol. Biol. **342**(1): 43-56.

Shen BW, Lambert A, Walker BC, Stoddard BL, Kaiser BK. 2016. The Structural Basis of Asymmetry in DNA Binding and Cleavage as Exhibited by the I-SmaMI LAGLIDADG Meganuclease. J. Mol. Biol. **428**(1): 206-220.

Shrivastav M, De Haro LP, Nickoloff JA. 2008. Regulation of DNA double-strand break repair pathway choice. Cell Res. **18**(1): 134-147.

Shub DA, Goodrich-Blair H, Eddy SR**.** 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. Trends Biochem. Sci. **19**(10): 402-404.

Sigel RKO. 2005. Group II intron ribozymes and metal ions - a delicate relationship. Eur. J. Inorg. Chem. 2281-2292.

Siegl T, Petzke L, Welle E, Luzhetskyy A. 2010. I-SceI endonuclease: a new tool for DNA repair studies and genetic manipulations in streptomyces. Appl. Microbiol. Biotechnol. **87**: 1525-1532.

Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, Lambowitz AM, Fire AZ. 2016. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science. **351**(6276): 4234.

Silva G, Poirot L, Galetto R, Smith J, Montoya G, Duchateau P, Pâques F. 2011. Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. Curr Gene Ther. **11**(1): 11-27.

Silva GH, Dalgaard JZ, Belfort M, Van Roey P. 1999. Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. J. Mol. Biol. **286**(4): 1123-1136.

Simossis VA, Heringa J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res. **33**: W289-294.

Singh P, Tripathi P, Silva GH, Pingoud A, Muniyappa K. 2009. Characterization of *Mycobacterium leprae* RecA intein, a LAGLIDADG homing endonuclease, reveals a unique

mode of DNA binding, helical distortion, and cleavage compared with a canonical LAGLIDADG homing endonuclease. J. Biol. Chem. **284**(38): 25912-25928.

Singh P, Tripathi P, Muniyappa K. 2010. Mutational analysis of active-site residues in the *Mycobacterium leprae* RecA intein, a LAGLIDADG homing endonuclease: Asp(122) and Asp(193) are crucial to the double-stranded DNA cleavage activity whereas Asp(218) is not. Protein Sci. **19**(1): 111-123.

Smith HO, Nathans D. 1973. Letter: A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. J. Mol. Biol. **81**(3): 419-423.

Soroldoni D, Hogan BM, Oates AC. 2009. Simple and efficient transgenesis with meganuclease constructs in zebrafish. Methods Mol. Biol. **546**: 117-130.

Southworth MW, Perler FB. 2002. Protein splicing of the *Deinococcus radiodurans* strain R1 snf2 intein. J Bacteriol. **184**(22): 6387-6388.

Sternberg SH, Doudna JA. 2015. Expanding the biologist's toolkit with CRISPR-Cas9. Mol Cell. **58**: 568-574.

Steuer S, Pingoud V, Pingoud A, Wende W. 2006. Chimeras of the homing endonuclease PI-SceI and the homologous *Candida tropicalis* intein: a study to explore the possibility of exchanging DNA-binding modules to obtain highly specific endonucleases with altered specificity. Chembiochem. **5**(2): 206-213.

Stoddard BL. 2006. Homing endonuclease structure and function. Q. Rev. Biophys. **38**: 49-95.

Stoddard BL. 2011. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. Structure. **19**: 7-15.

Stoddard BL. 2014. Homing endonucleases from mobile group I introns: discovery to genome engineering. Mob DNA. **5**: 7.

Stoddard B, Scharenberg AM, Monnat RJ Jr. 2008. Advances in engineering homing endonucleases for gene targeting: Ten years after structures. In: Bertolotti R, Ozawa K (eds.). Progress in Gene Therapy 3: Autologous and Cancer Stem Cell Gene Therapy, World Scientific Press. Hackensack. NJ. 135-167.

Storici F, Durham CL, Gordenin DA, Resnick MA. 2003a. Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. Proc. Natl. Acad. Sci. U.S.A. **100**: 14994-14999.

Storici F, Resnick MA. 2003b. *Delitto perfetto* targeted mutagenesis in yeast with oligonucleotides. Genet. Eng. (NY). **25**: 189-207.

Suzuki H, Kameyama T, Ohe K, Tsukahara T, Mayeda A. 2013. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. FEBS Lett. **587**: 555-561.

Swamy KH, Goldberg AL. 1982. Subcellular distribution of various proteases in *Escherichia coli*. J. Bacteriol. **149**(3): 1027-1033.

Takeuchi R, Lambert AR, Mak AN, Jacoby K, Dickson RJ, Gloor GB, Scharenberg AM, Edgell DR, Stoddard BL. 2011. Tapping natural reservoirs of homing endonucleases for targeted gene modification. Proc. Natl. Acad. Sci. USA. **108**(32): 13077-13082.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. **28**: 2731-2739.

Tao H, Liu W, Simmons BN, Harris HK, Cox TC, Massiah MA. 2010. Purifying natively folded proteins from inclusion bodies using sarkosyl, Triton X-100, and CHAPS. Biotechniques. **48**: 61-64.

Taylor GK, Petrucci LH, Lambert AR, Baxter SK, Jarjour J, Stoddard BL. 2012. LAHEDES: the LAGLIDADG homing endonuclease database and engineering server. Nucleic Acids Res. **40**(W1): W110-W116.

Taylor GK, Heiter DF, Pietrokovski S, Stoddard BL. 2011. Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family. Nucleic Acids Res. **39**(22): 9705-9719.

Tebas P, Sension M, Arribas J, Duiculescu D, Florence E, Hung CC, Wilkin T, Vanveggel S, Stevens M, Deckx H; ECHO, THRIVE Study Groups. 2014. Lipid levels and changes in body fat distribution in treatment-naive, HIV-1-Infected adults treated with rilpivirine or Efavirenz for 96 weeks in the ECHO and THRIVE trials. Clin. Infect. Dis. **59**(3): 425-434.

Thierry A, Dujon B. 1992. Nested chromosomal fragmentation in yeast using the meganuclease I-SceI: a new method for physical mapping of eukaryotic genomes. Nucleic Acids Res. **20**(21): 5625-5631.

Thakker C, Lin K, Martini-Stoica H, Bennett GN. 2015. Use of transposase and ends of IS608 enables precise and scarless genome modification for modulating gene expression and metabolic engineering applications in Escherichia coli. Biotechnol J. 10.1002/biot.201500205.

Thomas KR, Folger KR, Capecchi MR. 1986. High frequency targeting of genes to specific sites in the mammalian genome. Cell. **44**(3):419-428.

Thomas KR, Capecchi MR. 1987. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell. **51**(3): 503-512.

Thompson MD, Copertino DW, Thompson E, Favreau MR, Hallick RB. 1995. Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus Euglena. Nucleic Acids Res. **23**(23): 4745-4752.

Thompson AJ, Yuan X, Kudlicki W, Herrin DL. 1992. Cleavage and recognition pattern of a double-strand-specific endonuclease (I-CreI) encoded by the chloroplast 23S rRNA intron of Chlamydomonas reinhardtii. Gene. **119**(2): 247-251.

Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. 2011. Evolution of introns in the archaeal world. Proc. Natl. Acad. Sci. USA. **108**: 4782-4787.

Toor N, Keating KS, Taylor SD, Pyle AM. 2008. Crystal structure of a self-spliced group II intron. Science. **320**: 77-82.

Toor N, Robart AR, Christianson J, Zimmerly S. 2006. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem-loop motif of a transcriptional terminator. Nucleic Acids Res. **34**: 6461-6471.

Toor N, Zimmerly S. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. RNA. **8**(11):1373-1377.

Tourasse NJ, Helgason E, Økstad OA, Hegna IK, Kolstø AB. 2006. The Bacillus cereus group: novel aspects of population structure and genome dynamics. J. Appl. Microbiol. **101**: 579–593.

Truong DM, Hewitt FC, Hanson JH, Cui X, Lambowitz AM. 2015. Retrohoming of a mobile group II intron in human cells suggests how eukaryotes limit group II intron proliferation. PLoS Genet. **11**: e1005422.

Truong DM, Sidote DJ, Russell R, Lambowitz AM. 2013. Enhanced group II intron retrohoming in magnesium-deficient *E.coli* via selection of mutations in the ribozyme core. Proc. Natl. Acad. Sci. USA. **110**: 3800-3809.

Turmel M, Otis C, Côté V, Lemieux C. 1997. Evolutionarily conserved and functionally important residues in the I-CeuI homing endonuclease. Nucleic Acids Res. **25**(13): 2610-2619.

Ulge UY, Baker DA, Monnat RJ Jr. 2011. Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. Nucleic Acids Res. **39**: 4330-4339.

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature. **435**(7042): 646-651.

Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. 2010. Genome editing with engineered zinc finger nucleases. Nat. Rev. Genet. **11**(9): 636-646.

Vainstein A, Marton I, Zuker A, Danziger M, Tzfira T. 2011. Permanent genome modifications in plant cells by transient viral vectors. Trends Biotechnol. **29**(8): 363-369.

Valton J, Daboussi F, Leduc S, Molina R, Redondo P, Macmaster R, Montoya G, Duchateau P. 2012. 5'-Cytosine-phosphoguanine (CpG) methylation impacts the activity of natural and engineered meganucleases. J. Biol. Chem. **287**(36): 30139-30150.

Villate M, Merinoa N, Blanco FJ. 2012. Production of meganucleases by cell-free protein synthesis for functional and structural studies. Protein Expression and Purification. **85**(2): 246-249.

Walter TS, Meier C, Assenberg R, Au KF, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JM. 2006. Lysine methylation as a routine rescue strategy for protein crystallization. Structure. **14**(11): 1617-1622.

Wang XJ, Peng YJ, Zhang LQ, Li AN, Li DC. 2012. Directed evolution and structural prediction of cellobiohydrolase II from the thermophilic fungus Chaetomium thermophilum. Appl. Microbiol. Biotechnol. **95**: 1469-1478.

Wang Y, Zhou XY, Xiang PY, Wang LL, Tang H, Xie F, Li L, Wei H. 2014. The meganuclease I-SceI containing nuclear localization signal (NLS-I-SceI) efficiently mediated mammalian germline transgenesis via embryo cytoplasmic microinjection. PLoS One. **9**(9): e108347.

Waring RB, Davies RW, Scazzocchio C, Brown TA. 1982. Internal structure of a mitochondrial intron of Aspergillus nidulans. Proc. Natl. Acad. Sci. USA. **79**(20): 6332-6336.

Watanabe K, Breier U, Hensel G, Kumlehn J, Schubert I, Reiss B. 2016. Stable gene replacement in barley by targeted double-strand break induction. J. Exp. Bot. **67**(5):1433-1445.

Weiner AM. 1993. mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism? Cell. **72**: 161-164.

Westra ER, Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, Wurm R, Raine A, Mescher M, Van Heereveld L, Mastop M, Wagner EG, Schnetz K, Van Der Oost J, Wagner R, Brouns SJ. 2010. H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. Mol. Microbiol. **77**: 1380-1393.

Williams DA, Baum C. 2003. Medicine. Gene therapy-new challenges ahead. Science. **302**(5644): 400-401.

Wilkinson DL, Harrison RG. 1991. Predicting the solubility of recombinant proteins in Escherichia coli. Bio/Technology. **9**: 443-448.

Wilson GG. 1988. Cloned restriction-modification systems-a review. Gene. **74**(1): 281-289.

Wilson TE, Lieber MR. 1999. Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway. J. Biol. Chem. **274**(33): 23599-23609.

Windbichler DA, Papathanos PA, Catteruccia F, Ranson H, Burt A, Crisanti A. Homing endonuclease mediated gene targeting in *Anopheles gambiae* cells and embros. 2007. Nucleic Acids Res. **35**: 5922-5933.

Windbichler N, Papathanos PA, Crisanti A. 2008. Targeting the X chromosome during Spermatogenesis Induces Y Chromosome Transmission Ratio Distortion and Early Dominant Embryo Lethality in Anopheles gambiae. PLoS Genet. **4**(12): e1000291.

Wolfs JM, DaSilva M, Meister SE, Wang X, Schild-Poulter C, Edgell DR. 2014. MegaTevs: single-chain dual nucleases for efficient gene disruption. Nucleic Acids Res. **42**(13): 8816-8829.

Wong SM. 2004. SCE jumping: genetic tool for allelic exchange in bacteria. Crit. Rev. Eukaryot. Gene Expr. **14**(1-2): 53-64.

Wu B, Buljic A, Hao W. 2015. Extensive Horizontal Transfer and Homologous Recombination Generate Highly Chimeric Mitochondrial Genomes in Yeast. Mol Biol Evol. doi: 10.1093/ molbev/msv12.

Wu B, Hao W. 2014. Horizontal Transfer and Gene Conversion as an Important Driving Force in Shaping the Landscape of Mitochondrial Introns. G3 (Bethesda). **4**(4): 605-612.

Xiong Y, Eickbush T. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**: 3353-3362.

Xu SY, Kuzin AP, Seetharaman J, Gutjahr A, Chan SH, Chen Y, Xiao R, Acton TB, Montelione GT, Tong L. 2013. Structure determination and biochemical characterization of a putative HNH endonuclease from *Geobacter metallireducens* GS-15. PLoS One. **8**(9): e72114.

Yang J, Zimmerly S, Perlman PS, Lambowitz AM. 1996. Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. Nature. **381**: 332-335.

Yang J, Mohr G, Perlman PS, Lambowitz AM. 1998. Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites *in vitro*. J. Mol. Biol. **282**: 505-523.

Yang M, Djukanovic V, Stagg J, Lenderts B, Bidney D, Falco SC, Lyznik LA. 2009. Targeted mutagenesis in the progeny of maize transgenic plants. Plant Mol. Biol. **70**(6): 669-679.

Yang J, Mohr G, Perlman PS, Lambowitz AM. 1998. Group II intron mobility in yeast mitochondria: target DNA-primed reverse transcription activity of aI1 and reverse splicing into DNA transposition sites in vitro. J. Mol. Biol. **282**(3): 505-523.

Yao J, Lambowitz AM. 2007. Gene targeting in gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. Appl. Environ. Microbiol. **73**: 2735-2743.

Yao J, Truong DM, Lambowitz AM. 2013. Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. PLoS Genet. **9**: e1003469.

Yin LF, Hu MJ, Wang F, Kuang H, Zhang Y, Schnabel G, Li GQ, Luo CX. 2012. Frequent gain and loss of introns in fungal cytochrome b genes. PLoS One. **7**: e49096.

Yusufzai T, Kadonaga JT. 2010. Annealing helicase 2 (AH2), a DNA-rewinding motor with an HNH motif. Proc. Natl. Acad. Sci. USA. **107**(49): 20970-20973.

Zeng Q, Bonocora RP, Shub DA. 2009. A novel free-standing homing endonuclease that targets an intron insertion site in the psbA gene of cyanophages. Curr. Biol. **19**: 218-222.

Zetsche B, Volz SE, Zhang F. 2015. A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat. Biotechnol. **33**: 139-142.

Zhang L, Doudna JA. 2002. Structural insights into group II intron catalysis and branch-site selection. Science. **295**: 2084-2088.

Zhao L, Bonocora RP, Shub DA, Stoddard BL. 2007. The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. EMBO J. **26**(9): 2432-2442.

Zhong J, Karberg M, Lambowitz AM. 2003. Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. Nucleic Acids Res. **31**: 1656-1664.

Zhou H, Li P, Wu D, Ran T, Wang W, Xu D. 2015. EheA from Exiguobacterium sp. yc3 is a novel thermostable DNase belonging to HNH endonuclease superfamily. FEMS Microbiol. Lett. **362**(24): 204.

Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995a. Group II intron mobility occurs by target DNA-primed reverse transcription. Cell. **82**: 545-554.

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995b. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell. **83**: 529-538.

Zimmerly S, Semper C. 2015. Evolution of group II introns. Mobile DNA. **6**:7.

Zimmerly S, Wu L. 2015. An unexplored diversity of Reverse Transcriptases in bacteria. Microbiol. Spectrum. **3**(2): MDNA3-0058-2014.

Zinn AR, Butow RA. 1985. Nonreciprocal exchange between alleles of the yeast mitochondrial 21S rRNA gene: kinetics and the involvement of a double-strand break. Cell. **40**(4): 887-895.