Evolutionary dynamics of intein-encoded homing endonucleases: Characterizing the related inteins in r-Gyr and TopA in hyperthermophilic archaea

By

Miriam Shiffman

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts
in Molecular Biology

Advised by Professor Lenny M. Seligman

POMONA COLLEGE
Claremont, CA
May 3, 2013

TABLE OF CONTENTS

ABSTRACT

Homing endonuclease genes represent selfish genetic elements capable of catalyzing their own spread based on the activity of the enzymes they encode. As a result, homing endonucleases (HEs) face conflicting selective pressures for promiscuity and specificity. While cyclical evolution of intein-encoded HEs is hypothesized, much about the distribution and dispersal of these molecular parasites remains unknown. In this investigation, characterization of a previously unstudied group of related archaeal inteins in functionally distinct host genes is used to study factors affecting intein evolution. Here, I show that the inteins in *M. jannaschii* reverse gyrase, *T. kodakaraensis* reverse gyrase, and *T. kodakaraensis* topoisomerase I encode active HEs. Based on cleavage preferences for native and nonnative sites *in vitro*, these experiments reveal insight into different strategies for homing and enable me to propose a model for the uneven distribution of the related r-Gyr/TopA inteins *in vivo* among five species of archaea.

INTRODUCTION

Homing endonucleases

Homing endonucleases are named for their exceptional ability to "home" to a particular DNA sequence that they can recognize and cleave. Paradoxically, these enzymes are noted for both high specificity and promiscuity; they recognize exceptionally lengthy target sites (14-40 bp), enabling single cleavage within a sufficiently complex genome, yet are tolerant of many sequence variants (Burt & Koufopanou, 2004; Li *et al.*, 2012).

These characteristics are associated with the evolution of homing endonucleases (HEs) as streamlined mobile genetic elements, encoded by short yet efficient selfish DNA sequences capable of catalyzing their own spread. Homing endonuclease genes (HEGs) are generally contained within open reading frames in intervening sequences (introns or inteins) within host genes. Because the HE recognizes a target site composed of the DNA sequence flanking the intervening sequence in the host gene, only host alleles lacking the intervening sequence (termed HEG-) are susceptible to cleavage. The cleaved HEG- allele may then be repaired by the cell's endogenous machinery using the intact HEG+ allele as a template, resulting in gene conversion. Through this mechanism, homing can result in both super-Mendelian inheritance and horizontal transfer of the HEG, while self-splicing at the DNA or protein level avoids disruption to host protein function (Chevalier & Stoddard, 2001). While the simple and elegant homing process has

been elucidated, much about the evolution and dispersal of these molecular parasites remains mysterious.

Applications

The exceptional specificity of HEs, unmatched by restriction enzymes or other nucleases, has many useful applications. For example, HEs can be engineered to cleave a specific mutant human allele while leaving the rest of the genome intact, making them valuable for gene therapy (Takeuchi *et al.*, 2011). In addition, coupling HEs to regulatory genes in mosquitoes could be used to prevent insect populations from serving as vectors of disease (Windbichler *et al.*, 2011). Protein engineering of HE specificity is now viable, but a more thorough understanding of how HEs evolve and propagate in the real world is essential to moving them out of the laboratory.

The LAGLIDADG family of endonucleases

HEs are found in the genomes of all three domains of life and viruses, and are classified into five families according to their active site motifs (Li *et al.*, 2012). At least four of these protein families are thought to have originated independently, suggesting that these selfish entities are a fundamental property of biology. The HEs characterized in this study are related to the LAGLIDADG family of endonucleases. This family, also termed DOD or dodecapeptide, is the largest and best characterized, and includes HEs associated with group I introns, archaeal introns and inteins, and freestanding genes (Chevalier & Stoddard, 2001).

LAGLIDADG endonucleases are noted for little primary sequence homology but conserved folded topology, consisting of an αββαββα structure of the protein core. The first α-helix contains the titular LAGLIDADG motif, while the β-strands comprise an antiparallel sheet that forms the DNA-binding saddle. Each functional LAGLIDADG endonuclease contains two of these conserved motifs near the active sites, formed alternately by a homodimer (thus imposing restrictions on the symmetry of the target site) or by two LAGLIDADG motifs within a monomer. Tight packing of these motifs enables coordinated cleavage of both DNA strands, yielding 3’ terminal 4-bp overhangs (Chevalier & Stoddard, 2001). Multiple protein-DNA contacts, both direct and water-

mediated, between the binding saddle and target site facilitate the exquisite specificity for which HEs are known (Li *et al.*, 2012; Rosen *et al.*, 2006).

Despite similarities in form and function, LAGLIDADG endonucleases are overall highly divergent. Comparison of phylogenetic trees indicates frequent transposition events as well as shuffling of LAGLIDADG domains, which may explain the observed diversity of these HEs (Dalgaard *et al.*, 1997).

Intein propagation through homing

In the case of intein-encoded HEs, which are the subject of this investigation, intein propagation is dependent upon the presence of both functional protein splicing domains and a functional endonuclease domain. Following translation of the host protein, the intein catalyzes the splicing reaction that gives rise to itself and the ligated host extein. If the intein then comes into contact with a cognate HEG- allele, the intein will home to and cleave its target site. Repair by the cell's double-strand break machinery using the intact HEG+ allele as a template results in transfer of the intein to the empty site (Figure 1). Through this mechanism, the intein is spread among members of a species or even to other species (Chevalier & Stoddard, 2001). Although the mechanism by which HEG+ alleles and the inteins they encode come into contact with HEG- alleles is not fully understood (Koufopanou, Goddard, & Burt, 2002), horizontal transfer of an intein-encoded LAGLIDADG HE has been demonstrated *in vivo* in archaea (Naor *et al.*, 2011).

Notably, all inteins do not contain homing endonucleases and all homing endonucleases are not encoded within inteins or other intervening sequences (Pietrokovski, 2001). Phylogenetic analysis of LAGLIDADG HEs confirmed separate origins for the protein splicing and endonuclease domains of inteins (Dalgaard *et al.*, 1997). Thus, the residence of HEs within inteins represents a sort of symbiosis between selfish genetic elements: the former imparts mobility, while the latter minimizes the risk of loss from the host genome (Swithers *et al.*, 2009).

Cyclical evolution of inteins

The invasion of inteins by HEs has important implications for the evolution of these selfish entities. Within sequenced genomes, intein-encoded HEs are observed in

three conditions: absent, degenerate, and functional. These states are hypothesized to correspond to three stages within a dynamic cycle of transmission, fixation, and loss (Burt & Koufopanou, 2004).

In this model, self-propagation of an intein will eventually fill all of the target sites in a given population. However, successful propagation is ultimately detrimental to the maintenance of a functional HE; once no empty alleles are available as potential target sites, the intein no longer has selective pressure to maintain endonuclease activity. The HE domain may then accumulate deletions and other inactivating mutations (Burt & Koufopanou, 2004). However, intein splicing domains must remain functional to avoid negatively impacting host extein function (Pietrokovski, 2001).

Eventually, at a very low rate that may be relevant over evolutionary timescales, the intein may be lost through precise excision, as imprecise loss would be deleterious to the host. This event regenerates the empty site, which is then available for reinvasion by a homologous intein (Burt & Koufopanou, 2004). While a biological mechanism for precise intein loss is unknown, mathematical modeling of VDE intein evolution predicted a maximum waiting time of six million years between each gain and loss, implying that VDE has been dynamically lost and transferred over 100 times over the course of its evolutionary history (Koufopanou *et al.*, 2002).

However, the cyclical model of intein evolution is not universally accepted. Pietrokovski (2001) proposes the alternate explanation that inteins were originally common and evolved to serve a beneficial function in an ancient progenitor, but are gradually becoming extinct since this unknown benefit is no longer functional. As noted by the author, further study of intein dispersal to new integration sites is required.

Factors governing intein evolution and dispersal

While examples of HE domestication are known, most inteins serve no benefit to their hosts and may even pose a liability (Pietrokovski, 2001). No fitness difference, as measured by growth rate under lab conditions, was observed between HEG+ and HEG- strains of a halophilic species of archaea (Naor *et al.*, 2011). Thus, the presence of numerous inteins among a diverse variety of species must be explained by factors that favor the selfish persistence and propagation of HEGs.

Due to the threat of degeneration posed by successful homing and fixation, horizontal transmission of the intein is thought to be essential to the long-term persistence of HE functionality. The ubiquity of horizontal transfer is supported by different predicted phylogenies for related intein and extein sequences. Furthermore, although all LAGLIDADG HEs share a common evolutionary origin, phylogenetic trees with statistically significant branches could only be generated for inteins that share the same insertion site in homologous exteins from various species. Based on this finding, Perler created InBase, the NEB intein database, to catalogue these so-called allelic inteins, many of which contain conserved HE domains (Perler, 2002).

Thus, HE evolution is theorized to be governed by opposing forces: HEs must be promiscuous enough to cleave homologous sites in other species to ensure horizontal transmission, yet specific enough to avoid cleavage of ectopic sites within the host genome. HEs are therefore most likely to persist if their molecular characteristics enable homing strategies that strike a balance between maximizing the pool of potential target sites and avoiding host toxicity (Burt & Koufopanou, 2004).

This balance may explain why inteins tend to interrupt highly conserved regions of highly conserved genes (Swithers *et al.*, 2009). The likelihood of horizontal transfer is maximized by the presence of inteins in highly conserved sequences, which are more likely to be present in other species and similar enough to cleave. Additionally, excision of the intein may be minimized through insertion in highly conserved sequences, which would be more sensitive to imprecise loss (Koufopanou *et al.*, 2002; Swithers *et al.*, 2009). Also, many conserved genes are translated during DNA synthesis, which may provide the intein with better access to chromosomal DNA (Pietrokovski, 2001). While these other factors may also be operative, Koufopanou *et al*. (2002) provide experimental evidence that the VDE intein has evolved to recognize and cleave highly conserved amino acids in highly conserved regions of highly conserved genes, substantiating its adaptation for horizontal transfer.

Once transmitted to a new species, the HE may then evolve tighter specificity to its new insertion site if retaining broad specificity causes toxic ectopic cleavage. Such evolution is not unlikely, since single amino acid mutations can alter target specificity (Rosen *et al.*, 2006). However, the native site is not necessarily expected to be the

optimal substrate for an HE, both as a consequence of the dynamic evolutionary cycle and since selective pressure for mutation is only exerted to the extent that the HE avoids toxicity (Scalley-Kim *et al.*, 2007).

Over time, HEs may also co-evolve with their cognate sites. Multiple lines of experimental evidence suggest that the importance of each DNA base for HE recognition and cleavage is correlated with its degree of conservation. In other words, HEs demonstrate reduced specificity for bases encoding nonconserved amino acids or bases at wobble positions that are more likely to accumulate mutations due to the degeneracy of the genetic code and the location of inteins in sequences encoding highly conserved proteins essential to host fitness. Thus, to ensure vertical transmission, HE specificity may be attuned to accommodate genetic drift within the evolutionary constraints of the target site (Koufopanou *et al.*, 2002; Scalley-Kim *et al.*, 2007).

Despite possessing similar native target sites, allelic inteins often possess low sequence identity (Chute *et al.*, 1998; Dalgaard *et al.*, 1997) and may make disparate DNA contacts, resulting in altered specificity profiles (Rosen *et al.*, 2006). Characterization of two intron-encoded LAGLIDADG HEs with similar target sequences revealed that the two HEs exhibited very different binding and cleavage specificities. The authors concluded that these properties may be the result of different adaptive strategies: multiple DNA-protein contacts enable the evolution of higher cleavage activity without deleterious ectopic cleavage, while fewer DNA-protein contacts enable a wider range of host sites, mitigated by lower cleavage activity to reduce host toxicity (Li *et al.*, 2012).

While trends in the evolutionary dynamics of inteins are beginning to be elucidated – including the importance of horizontal transfer, their location in highly conserved sequences, co-evolution with their target sites, and the use of diverse strategies for homing – it remains poorly understood how the combination of these factors plays into the evolution and observed dispersal of inteins.

The r-Gyr/TopA group of inteins

Among the inteins listed in InBase, those in the reverse gyrase and topoisomerase I genes in some archaea represent the only apparent case of allelic inteins present in functionally distinct genes. Chute *et al.* (1998) reported the discovery of the first topoisomerase I (TopA) intein, and suggested that this intein in the TopA gene in

Pyrococcus furiosus is homologous to the intein in the reverse gyrase (r-Gyr) gene in *Methanococcus jannaschii*. These inteins are inserted in the same site in their respective exteins, interrupting a highly conserved region of the topoisomerase sequence two residues short of the catalytic residue. However, the *P. furiosus* TopA and *M. jannaschii* r-Gyr inteins share only 31% sequence homology and have the same codon usage as their respective host genomes. The authors postulate that these inteins are related by ancient horizontal transfer via homing, and thus represent the only known example of homologous inteins present in functionally distinct genes (Chute *et al.*, 1998).

The r-Gyr and TopA genes share a common evolutionary origin but are distributed differently among organisms and serve discrete functions. Early in evolutionary history, the ancestral TopA gene underwent duplication and subsequent fusion with a helicase module, giving rise to r-Gyr (Figure 2). Because r-Gyr is the only known gene limited to hyperthermophilic archaea and bacteria, this event is thought to be key to the existence of life at temperatures above 50°C. Reverse gyrase introduces positive supercoils in DNA to increase the melting temperature of double strands and also serves other functions to maintain DNA at high temperatures (Valenti *et al.*, 2008). On the other hand, the more ancient topoisomerase I functions to maintain a global balance of DNA supercoiling by relaxing negative supercoils, and is found among diverse species from all three domains of life (Viard *et al.*, 2001).

The inteins in r-Gyr and TopA represent a unique case of allelic inteins horizontally transferred between functionally distinct gene paralogs, rather than a case of vertical inheritance from the ancestral TopA gene, based on the known mobility of inteins and the number of precise deletions that would be required to explain the absence of the intein in many sequenced r-Gyr and TopA genes. Notably, *P. furiosus* and *M. jannaschii* possess copies of both genes, with an HEG+ allele at one insertion site and an HEG- allele at the other. Thus, the authors raise the question of why these inteins, if capable of homing, have not spread to the other site in these species (Chute *et al.*, 1998).

Since the discovery of the first TopA intein, many more genomes have been sequenced and annotated. Today, InBase lists five allelic inteins within the r-Gyr or TopA genes of four archaeal species: *Methanococcus jannaschii*, *Pyrococcus furiosus*, *Pyrococcus horikoshii*, and *Thermococcus kodakaraensis*. Notably, the intein is entirely

absent in some hyperthermophiles that share other allelic inteins with these species, such as *Pyroccus abyssi* (Perler, 2002). This record, though incomplete, emphasizes the uneven distribution of the r-Gyr/TopA intein among archaea (Figure 3).

In general, exposure of an HEG- allele to the HEG+ allele and corresponding intein is thought to be sufficient for transmission of the HEG. The existence of archaea with only one of these allelic inteins represents the atypical stable co-existence of HEG+ and HEG- alleles. This phenomenon has previously been described in a tetraploid yeast strain harboring an intein+ allele and a mutated intein- allele, which was only susceptible to HE cleavage under relaxed conditions (Gimble, 2001). Whether the r-Gyr/TopA group of inteins represents an evolutionary anomaly or a common step in the dispersal of HEGs to new integration sites, investigation of these inteins may provide insight into factors governing HE evolution and distribution.

Specific aims

In this study, I aim to characterize the related r-Gyr and TopA inteins through *in vitro* and *in silico* analysis. Notably, a high degree of genomic plasticity has been reported among several of the archaeal species in this study, potentially due to their hyperthermophilic lifestyle, complicating sequence analysis (Lecompte *et al.*, 2001). Thus, *in vitro* characterization of these HEs – which is particularly lacking in the study of archaeal inteins – will be an important component of this investigation.

First, among the members of this group in InBase, I hypothesize that the inteins in *M. jannaschii* r-Gyr and *T. kodakaraensis* r-Gyr and TopA have HE activity based on their size and the presence of conserved HE domains. Thus, I will assess their cleavage activity through *in vitro* digests using cloned inteins and target sites based on the 40 bp sequences that flank the intein *in vivo*.

In order to determine whether intein distribution corresponds to HE specificity, I will compare the cleavage specificities of the *Mja* r-Gyr, *Tko* r-Gyr, and *Tko* TopA inteins. Competition assays will be used to quantify the relative affinity of each HE for various nonnative sites in both r-Gyr and TopA, some of which contain inteins *in vivo*. HE affinity *in vitro* will then be compared to intein distribution *in vivo* in order to determine whether cleavage specificity explains the uneven distribution of the r-Gyr/TopA inteins among archaea. Characterization of the relative promiscuity of the *Tko*

r-Gyr and *Tko* TopA HEs may also indicate whether one of these HEs gave rise to the other through ectopic cleavage in this species or whether they were acquired by *T. kodakarensis* independently.

Furthermore, I aim to explore the basis for HE specificity among these inteins by determining the minimal target sequence required for HE cleavage. Comparison of narrowed site sequences with the ability of each HE to cleave each site may then elucidate base pairs that are important for homing. Furthermore, comparison of specificity profiles among the *Mja* r-Gyr, *Tko* r-Gyr, and *Tko* TopA HEs will be used to assess whether these related proteins have evolved different strategies for site recognition and cleavage. Thus, through investigation of the related inteins in r-Gyr and TopA, I aim to further clarify general principles regarding the evolution and dispersal of inteins, as well as the co-evolution of homing endonucleases and their target sites.

MATERIALS & METHODS

Bacterial strains and media

Escherichia coli strains DH5α and CC117 were used for molecular cloning, and TOP10 was used for heterologous protein expression. Standard LB growth media (10 g tryptone, 5 g yeast extract, 8 g NaCl per 1 L media) were used, with 15 g agar added for solid media. Media were supplemented with ampicillin (200 mg/L), tetracycline (100 mg/L), and/or X-gal (40 mg/L) as appropriate.

Amplification and cloning of inteins

In order to clone homing endonuclease sequences, inteins were amplified by polymerase chain reaction from archaeal genomic DNA. *Methanococcus jannaschii* genomic DNA was supplied by the American Type Culture Collection (product ATCC 43067D; atcc.org) and *Thermococcous kodakaraensis* genomic DNA was supplied by Professor EJ Crane (Pomona College). PCR primers were designed based on the first and last 22 bases of the intein sequence using the NEB Intein Database (Perler, 2002). Primers were designed to introduce a 5' in-frame NcoI site and a 3' in-frame NotI site. PCR reactions were set up using standard protocol with Taq polymerase (Promega; promega.com), with 2 uL each 5 uM primer and 10 ng DNA per 20 uL reaction. The reaction was carried out as follows: initial denaturation for 3 minutes at 94°C, followed by 34 cycles of denaturation, annealing, and elongation (30 seconds at 94°C, 30 seconds at 50°C, 2 minutes at 72°C), followed by final elongation for 5 minutes at 72°C.

Amplified inteins were cloned into pGEM-T Easy (Promega) following standard protocol (promega.com). This reaction was then transformed into competent cells and plated on selective media (Amp/X-gal). Colonies were screened for those containing plasmids with inserts (white in a blue/white screen), and plasmids were isolated using standard miniprep

protocol (Qiagen; qiagen.com). To confirm insertion of the intein, plasmids were diagnostically digested with EcoRI and then sequenced.

In order to construct plasmids for heterologous intein expression, inteins were then cloned into the ampicillin-resistant plasmid pBE, which contains an arabinose-inducible promoter for expression and a 6X-His tag for protein purification (Seligman *et al.*, 2002). Both plasmids (pGEM containing the intein insert and pBE) were digested with NcoI and NotI, and desired bands were excised following 1% TAE gel electrophoresis. DNA was purified using the Gel Extraction kit (Millipore; millipore.com) and ligated with Quick Ligase (New England Biolabs; neb.com). Following transformation and plating on selective media (Amp), plasmid constructs were miniprepped, diagnostically digested (NcoI/NotI), and sequence verified.

Heterologous expression and purification of intein proteins

To make homing endonuclease protein, overnight cultures were prepared of *E. coli* strain TOP10 containing the HE construct in the pBE expression vector. Cultures were used to inoculate 50 mL LB (with Amp) for an increase in absorbance of 5-10 Klett units. After 37°C incubation with shaking for 1-3 hours, until absorbance increased by 100 Klett units, expression of the HE construct was induced with arabinose (0.2% w/v). Following further incubation for 3-4 hours, cells were pelleted by centrifugation (5 minutes at 8000 rcf, 4°C) and frozen dry at -80°C.

In order to purify HE protein from the cell pellet, lysis, wash, and elution buffers were prepared (50 mM NaH_2PO_4, 300 mM NaCl, and either 10, 20, or 300 mM imidazole, respectively; all pH 7.5). Pelleted cells were thawed on ice, resuspended in 1-2 mL lysis buffer with lysozyme (1 mg/mL), and incubated on ice for 30 minutes. After vortexing three times for 10 seconds with 5 second pauses on ice between steps, cell lysate was centrifuged (30 minutes at 10,000 rcf, 4°C) and the supernatant was collected as the crude cell lysate. The Ni-NTA spin kit and protocol (Qiagen) were followed to isolate poly-His-tagged protein from the crude cell lysate. In brief, lysis buffer was used to equilibrate the Ni-NTA column (centrifuging for 2 minutes at 5,000 rcf) before crude cell lysate was loaded onto the column. After two wash steps, two eluates were collected in fresh tubes. Proteins were stored in glycerol (33.3%) at -80°C.

Construction of target site plasmids

Homing endonuclease target site plasmids were constructed in pGEXX or oXXI. Plasmid pGEXX is a derivative of the ampicillin-resistant pGEM-T Easy (Promega), with added XhoI and XbaI restriction sites separated by a small spacer. Plasmid oXXI (also known as pBR-O-Xho) is a tetracycline-resistant derivative of pBR322 (New England Biolabs), as described previously (Seligman *et al.*, 2002). Stocks of oXXI were prepared in the Dam- strain S1540 to avoid methylation of restriction sites in plasmids.

Putative HE target sites were designed using the NEB Intein Database (Perler, 2002). Initially, predicted homing sites were selected based on the 20 bp flanking each side of the intein *in vivo* in each species of interest (or flanking the predicted site of intein insertion in species lacking the intein). For site narrowing, bases were subtracted from either end of the target site. To make site plasmids, complementary oligonucleotides were

designed based on putative homing sites, with added 5' sequences to complement XhoI and XbaI sticky ends but avoid recreating these restriction sites. Oligonucleotides (synthesized by Eurofins MWG Operon) were annealed by combining 1X NEB buffer 2 and 2 uL each 5 uM oligonucleotide per 10 uL reaction, heating to 80°C for 5 minutes, and cooling to 4°C at 0.1%. Annealed oligos were subsequently ligated into *Xho*I/*Xba*I-digested plasmids (pGEXX or oXXI) using standard protocol for the Quick Ligation kit (New England Biolabs).

Homing endonuclease* in vitro *digests

In vitro homing endonuclease digests of target sites were carried out in 20:9 buffer (25 uL 1M Tris 9, 12.5 uL 1M MgCl2, 1.25 uL 1M DTT, 6.25 uL 100X BSA per 125 uL 10X stock). Target sites were linearized at 37°C with the appropriate HE (generally XmnI for oXXI and BsRBI for pGEXX plasmids). Immediately prior to addition, the hyperthermophilic homing endonuclease protein was heated to 90°C, 20 minutes to denature contaminating *E. coli* proteins. HE digests were then carried out at 80°C for 30 minutes or 1 hour, and analyzed on 0.8% agarose TBE gels. In order to prevent HEs from interfering with DNA migration, samples were loaded with dye containing SDS (0.06 g Bromophenol blue, 15 mL 50% glycerol, 7.5 mL 10% SDS per 25 mL 6X stock). Promega 1 kb ladder was also loaded for analysis. Following electrophoresis, gels were stained with EtBr for 25 minutes, destained in dH_2O for 10 minutes, and imaged with ImageJ.

Competition assays

In order to determine relative affinities of homing endonucleases for nonnative sites, *in vitro* competitive cleavage assays were carried out using the native and nonnative target site plasmids. First, the native site in oXXI and nonnative site in pGEXX were linearized in 20:9 buffer with, respectively, either XmnI or BsRBI. Following 2.5 hours 37°C digest and 30 minutes 80°C heat inactivation of restriction enzymes, equimolar linearized sites were pooled (generally 75 ng oXXI and 60 ng pGEXX per 20 uL pooled reaction) and split into one 40 uL and the desired number of 20 uL fractions. After the 90°C heat step, homing endonuclease protein was added to the 40 uL fraction and diluted by a series of 1:2 dilutions into the other fractions. Some 1:4 dilutions were necessary to assess cleavage by the *Tko* TopA HE on r-Gyr sites for which it had low affinity. Following digest at 80°C (for 30 minutes, 1 hour, or 2 hours and 15 minutes), competition assays were analyzed by gel electrophoresis. ImageJ was used to quantify percent cleaved for each site in each lane based on band intensity after subtracting background. Relative HE concentration required for 50% cleavage of each site was then calculated based on the values on either side of this benchmark. Relative preference '*C*' of the homing endonuclease for the nonnative site, compared to the native site, was calculated as follows.

$$C = \frac{\text{relative HE concentration for 50\% cleavage of nonnative site}}{\text{relative HE concentration for 50\% cleavage of native site}}$$

Sequence analysis
Computation of conservation among sites and inteins was done using the standard alignment tools in CLC Viewer. Sequence logos were generated using WebLogo (Crooks & Hon, 2004).

RESULTS

The M. jannaschii *r-Gyr,* T. kodakaraensis *r-Gyr, and* T. kodakaraensis *TopA inteins all have homing endonuclease activity*

In order to study the r-Gyr/TopA group of inteins, the *Methanococcus jannaschii* (*Mja*) r-Gyr intein and the *Thermococcus kodakaraensis* (*Tko*) r-Gyr and TopA inteins were selected for characterization. These inteins are the largest among members of this group listed in InBase (Figure 3) and contain many conserved intein blocks, including LAGLIDADG motifs; thus, these three inteins were predicted to be most likely to retain homing endonuclease activity. Each selected intein was amplified by PCR of genomic DNA from its respective host organism and cloned into an expression vector with an inducible promoter and a poly-His tag (Figure 4). Cloned inteins were heterologously expressed in *Escherichia coli* and purified by Ni++ affinity column.

In order to assess whether cloned inteins had endonuclease activity, oligonucleotides of each native homing site were cloned into plasmids (Figure 5). Predicted target sites were selected based on the 20 bp flanking each side of the intein *in vivo* (i.e. the sequence of an HEG- allele in the native species). *In vitro* HE digests of linearized site plasmids were conducted at 80°C since recombinant proteins were derived from hyperthermophiles.

The *Mja* r-Gyr, *Tko* r-Gyr, and *Tko* TopA inteins all cleaved their native sites *in vitro* (Figure 6). Heating the endonucleases prior to digest (90°C, 20 minutes) was found to improve the strength and clarity of bands on the gel, indicating that contaminating *E. coli* proteins – which are denatured by the heating step – are present in purified HE protein isolates. Additionally, the notable loss of DNA in the *Tko* r-Gyr digest, which was observed consistently, may be due in part to increased contaminants (salts and protein) from the higher volume of HE used in this digest in order to counter the lower activity of the *Tko* r-Gyr protein sample.

All three inteins cleave homologous sites in other species and in both related genes, with notable exceptions

In order to assess the broad target profiles for these HEs, site plasmids were prepared based on the r-Gyr and TopA genes from five species of hyperthermophilic archaea, some of which possess inteins at these loci *in vivo* (Figure 3). In addition to *Mja* and *Tko*, these species include three members of the genus *Pyrococcus*: *P. furiosus* (*Pfu*), *P. horikoshii* (*Pho*), and *P. abyssi* (*Pab*). Alignment of putative target sites (Figure 7) revealed a relatively high degree of divergence at the DNA level yet conservation at the intragenic protein level. Intergenic protein conservation was less pronounced, with 5/12 residues differing between the r-Gyr and TopA consensus sequences.

In vitro digests of each of the ten linearized site plasmids by each of the three HEs revealed distinctive specificity profiles for each HE (Figure 8). Notably, every HE cleaved the majority of sites tested, including most homologues and some sites from the other gene. The *Mja* TopA site was the most resistant to cleavage, and exhibited only low-level cleavage by both r-Gyr HEs. Disappearance of linearized site plasmid without appearance of product bands (e.g. the *Tko* r-Gyr HE digest of *Tko* TopA site in Figure 8A) could indicate failure of the HE to unbind the DNA, preventing proper migration on the gel. However, samples were run with loading dye containing SDS, casting doubt on this explanation. The *Tko* TopA HE cleaved ectopic sites on both types of site plasmids under the reaction conditions, but only when affinity for the intended target site was sufficiently low (see first five lanes of *Tko* TopA gels in Figures 8A and 8B). Thus, it was difficult to ascertain whether the *Tko* r-Gyr and TopA HEs were capable of cleaving the reciprocal site in the other gene. These results showed that the *Mja* r-Gyr, *Tko* r-Gyr, and *Tko* TopA HEs were each capable of cleaving multiple sequences besides the cognate site, leading to the question of how affinity for these foreign sequences would compare to affinity for the native site at which each intein is inserted in its respective host genome.

Each intein displays a distinctive specificity profile, cleaving many nonnative sites with comparable affinity to the native site and preferring the native site to others

In order to assess the relative cleavage affinities of cloned inteins for various non-cognate sequences, a competitive cleavage assay was conducted for each of the three HEs with each of the nine nonnative sites from five archaeal species (Figures 9-35). For each

assay, the native and nonnative target sites were linearized separately, combined in equimolar amounts, and split into a number of fractions. Following a 90°C heating step to ensure denaturation of *E. coli* proteins, homing endonuclease protein was added to one fraction and serially diluted into the others in order to assess cleavage preference at a range of HE concentrations. Digest conditions (starting amount of HE and digest time, from 30 minutes to 2 hours, 15 minutes) were manipulated in attempt to ensure that samples ran the gamut from 100% cleavage of both sites to no cleavage. Following incubation at 80°C, digests were visualized by gel electrophoresis. Because the native and nonnative sites were cloned into two different plasmids and linearized by different restriction enzymes, substrate and product bands for the two sites could be resolved and used to calculate the percent cleaved for each site in each lane. Cleavage preference was quantified from each assay by calculating *C*, the HE concentration required for 50% cleavage of the nonnative site relative to that required for equivalent digest of the native site. Thus, a *C* value close to one corresponded to equal affinity of the homing endonuclease for the two sites, while $C > 1$ indicated preference for the native site, and $C < 1$ indicated preference for the nonnative site.

Competition assays with the *Mja* r-Gyr HE indicated that this enzyme cleaved all sites tested, with notably higher preference for sequences from r-Gyr over those from TopA. The HE displayed similar affinity for all r-Gyr sites tested compared to its cognate site (Figures 9-12), with ~1.25-fold (i.e. C^{-1}) preference for the nonnative *Pfu* and *Pab* r-Gyr sites (Figures 10, 12). On the other hand, the HE displayed lower affinity for all TopA sites tested (Figures 14-17). Magnitude ranged from ~2-fold preference for the native site over *Tko* and *Pab* TopA (Figures 14, 17), to ~5-6-fold preference for the native site over *Pfu* and *Pho* TopA (Figures 15, 16), to >20-fold preference for the native site over the TopA sequence from the native *Mja* host genome (Figure 13). The latter competition assay is also notable because >60% cleavage of the *Mja* TopA site was obtained by increasing HE quantity and digest time, despite very low-level cleavage noted in previous digests (see *Mja* r-Gyr gel, MT lane in Figure 8A).

Competition assays with the related *Tko* r-Gyr HE revealed a different profile of cleavage preferences. The HE displayed ~2-fold preference for the native site over the *Mja* r-Gyr site (Figure 18), but similar affinity for the three *Pyrococcal* r-Gyr sites

(Figures 19-21). For sequences from the other gene, cleavage preferences ranged widely. The HE was unable to cleave both the *Mja* TopA site and the *Tko* TopA site from its host genome under all digest conditions tried (Figures 22, 23). On the other hand, the HE displayed ~2-2.5-fold preference for the native site over the *Pfu* and *Pho* TopA sites (Figures 24, 25), and cleaved *Pab* TopA with comparable affinity to its native site (Figure 26).

Results from competition assays using the *Tko* TopA HE were strikingly different from those using the two r-Gyr HEs. Only low-level cleavage of the *Mja*, *Tko*, *Pfu*, and *Pho* r-Gyr sites was observed despite increasing HE concentration and digest time (Figures 27-30). Because calculating *C* requires >50% cleavage of the native and nonnative sites, relative preference could not be quantified for these sequences. Affinity for the *Pab* r-Gyr site was also low, but up to 80% cleavage was observed, yielding a calculation of ~10-fold preference for the native site over this sequence. On the other hand, affinity for all TopA sites was comparable to that for its native site (Figures 32-35).

Results from all 27 competition assays were aggregated in order to facilitate recognition of trends in HE specificity profiles (Table 1). Upon comparison of results, the *Mja* TopA target was noted as the site most resistant to cleavage overall, inviting further investigation.

Ile-to-Cys mutation of the **M. jannaschii** *TopA site does not alter susceptibility to HE cleavage*

The dramatically poor cleavage of the *Mja* TopA site in competition assays with both r-Gyr HEs confirmed previous results of *in vitro* digests (see Figure 8A, MT lane in *Mja* r-Gyr and *Tko* r-Gyr gels). Upon reexamination of target site alignments (Figure 7), *Mja* TopA was discovered to be the only cloned sequence with an isoleucine residue following the hypothetical site of intein insertion. All other r-Gyr and TopA targets had cysteine, serine, or threonine at this position – the only residues that would be capable of facilitating the splicing reaction that separates an intein and an extein into two mature proteins (Perler, 2002).

In order to investigate the effect of this codon on this sequence's susceptibility to HE cleavage, the *Mja* TopA (I→C) target site was cloned. This site introduced an Ile-to-Cys mutation at the first residue of the extein downstream of the hypothetical site of

intein insertion, using the corresponding codon from *Mja* r-Gyr (Figure 36A). If the presence of the Ile codon were solely responsible for this site's resistance to cleavage, the introduction of this mutation was expected to improve cleavage of the *Mja* TopA target site.

To this end, the *Mja* TopA (I→C) plasmid was linearized and digested with each of the three HEs (Figure 36B). Whereas the *Tko* TopA HE cleaved the mutated site to near completion, the *Mja* r-Gyr HE exhibited partial cleavage under digest conditions. The *Tko* r-Gyr HE did not cleave the site at all, and substantial DNA loss was noted. These results were consistent with the relative cleavage affinities for wild-type *Mja* TopA in competition assays (see Table 1 for summary and Figures 13, 22, 32 for assays).

In order to directly compare cleavage of the mutant sequence to cleavage of the wild-type sequence, both versions of the *Mja* TopA target site were digested with each HE under identical conditions (Figure 36C). All three HEs cleaved (or failed to cleave) both sites equally. Notably, the *Mja* r-Gyr HE did not cleave either site to any measurable extent, whereas partial cleavage of the wild-type or mutant site was observed previously (see Figure 8A, *Mja* r-Gyr gel, MT lane and Figure 36B, respectively). The only differences in reaction conditions were minor variations in quantities of DNA and enzyme. Thus, the *Mja* r-Gyr HE inconsistently cleaves both *Mja* TopA sites similarly, and both HEs from *Tko* also have equal affinities for the mutant and wild-type sites.

The minimal homing sequence for the* M. jannaschii *r-Gyr HE is 15 to 23 bases long

In order to determine the minimal target required for homing and cleavage by the *Mja* r-Gyr HE, native site plasmids were constructed with narrowed target sites and assessed for cleavage following HE digest. After narrowing eight bases from either the 3' or 5' end of the initial 40 bp target site, the *Mja* r-Gyr HE successfully cleaved both the 1-32 and 9-40 narrowed sites (Figure 37). Thus, both truncations were combined, and an additional four bases were narrowed from either end. However, both the 9-28 and 13-32 site plasmids were resistant to cleavage (Figure 38). The minimal target site required for cleavage therefore lies from some base between bases 9 and 13, to some base between bases 28 and 32 (with bases numbered according to their position in the original 40 bp target site, Figure 7).

DISCUSSION

Confirmation of sequence-based prediction of homing endonuclease activity

This investigation conclusively established that the inteins from *M. jannaschii* r-Gyr, *T. kodakaraensis* r-Gyr, and *T. kodakaraensis* TopA are active homing endonucleases. Previously, these inteins were predicted to have HE activity based on the presence of conserved protein domains, including the LAGLIDADG motif from which this family of HEs derives its name (Chute *et al.*, 1998; Perler, 2002). Here, I show via *in vitro* analysis that these sequence-based predictions are accurate (Figure 6).

This finding invites the question of how these inteins currently fit into the stages of cyclical intein evolution. According this simplified model, inteins are either in the process of actively spreading throughout the population, or they have reached fixation and are in a state of gradual degeneration (Burt & Koufopanou, 2004). It is unknown how the representative archaeal genomes from which these sequences were cloned reflect the dynamic distribution of these inteins among populations. Thus, we do not know whether these inteins have saturated the available sites in their native species. Furthermore, limited gene flow has been observed among spatially segregated populations of hyperthermophilic archaea in hydrothermal vents (Whitaker, Grogan, & Taylor, 2003), which all of the species in this investigation co-habit (Takai & Nakamura, 2011). Thus, infrequent gene exchange among sub-populations could be sufficient to maintain these inteins in a state of active spread, even if they have reached local fixation. Future studies could investigate the population dynamics of intein propagation by sequencing multiple representative genomes within a single population, as well as sampling across segregated sub-populations, such as those found in deep-sea vents.

Confirmation of HE activity suggests that the inteins in this investigation may still be in the midst of propagation throughout the population. However, the poorer activity of the *Tko* r-Gyr HE (see, e.g., Figure 6) could indicate that this intein is in the process of degeneration, meriting further examination of sequence alignments. In support of this, the *Tko* r-Gyr HE has the shortest sequence of the three cloned, although the *Mja* r-Gyr HE – at 494 residues – is only 5 amino acids longer. Alternately, this observation could be a function of the protein sample used in this investigation. Tagged inteins were prepared identically, but some sequences could be more amenable to heterologous expression and

protein purification. Quantification of HE protein, as well as repetition of digests with fresh protein samples, may begin to resolve this issue.

Now that three of the members of the r-Gyr/TopA group of inteins have been cloned and characterized, *in vitro* analysis of homologues in other species will be informative. For example, the status of the other inteins from the loci in this investigation – namely, the *P. horikoshii* r-Gyr and *P. furiosus* TopA inteins – remains unknown. While these inteins have accumulated significant deletions (on the order of ~100 amino acids versus those in this investigation; Figure 3), future studies should clone these proteins to confirm their activity status. If active, comparison of specificity profiles based on similar *in vitro* assays could provide further insight into the evolution and uneven distribution of this group of inteins.

Furthermore, InBase, the NEB Intein Database (Perler, 2002) represents an incomplete record of the inteins in this group. Protein BLAST of the *Mja* r-Gyr intein reveals multiple hits in sequenced genes from other archaea (Table 2). Analysis of these sequences will yield a more complete picture of the uneven distribution of the r-Gyr and TopA inteins among hyperthermophiles. In addition, some of these inteins may also be candidates for future cloning and *in vitro* characterization.

In vitro *cleavage preferences and intein distribution*

The findings of this investigation support the idea that HE specificity is a significant factor in intein distribution, but is certainly not the sole factor at work. Comparison of cleavage preferences, as measured in competition assays, with the distribution of the r-Gyr/TopA inteins *in vivo* reveals both interesting overlaps and interesting exceptions.

In general, sites with inteins were not necessarily more amenable to cleavage. For example, the *Mja* r-Gyr site, which contains an intein, was the r-Gyr site cleaved most poorly by the *Tko* r-Gyr HE, and was barely cleaved at all by the *Tko* TopA HE. On the other hand, both *Pab* sites, which lack inteins, were cleaved best among other sites in the same gene by all three HEs (Table 1).

However, results support my hypothesis that HE specificity is responsible for the uneven distribution of the intein in *M. janaschii* (i.e. in *Mja* r-Gyr but not *Mja* TopA). The *Mja* r-Gyr HE cleaved the *Mja* TopA site inconsistently, with low affinity, and only

under specific digest conditions (Figures 8A, 36C). In fact, the HE had the lowest preference for this site among all tested (Table 1). Since *in vitro* digest conditions are already extreme compared to conditions in a cell, low-level cleavage of the nonnative site is likely not relevant *in vivo*. Thus, the inability of the *Mja* r-Gyr HE to cleave the *Mja* TopA site explains the lack of an intein at that locus. This is similar to a previous report of the stable coexistence of HEG+ and HEG- alleles, in which a tetraploid yeast strain harbored both an intein+ and mutant intein- allele, which was only cleaved under relaxed digest conditions (Gimble, 2001).

In the case of *P. furiosus* and *P. horikoshii*, results for these target sites reveal other insights into intein distribution in these species. Each of the three cloned HEs displayed similar affinity for the *Pfu* and *Pho* r-Gyr sites, and for the *Pfu* and *Pho* TopA sites (Table 1). Thus, inherent sequence differences between the two species do not explain why inteins are distributed oppositely (i.e. in *Pfu* TopA but not r-Gyr, and in *Pho* r-Gyr but not TopA). Rather, my results suggest that these species inherited different progenitor inteins with lopsided specificity profiles. For example, the *Tko* TopA HE cleaves *Pfu* TopA but not r-Gyr; thus, invasion of *Pfu* by this HE would likely result in the observed distribution of the intein in this species. On the other hand, the *Tko* r-Gyr HE has relatively high preference for both loci in *Pfu* and *Pho*, so (in its current state of evolution) is unlikely to be the progenitor intein in these species.

The results of this investigation do not support the theory that the intein spread from r-Gyr to TopA (or vice versa) in *T. kodakaraensis*. Contrary to expectations, the *Tko* r-Gyr and *Tko* TopA inteins do not cleave one other's sites (Figures 23, 28). Furthermore, sequence homology between these inteins is only 10.7%. Low homology is often observed among related inteins and may reflect accelerated evolution of inteins compared to host proteins (Chute *et al.*, 1998), and – in this case – increased genomic plasticity among hyperthermophiles (Lecompte *et al.*, 2001). However, this unusually low figure casts doubt on the direct relatedness of these inteins. Note, however, that the *Tko* r-Gyr and TopA inteins are still considered homologous since the defining characteristic of allelic inteins is sharing a common insertion site (Chute *et al.*, 1998) – which these inteins do with respect to host gene structure, if not function.

Thus, it seems likely that the *Tko* r-Gyr and TopA inteins originated in this species through two independent invasions by related inteins. This situation could have occurred through invasion by two progenitor inteins with high affinity for only one locus. Alternately, following invasion by the first intein, *Tko* could have been invaded by an intein that is capable of cleaving both sites, such as the *Mja* r-Gyr HE. Since one locus would already contain an intein, just one of the target sequences would be intact, and the invading intein would only be transmitted to the empty site. In fact, the theory of two invasion events could explain the unusually low affinity of *Tko* r-Gyr for the TopA site. If there were a period of time between invasions during which *Tko* only possessed an intein in r-Gyr, this intein would be under selective pressure to avoid cleavage of *Tko* TopA if ectopic cleavage were lethal and failed to produce intein transmission. On the other hand, the *Mja* r-Gyr intein – which would not have experienced this selective pressure – cuts *Tko* TopA relatively well, with only ~2-fold preference for the native site (Figure 14). Because the cleavage profiles of both r-Gyr inteins otherwise resemble one another (Table 1), the inability of the *Tko* r-Gyr HE to cleave the TopA site supports the theory of particular selective pressures rather than broad evolution of tighter specificity.

If the intein did not spread from one gene to the other in *Tko*, this raises the question of how this jump did occur. Did the intein spread within species or across species? *Tko* is the only species identified with inteins at both loci, even among the expanded catalogue of r-Gyr/TopA inteins (Table 2). However, the jump could have occurred within some yet-unknown species with both inteins. On the other hand, the intein could have been transmitted across species, from one locus in one species to the other locus in another species. This mode of transmission would require DNA repair using a functionally distinct gene paralog from another species as a template. The feasibility of this event is supported by evidence of gene conversion between divergent paralogs in archaea (Archibald & Roger, 2002), as well as interspecific homologous recombination between distant archaeal lineages (Inagaki *et al.*, 2006). Notably, to preserve uneven distribution, an interspecific jump would have to be catalyzed by an HE with higher affinity for the nonnative site in the opposite gene than for the nonnative site in the same gene, which was not observed in this investigation. Alternately, the intein could have spread within species and subsequently been lost in one gene through precise

loss, which is unlikely but possible over evolutionary timescales (Burt & Koufopanou, 2004; Posey *et al.*, 2004). *In vitro* and *in silico* analysis of HEs and sites from the expanded list of r-Gyr/TopA inteins may provide insight.

In contrast to the previous examples, cleavage preferences do not seem to contribute to the lack of these inteins in *P. abyssi*, which possesses both loci. All three cloned HEs exhibited relatively strong preferences for both *Pab* sites versus other nonnative sites in the same gene (Table 1). Thus, the absence of the r-Gyr/TopA inteins in *Pab* is not due to the inability to cleave these sites, meriting consideration of other factors governing intein dispersal. For example, the absence of these inteins in *Pab* may be due to a lack of exposure to the intein. However, the five species in this investigation all share other inteins in common (Figure 39) and live in proximity in hydrothermal vents (Takai & Nakamura, 2011). Species differences in the ability to uptake naked DNA, which is required for intein transmission, may also be at work. For example, *T. kodakaraensis* is naturally competent (Sato *et al.*, 2003), as is a particular mutant of *P. furiosus* but not all members of the species (Lipscomb *et al.*, 2011). Intein transmission may also be affected by the locations of the r-Gyr and TopA genes on the chromosome, which differ among species in strand polarity and relative distance (Chute *et al.*, 1998). Finally, intein dispersal may be affected by the length of flanking homology between donor and recipient loci, which is correlated with efficiency of intein transfer (Naor *et al.*, 2011). This factor may be especially relevant in the case of r-Gyr and TopA, since cleaved sites may rely on repair using a functionally distinct gene as a template.

In vitro *cleavage preferences and homing strategies*

Although this investigation does not directly address the molecular basis for homing specificity, different patterns of cleavage preference are the result of underlying differences in homing strategy based on different protein-DNA contacts (Rosen *et al.*, 2006). Thus, the preference profiles generated by competition assays may indirectly provide insight into molecular mechanisms of homing.

Comparison of trends in specificity profiles (Table 1) reveals distinctive strategies for cleavage. While all three HEs were promiscuous, cleaving the majority of sequences tested, the r-Gyr HEs were capable of acting on a wider pool of substrates. Though *Tko* TopA exhibited ectopic cleavage on both site plasmids (see first five lanes of *Tko* TopA

gels, Figure 8), this behavior is likely an artifact of *in vitro* digest conditions rather than an indication of promiscuity. Thus, the distinction in relative promiscuity among these HEs may reflect distinctive evolutionary strategies: higher promiscuity mitigated by decreased activity versus tighter specificity and increased activity (Li *et al.*, 2012).

Furthermore, order of site preferences of both r-Gyr HEs is nearly identical, and is almost diametrically opposed to that of the *Tko* TopA HE. Thus, both r-Gyr HEs likely make similar base pair contacts, other than some aforementioned difference affecting recognition and cleavage of the *Tko* TopA site. On the other hand, the *Tko* TopA HE seems to make very different base pair contacts. This hypothesis is consistent with the low sequence identity between *Tko* TopA and the other inteins (11% with *Tko* r-Gyr and 8.8% with *Mja* r-Gyr, versus 38% between the r-Gyr inteins).

However, one limit of competition assays is that results are relative to the affinity of each HE for its native site, which is not necessarily the optimal substrate (Scalley-Kim *et al.*, 2007). Thus, future studies could build on this investigation by biochemically characterizing these enzymes with standardized parameters, such as substrate affinity, rate constant, and binding coefficient, in order to directly compare cleavage across HEs.

Consideration of cleavage preferences in light of base pair differences between sites yields additional insight into the molecular basis for specificity. Overall, the association between divergence from the native site and relative preference for that site was slightly negative, as might be expected. For example, the *Tko* TopA HE cleaved the *Pab* site best among r-Gyr sites, and this site was the least divergent from the 40 bp native site (14 versus 15-19 base pair differences). However, this trend was very weak; in general, base pair difference between native and nonnative sites was *not* strictly predictive of cleavage preference. For example, the *Mja* r-Gyr and *Pab* TopA sites are the most highly divergent pair (21 base pair differences), but the nonnative site is cleaved over half as well as the native site (Figure 17). The lack of a simple relationship between site sequence homology and cleavage preference reinforces the nonrandom nature of DNA contacts made by these HEs.

In fact, the enzymes in this investigation often exhibited higher cleavage preference for homologous sites in the same gene than for sites in the opposite gene despite comparable differences in magnitude of divergence. For example, the 40 bp *Mja*

r-Gyr native site differs similarly from *Tko* and *Pfu* r-Gyr as it does from *Mja* TopA (at 15 and 16 bases, respectively), but this HE cleaves the former two equally to or better than the native site yet barely cleaves the latter. Similarly, the *Tko* TopA site differs from both sites in *Pho* by 17 base pairs, but the intein only cleaves the sequence from the TopA gene. In fact, this phenomenon is most pronounced in the *Tko* TopA HE, which cleaves the *Mja* and *Pab* TopA sites well (16 base pair differences) yet is barely capable of cleaving the *Mja*, *Pfu*, and *Pab* r-Gyr sites at all (16 or fewer differences). Thus, gene identity was more predictive of cleavage preference than was DNA sequence homology alone. Since consensus at the protein level differs at 5/12 residues between the r-Gyr and TopA sites (Figure 7), this finding is in line with previous evidence that HEs co-evolve with their target sites in order to accommodate genetic drift within the evolutionary constraints of their respective host genes (Koufopanou *et al.*, 2002; Scalley-Kim *et al.*, 2007).

In order to examine the molecular basis for specificity in more detail, sequences cleaved with high preference were aligned and contrasted to sequences cleaved with low preference. For each enzyme, sequence logos were generated based on the sites cleaved with similar affinity as the cognate site in order to identify conserved bases and quantify information content at each position (Figures 40A, 41A, 42A). Sequence logos were aligned with poorly cleaved sites in order to identify individual bases that deviated from all of the best-cleaved sequences (Figures 40B, 41B, 42B). Positions along the 40 bp target were hypothesized to contribute to HE specificity if they were both conserved among sites amenable to cleavage and changed in sites resistant to cleavage.

This analysis yielded 15 sites of potential DNA-protein contact with the *Mja* r-Gyr HE, 12 with the *Tko* r-Gyr HE, and 7 with the *Tko* r-Gyr HE. Notably, due to the limited scope and nonrandom nature of sites tested, these numbers are not necessarily predictive of the relative number of contacts made by each HE. Among these sites, the -4 position (in relation to the site of intein insertion) was the only prediction common to all three HEs. Future investigations should test affinity for -4 site mutants in order to determine whether this position represents a conserved contact for these HEs. On the other hand, approximately 30% of the identified bases for each HE were unique, representing predictions for the molecular basis of distinctive specificity.

Surprisingly, many of these bases identified as important for cleavage corresponded to wobble positions. This was the case for 6/15 potential contacts for *Mja* r-Gyr, 6/12 for *Tko* r-Gyr, and 3/7 for *Tko* TopA. These numbers are in contrast to the null hypothesis of one in three, and even fewer if these HEs have evolved to tolerate variability at wobble positions. Thus, this analysis does not support the hypothesis that HE specificity is correlated to genetic code degeneracy (Scalley-Kim *et al.*, 2007). However, this result likely reflects the limited scope of this experiment rather than a refutation of this hypothesis. For example, in order to be identified as important for cleavage, base pairs had to differ from those in better-cleaved sites. Since wobble positions are more likely to differ among homologues in general, this analysis may have been biased toward identifying bases in these positions. On the other hand, bases that are highly conserved across all ten sites tested – prime potential contributors to HE specificity – would not be identified. In order to test this hypothesis without bias, future experimenters could directly measure the importance of each DNA base for cleavage based on HE digest of site libraries with random point mutations to determine whether these results are correlated with the degree of conservation of each base pair in r-Gyr or TopA. Furthermore, determination of the crystal structures of these enzymes bound to cognate and non-cognate substrates would yield invaluable insight into DNA-protein contacts.

Despite aforementioned limitations, this analysis provides a preliminary look at providing a molecular explanation for cleavage preferences among these HEs. For example, the *Mja* TopA site, which was exceptionally resistant to cleavage by both r-Gyr HEs, differs at only unique two positions (-5 and +2) from other sites that are cleaved with better affinity by these enzymes (Figures 40B, 41B). Ile-to-Cys mutation of the codon following the hypothetical site of intein insertion does not affect the site's amenability to cleavage (Figure 36), indicating that the identity of the +2 position may not actually be critical to cleavage. Taken together, these data suggest that the *Mja* TopA site's resistance to cleavage could derive from the presence of cytosine at the -5 position. This hypothesis could be tested by comparing cleavage preference for the *Mja* TopA wild-type site versus sites with point mutations at this position. Thus, these data provide

easily testable predictions for bases that contribute to the unique site specificity of each HE, laying the groundwork for future investigations.

Increased selective pressures on the **M. jannaschii** *TopA site*

The results of this investigation suggest that the *Mja* TopA site, the target site most resistant to cleavage overall, may represent a case of exceptional selective pressure to resist intein invasion. The *Mja* TopA site was cleaved extremely poorly by *Mja* r-Gyr and was not cleaved at all by *Tko* r-Gyr (Table 1). Due to the uneven distribution of the intein in *Mja*, the TopA site may be continuously at risk of cleavage. Notably, this site is the only one of the ten homologues to contain an Ile residue immediately downstream of the hypothetical site of intein insertion (Figure 7). The intein splicing reaction is known to rely on nucleophilic attack by Cys, Ser, or Thr at this position in the extein (Perler, 2002) – the residues conserved in all nine other loci. Thus, invasion of the *Mja* TopA site would be toxic since the intein would not be properly spliced out of the extein, an essential host gene that regulates DNA topology (Chute *et al.*, 1998). Due to these higher stakes for cleavage, I argue that the *Mja* TopA site is under increased selective pressure to accumulate mutations that resist cleavage.

Examination of differences in cleavage preference and base pair sequence between sites supports this conclusion. *Mja* r-Gyr differs from the TopA site at 15 out of 40 bases, equal to or fewer than all other TopA sites, yet the *Mja* TopA site is cleaved the most poorly by an order of magnitude (Table 1). This finding is consistent with the theory that this site is enriched in random mutations that decrease HE affinity.

This theory is also consistent with the discovery that cleavage of the *Mja* TopA site was unchanged by the replacement of Ile with the conserved Cys (using the codon from the native *Mja* r-Gyr site) (Figure 36). If the mutated site were cleaved with high affinity, this would imply that the Ile codon was largely responsible for the *Mja* TopA site's resistance to cleavage, meaning that other mutations were incidental. Though digests should be repeated due to inconsistencies in HE cleavage, it remains clear that the Ile codon is not solely responsible for the site's resistance to cleavage. While this negative result does not conclusively show that these mutations have accumulated as a

result of increased selection, it discredits an opposing hypothesis that the Ile mutation alone is sufficient to prevent toxic invasion of the *Mja* TopA site.

In fact, the *Mja* TopA site is the most highly divergent sequence at the protein level, which is notable since it occurs around the catalytic residue in a highly conserved region of a highly conserved gene (Chute *et al.*, 1998). Has *Mja* taken a hit in gene function to avoid the detrimental fitness cost of cleavage? Future experiments could evaluate this theory by aligning all sequenced TopA loci and using the consensus sequence to determine the rate of mutation in *Mja* TopA at the target site compared to the rest of the gene.

Minimal requirements for HE cleavage

Because the *Mja* r-Gyr HE cleaved sites with eight bases narrowed from either end, but did not cleave sites with these truncations and an additional four bases narrowed, the minimal HE recognition sequence was determined to be 15 to 23 bases long (Figures 37, 38). However, this conclusion assumes that cleavage of the 1-32 and 9-40 sites implies cleavage of the 9-32 site, although it is possible that the HE could require more bases on one side if the other side were truncated. Furthermore, eliminating bases that serve as points of DNA-HE contact may still permit cleavage, though affinity would be decreased. Thus, in addition to continuing the site narrowing protocol to determine the minimal site that permits cleavage, future experimenters should also determine the minimal site that does not affect HE affinity. The latter could be accomplished by using competition assays to compare cleavage preferences for narrowed sites to the full 40 bp site.

Once both types of minimal sites are determined for the *Mja* r-Gyr HE, similar narrowed sites should be tested for the *Tko* r-Gyr and TopA HEs in order to compare conservation of cleavage requirements among related HEs. Furthermore, determination of minimal sites will facilitate comparison of base pair differences between sites with HE affinity for these sites in order to determine bases that are important for cleavage. While full 40 bp sites were analyzed in this discussion for the reasons detailed above, further characterization of minimal sites will enhance this analysis.

Expanding on future directions

In addition to future experiments that directly build upon this thesis, as detailed in previous discussion sections, other lines of investigation may prove fruitful in the study of HE evolution and spread.

The classification of homing endonucleases into LAGLIDADG and other families implies that related HEs arose from common 'ancestors' (Chevalier & Stoddard, 2001; Dalgaard *et al.*, 1997) Does the presence of homologous HEs in functionally distinct genes represent a common step in the dispersal of HEs to new insertion sites? This idea could be explored by studying intein insertion sites in different host exteins to determine whether it is possible to cluster inteins by the closeness of their insertion site sequence. Similar analysis could also be conducted on intein sequences themselves. As illustration of this possibility, inteins in unrelated host proteins also turned up in the BLAST search of the *Mja* r-Gyr intein (Table 2).

Another mystery that remains is why species – like *P. abyssi* and the four other archaea in this investigation – share some inteins but not others (Figure 39). In order to investigate the dynamics of intein spread, species containing inteins could be analyzed for patterns in distribution. How do species compare in the number and identities of inteins that they contain? Are there clusters of particular shared inteins? Are clusters discrete, or do some species serve as hubs?

The experiments described here better characterize a unique case of related homing endonucleases inserted in functionally distinct genes, providing a foundation for future exploration of how the properties of HEs and their target sites relate to intein evolution and distribution *in vivo*.

ACKNOWLEDGEMENTS

Thank you to Professor Seligman for mentorship and encouragement throughout all four years at Pomona.

Thank you to Nick, Duncan, Francis, and Claudia for lab camaraderie.

Thank you to my parents, for listening to lab gripes and triumphs alike.

I'd also like to thank Walter J. Stutzman and Pomona College SURP for summer research funding, Rancho Santa Ana Botanic Garden for sequencing, Matt Reilly for help quantifying competition assays, Professor Sazinsky for biochemistry advice, Zack Mirman for mentorship, and 8tracks for quality tunes.

REFERENCES

Archibald, J., & Roger, A. (2002). Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *Journal of molecular biology*. doi:10.1006/jmbi.2001.5409

Burt, A., & Koufopanou, V. (2004). Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Current opinion in genetics & development*, *14*(6), 609–15. doi:10.1016/j.gde.2004.09.010

Chevalier, B., & Stoddard, B. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic acids research*, *29*(18), 3757–3774.

Chute, I. C., Hu, Z., & Liu, X. Q. (1998). A topA intein in Pyrococcus furiosus and its relatedness to the r-gyr intein of Methanococcus jannaschii. *Gene*, *210*(1), 85–92.

Crooks, G., & Hon, G. (2004). WebLogo: a sequence logo generator. *Genome Research*, 1188–1190. doi:10.1101/gr.849004.1

Dalgaard, J. Z., Klar, a J., Moser, M. J., Holley, W. R., Chatterjee, a, & Mian, I. S. (1997). Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic acids research*, *25*(22), 4626–38.

Gimble, F. S. (2001). Degeneration of a homing endonuclease and its target sequence in a wild yeast strain. *Nucleic acids research*, *29*(20), 4215–23.

Inagaki, Y., Susko, E., & Roger, A. J. (2006). Recombination between elongation factor 1alpha genes from distantly related archaeal lineages. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4528–33. doi:10.1073/pnas.0600744103

Koufopanou, V., Goddard, M. R., & Burt, A. (2002). Adaptation for horizontal transfer in a homing endonuclease. *Molecular biology and evolution*, *19*(3), 239–46.

Lecompte, O., Ripp, R., & Puzos-Barbe, V. (2001). Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. *Genome Research*, *11*, 981–993. doi:10.1101/gr.165301.1

Li, H., Ulge, U. Y., Hovde, B. T., Doyle, L. a, & Monnat, R. J. (2012). Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. *Nucleic acids research*, *40*(6), 2587–98. doi:10.1093/nar/gkr1072

Lipscomb, G. L., Stirrett, K., Schut, G. J., Yang, F., Jenney, F. E., Scott, R. a, Adams, M. W. W., *et al.* (2011). Natural competence in the hyperthermophilic archaeon Pyrococcus furiosus facilitates genetic manipulation: construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases. *Applied and environmental microbiology*, *77*(7), 2232–8. doi:10.1128/AEM.02624-10

Naor, A., Lazary, R., Barzel, A., Papke, R. T., & Gophna, U. (2011). In vivo characterization of the homing endonuclease within the polB gene in the halophilic archaeon Haloferax volcanii. *PloS one*, *6*(1), e15833. doi:10.1371/journal.pone.0015833

Perler, F. B. (2002). InBase: the Intein Database. *Nucleic acids research*, *30*(1), 383–4.

Pietrokovski, S. (2001). Intein spread and extinction in evolution. *Trends in genetics : TIG*, *17*(8), 465–72.

Posey, K. L., Koufopanou, V., Burt, A., & Gimble, F. S. (2004). Evolution of divergent DNA recognition specificities in VDE homing endonucleases from two yeast species. *Nucleic acids research*, *32*(13), 3947–56. doi:10.1093/nar/gkh734

Rosen, L. E., Morrison, H. a, Masri, S., Brown, M. J., Springstubb, B., Sussman, D., Stoddard, B. L., *et al.* (2006). Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic acids research*, *34*(17), 4791–800. doi:10.1093/nar/gkl645

Sato, T., Fukui, T., Atomi, H., & Imanaka, T. (2003). Targeted Gene Disruption by Homologous Recombination in the Hyperthermophilic Archaeon Thermococcus kodakaraensis KOD1. *Journal of bacteriology*. doi:10.1128/JB.185.1.210

Scalley-Kim, M., McConnell-Smith, A., & Stoddard, B. L. (2007). Coevolution of a homing endonuclease and its host target sequence. *Journal of molecular biology*, *372*(5), 1305–19. doi:10.1016/j.jmb.2007.07.052

Seligman, L., Chisholm, K., Chevalier, B., Chadsey, M., Edwards, S., Savage, J., & Veillet, A. (2002). Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic acids research*, *30*(17), 3870–3879.

Swithers, K. S., Senejani, A. G., Fournier, G. P., & Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC evolutionary biology*, *9*, 303. doi:10.1186/1471-2148-9-303

Takai, K., & Nakamura, K. (2011). Archaeal diversity and community development in deep-sea hydrothermal vents. *Current opinion in microbiology*, *14*(3), 282–91. doi:10.1016/j.mib.2011.04.013

Takeuchi, R., Lambert, A. R., Mak, A. N.-S., Jacoby, K., Dickson, R. J., Gloor, G. B., Scharenberg, A. M., *et al.* (2011). Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(32), 13077–82. doi:10.1073/pnas.1107719108

Viard, T., Lamour, V., Duguet, M., & Bouthier de la Tour, C. (2001). Hyperthermophilic topoisomerase I from Thermotoga maritima. A very efficient enzyme that functions independently of zinc binding. *The Journal of biological chemistry*, *276*(49), 46495–503. doi:10.1074/jbc.M107714200

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, *301*(5635), 976–8. doi:10.1126/science.1086909

Windbichler, N., Menichelli, M., Papathanos, P. A., Thyme, S. B., Li, H., Ulge, U. Y., Hovde, B. T., *et al.* (2011). A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature*, *473*(7346), 212–5. doi:10.1038/nature09937