Discovering and Characterizing New Homing Endonucleases for Genome Engineering

Kyle M. Jacoby

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

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Dr. Andrew Scharenberg, Chair
Dr. Stanley Fields
Dr. Philip Bradley

Program Authorized to Offer Degree:
Molecular and Cellular Biology

Abstract

University of Washington

Discovering and Characterizing
New Homing Endonucleases for
Genome Engineering

Kyle M. Jacoby

Chair of the Supervisory Committee:
Dr. Andrew Scharenberg
Adjunct Associate Professor
Department of Immunology

LAGLIDADG Homing Endonucleases (LHEs) are a family of highly specific DNA-cutting enzymes capable of recognizing target sequences of ~20 bp. In many eukaryotes, including humans and yeast, double-strand breaks induced by LHEs stimulate repair by Homologous Recombination, which can be used to alter or repair a gene if the template is supplied in trans, and Non-Homologous End Joining, which can be used to knock out a gene. The potential for such precise genome editing would reduce worry about insertional mutagenesis or misregulation, as only the specific gene under its native promoter would be targeted. Thus, LHEs have drawn intense interest for their research, biotech and clinical applications.

Methods for rational engineering of LHEs have been limited by a small number of high quality starting enzymes, and an extremely restricted understanding of how to modify them to

create novel enzymes that efficiently cleave hybrid target sequences. Here I describe my attempts to address these limitations by using a homology-directed search method to acquire, characterize, and engineer a robust set of I-OnuI-related LHEs which recognize a diverse set of target sequences. A system of iterative binding selection using yeast surface display was also developed to identify target sites for, and perform non-directed analysis of, previously uncharacterized enzymes. This diverse family of LHEs will serve both as a platform from which to launch short-distance designs, and a dataset to improve our understanding of protein-DNA interactions.

Table of Contents

List of Figures

List of Abbreviations

The following table describes the significance of various abbreviations and acronyms used throughout the thesis. The page on which each one is defined or first used is also given.

Chapter 1
Introduction to Genome Engineering

1.1 The impetus for genome engineering

At a very fundamental level, an organism is defined by its genome. These genomes consist of strings of nucleic acids that govern their environment directly via molecular interactions, and indirectly by coding for effector molecules such as ribonucleic acids (RNA) and proteins. These RNAs and proteins, and their interactions with the genome itself, create the structure and carry out the functions that embody that particular organism. Hence, an organism's genome gives rise to its very being. All cellular organisms including plants, mammals, yeast, bacteria, fungi, and archaea – and even many viruses – have deoxyribonucleic acid (DNA) genomes. It is no great wonder then why scientists from a broad range of biological fields covet technologies that would bestow upon them the transcendent ability to manipulate DNA in a genomic context. Such a technology would represent the ability to reprogram organisms at will.

Potential applications of DNA-editing technologies in research and biotechnology are plentiful. Researchers commonly wish to add (knock in), interrupt (knock out), or alter a gene in order to study its role in the organism. Knocking out a gene is one of the most common genetic manipulations; it is often the first step in investigating a gene's role in a signaling or metabolic pathway. In fact, many scientific publications are primarily descriptions of the phenotype of an organism after a gene of interest has been disrupted[1–3]. Once a gene's role is known, scientists in biotechnology and synthetic biology will need to knock in genes or tune existing metabolic pathways[4] in order to optimize production of their desired product. For example, once the gene responsible for insulin production was identified, it was then cloned into *Escherichia coli* to

produce a drug to treat diabetes in humans[5]. As for tuning metabolic pathways[6–9], one can clearly imagine the use and value of a yeast strain with increased ethanol production capacity.

Biologists have been able perform genetic manipulations with relative ease in model organisms such as *E. coli* and *Saccharomyces cerevisiae* for many decades, yet this same task has remained a formidable undertaking in most other organisms. A number of genetic tools, such as viruses, transposons, and self-replicating DNA elements, can be used to allow additions to the *E. coli* genome. In fact, these very tools allow cloning and other simple genetic experiments to be carried out routinely. Unfortunately, these tools cannot be used in the vast majority of other organisms, and they cannot be used to consistently target a particular locus; in this context "targeting" would generally require screening or selecting for the desired random event. Furthermore, genetic elements used in *E. coli* cannot propagate in or invade other organisms due to their reliance on their specific host's machinery (or lack thereof). This unfortunate lack of host-independent tools renders most other organisms refractory to scientific research or utilization.

While the desire to manipulate the genomes of organisms for biotechnological and research purposes is clear, the prospect of modifying the human genome is perhaps even more alluring. Gene therapy, the application of genome engineering to cure human disease, became a realistic goal for clinicians and researchers with the discovery of DNA-manipulating tools that could function in human cells. The first real hope of editing genes in the context of human genomes came in the form of a technique called "gene targeting" in the 1970's[10,11]. This technique relies on recombination of chromosomal DNA sequences with exogenous DNA that has long stretches of homology to the chromosomal target (**Figure 1**). Importantly, the

machinery required for this process is present in all eukaryotes, including humans. Furthermore, this method allows precise targeting to a particular genomic locus based on the sequence of the provided DNA. Unfortunately, while the simple act of providing homologous DNA to cells stimulates high levels of recombination in some lower eukaryotes such as yeast, it works far less efficiently in higher eukaryotes such as mice, and prohibitively infrequently in human cells[12,13].

Later in the 70's and 80's, researchers showed that human-tropic DNA- and retro-viruses had the potential to replicate alongside or integrate into the human genome[14–19]. Utilizing these viruses in humans would allow genetic manipulations akin to those allowed by *E. coli*-specific viruses: adding back a wild-type copy of a defective gene. Unfortunately this method of gene addition would not allow targeted mutations as with homologous recombination. Nonetheless, the first U.S. gene therapy trial was carried out in 1990 using this technology[20]. Adenosine deaminase severe combined immunodeficiency (ADA-SCID) is a fatal disease caused by a defect in a single gene. The wild-type coding sequence for this gene was inserted into a lentiviral delivery vector, and subsequently integrated into the patients' genomes. Though trials using this methodology met with some initial success, this tactic of non-specific integration proved dangerous. The modified viral genomes' nonspecific insertions resulted in transformation (if they inserted proximal to a proto-oncogenic gene) and genetic instability. The first patient fatality in a SCID trial occurred in 1999, followed closely by others in similar trials[21–23]. Although the viral gene therapy trials gave new hope for the functional recovery of patients with genetic defects, they underscored the danger of random integrations and consequential need for targeted genome engineering.

1.2 Targeted genome engineering

We have already established that random genetic integration (as with viruses and transposons) is too imprecise, especially in a clinical setting. Not only can important genes be disrupted, but the randomly inserted and neighboring genes can become dysregulated as well. These same factors limit the usefulness of these tools in research settings; editing genes in their natural genomic context, under their native regulatory elements would be far more informative. Conjointly, the ability to target a specific locus would enable true deletions. Genetic additions caused by viruses and transposons can only cause gene disruptions through their random insertion into and disruption of the gene of interest. Not only does disruption-based knock-out fail to create a true deletion, it also necessitates screening for the desired insertion, which is not always possible for a given organism or genetic outcome. Hence, random-addition technologies such as viruses, transposons, and episomal gene-addition-based tools are severely limited due to their tropism, their lack of precision, and their inability to perform all desired manipulations.

The ideal genome engineering tool would allow robust additions, deletions, and alterations at predetermined genomic sites in any organism. Gene targeting by homologous recombination (HR) provides a method for all three types of modifications in many organisms (**Figure 1**), but it is not robust enough in most of them. The baseline rate of homologous recombination in most organisms is low and dictates the need for additional technologies, or some improvement upon the standard gene targeting technique. Recombinases are a class of enzymes capable of catalyzing recombination in a site-specific manner, but only at a static sequence[24]. Since this static sequence specificity must be altered to obtain recombination in the gene of interest, there have been many attempts to change the specificity of recombinases. Unfortunately, no engineering attempts have met with enough success to allow efficient

recombinase use except at their native site[24]. Still, this technology remains alluring if it can be improved[25,26].

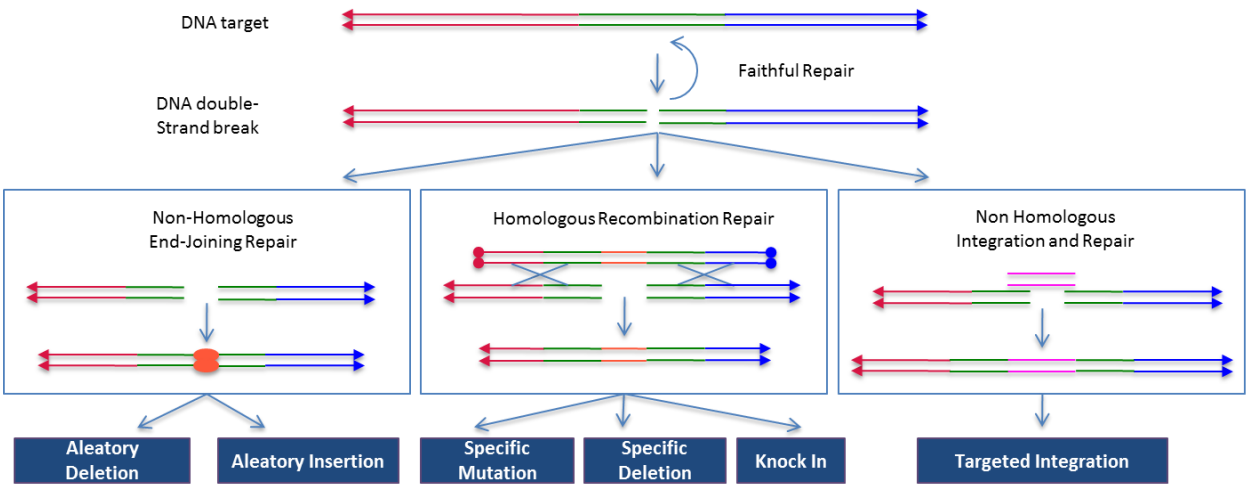


Figure 1. Site-specific genome engineering. Shown center, homologous recombination (HR) can be used to repair a DNA break, and in some organisms can be triggered spontaneously without a break. HR can lead to insertions, deletions, or mutations, depending on the nature of the sequence between the flanking homologous sequences (center). Non-homologous end joining (NHEJ) can faithfully repair a break (top), as is most often the case. Alternatively, insertions or deletions of varying length can be created before the ends are ligated (left), or ends of exogenous DNA can be ligated to create specific insertions (right).

In 1994 researchers showed that a double-strand break could stimulate increased levels of HR at the site of the break[27]. This result was particularly significant given that the experiments were carried out in mice, in which the rate of HR events was otherwise undetectable. This finding renewed enthusiasm for engineering by HR in mammalian cells by using DNA breaks to induce targeted mutagenesis given the shortcomings of alternative genome engineering methods. Furthermore, double-strand breaks can generate deletions and integrations at the site of the break as well, depending which pathway that the cell uses to repair the break[28,29] (**Figure 1**). Although the non-homologous end joining (NHEJ) pathway responsible for creating these deletions typically repairs enzymatically generated breaks faithfully by religation, deletions can occur, resulting in knock-outs[30]. This pathway is often dominant to HR, and could provide another

method for quickly and easily creating true deletions; not just disruption by insertion. Compelled by the sometimes-unwanted mutagenicity of the NHEJ pathway, researchers also discovered that single-strand site-specific breaks could also stimulate HR, while significantly reducing levels of mutagenic NHEJ[31,32]. Combining the proper method of generating a targeted break with the proper selection can yield a wide range of desirable mutations.

1.3 Generating targeted DNA breaks

The ideal engineering technology would have the following properties: first, it should be easy to program its specificity to target the desired locus. Second, it should be highly specific for its target with little if any off-target cutting. Third, it should be efficiently delivered by an existing, robust technology. For example, compact, non-repetitive coding sequences are required when delivering by lentiviral vectors (one of the highest efficiency delivery modalities for mammals). Finally, it should have the capacity to be multiplexed when multiple loci are targeted in the same experiment. Given the engineering potential of DNA-breaks, scientists have developed four site-specific endonuclease technologies in the last decade. Each technology has its own strengths and weaknesses. Unfortunately none of the nuclease technologies developed thus far fully incorporates all of these properties.

The first two technologies couple a single, modular DNA-binding protein domain to each half of a dimerizing non-specific nuclease (FokI) to create targeted breaks. The more established technology uses strings of zinc finger domains to achieve its targeting specificity (**Figure 2a**). While the engineering of zinc finger nucleases is relatively easy given the pre-determined DNA triplet recognition of each zinc finger, the technology is not without its limitations. For example,

zinc fingers are limited in their nucleotide recognition motifs, particularly outside of 5'-GNN-3', and therefore lack the ability to efficiently recognize all 64 triplets. Furthermore, binding of the domains is not completely independent, leading to unanticipated aggregate binding properties after the modular domains are assembled[33]. Consequently, the ability to correctly generate zinc finger nucleases with a desired binding affinity and specificity decreases with increased numbers of zinc finger modules[34], which can lead to off-target cutting and subsequent toxicity when using this technology[35–37]. The second technology uses a different DNA-binding protein with superior, single-base modularity derived from transcription activator-like effectors (TALEs)[38,39] (**Figure 2b**). TALEs have a more modular and more predictable binding pattern when combined, but are about three times larger per base recognized compared to zinc fingers (2x 3 kb units versus 2x 1 kb units). Furthermore, the highly repetitive nature of TALE nucleases has so far made lentiviral vector packaging impossible[40,41]. Both technologies have moderate to high levels of specificity, depending on the particular target, generating few if any off-target cleavage events. Unfortunately, neither of these nuclease platforms cannot be multiplexed effectively given their large size and heterodimerizing method of action; dimers from each of the two (or more) pairs could cross-pair, creating more possible recognition sequences and reducing overall concentrations of the proper dimers. Overall, these two technologies, especially TALE nucleases, lend the ability to quickly generate site-specific breaks, though their delivery modalities and multiplexed use may be limited.

The third nuclease platform is exceedingly easy to engineer by virtue of its use of RNA to guide the nuclease to the complementary genomic target (**Figure 2c**). This is the clustered regularly interspaced palindromic repeat and Cas nuclease system (CRISPR/Cas)[42]. Unfortunately, the specificity of guide RNAs seems to be relatively low, and targets must be

chosen very carefully in order to reduce off-target cleavage[43]. Although the size of the Cas nuclease is very large (4 kb), the ability to multiplex is high given that only short guide RNAs need to be added once the nuclease has been delivered[43–45]. Due to their homologous nature, whether or not multiple guide RNAs can be packaged efficiently into retroviral vectors remains to be seen. The new CRISPR/Cas system therefore warrants more research; while it has several advantages over other nuclease systems, its potential for off-target effects *in vivo* must first be thoroughly characterized.

The fourth class of site-specific nucleases is a family of compact, highly specific enzymes called homing endonucleases (**Figure 2d**). A subclass, LAGLIDADG homing endonucleases (LHEs), has the greatest amount of specificity, owing to their extended recognition sequences. Since an LHE's canonical recognition sequence is ~20 base pairs (bp), it appears on average only once every ~10^{12} bp (the human genome is under 10^{10} bp). In fact, LHEs' natural method of propagation relies on specific cleavage within a large genome, and subsequent induction of the HR; the very biology we wish to recapitulate. Genes coding for LHEs are mobile genetic elements. The rare cleaving DNA enzymes that they encode are in turn are responsible for catalyzing their ORF's mobility in a process known as homing. Homing relies on the generation of DNA double-strand breaks in an allele lacking the LHE gene insertion which stimulates homologous recombination using the LHE-containing allele as the template [46,47]. The LHEs must ensure that their insertion into the host gene does not disrupt its functionality; as such LHEs typically exist as introns, inteins, or in-frame fusions within their host genes. LHEs are also small (1 kb) and non-repetitive in sequence, making their delivery and multiplexing straight forward. Consequently, LHEs have drawn attention for use in site-specific genome engineering applications, particularly for organisms with large genomes [27,48,49].

Although LHEs have successfully been engineered against targets of interest[50–59], the primary drawback to this platform is the difficulty in altering the native specificity of the enzyme to new targets. Mitigating this drawback by facilitating the engineering process has been the goal of my research.

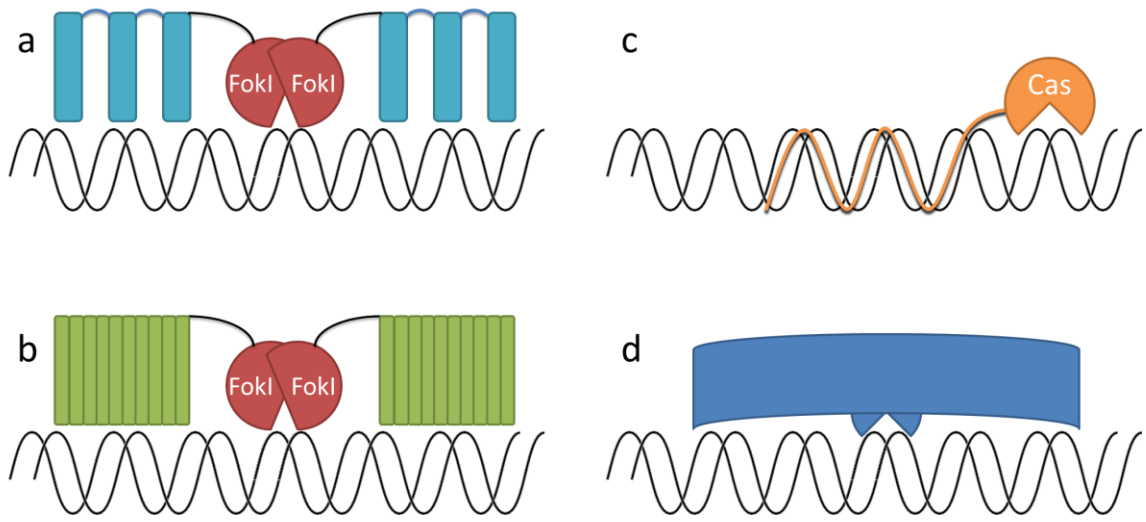


Figure 2. Site-specific nuclease platforms. (**a**) Zinc finger nucleases are chains of zinc fingers, which recognize 3 bp each. Each chain recognizes a different half-target, and is linked to a FokI nuclease domain which cleaves upon dimerization. (**b**) TALE nucleases are similar to zinc finger nucleases in their overall architecture, but each repeat domain recognizes a single base. (**c**) CRISPR RNAs bind complementary sequences and recruit Cas nuclease to cleave downstream of the binding site. (**d**) Homing endonucleases' binding and cleavage are integrated in a single protein.

1.4 Modifying homing endonucleases to target genomic loci

Given the benefits of using homing endonucleases for genome engineering, there has been a great deal of work aimed at facilitating their redesign – their primary limitation. If homing endonucleases are to enjoy widespread use, it will be important to optimize their engineering and selection. Below is a brief overview of the various approaches to redesigning proteins.

Computational and rational design have been used to successfully alter the specificity of LHEs[51,57,59–61]. Crystallographic data are often used in conjunction with computer modeling

programs to identify key residues that can be changed to generate an LHE with locally altered nucleic acid preferences[50,62]. These programs can also suggest alterations generated from previously observed binding motifs. However, many DNA-binding motifs appear to be enzyme family-specific[63], and LHE's DNA interactions have not been thoroughly characterized. As such, computer-based designs have had some success, but often fail to compile multiple adjacent mutations while accurately predicting the effect aggregate effect[50]. In practice, more progress has been made by employing "pocket designs," in which multiple possible solutions are screened; possible solutions involve random substitution of multiple residues surrounding the target base to be changed. This approach limits constraints of potentially conflicting independent mutations[52], though unforeseen changes may still be elicited[50]. Another shortcoming of purely computational design is that it relies on accurate crystallography data, which can be difficult and time-consuming to obtain for a given LHE. For these reasons, current computational redesign of LHEs, although powerful, should be used as an aid to other methods of modifying LHE specificity.

Random mutagenesis can address our limited predictive abilities. However, this method often under-samples a library given the frequent need for compensatory and epistatic mutations. Pairing random mutagenesis with rational design has augmented the success of both processes. For example, random mutagenesis was used to improve chimeric enzyme activities[57,64], and randomization targeted to regions based on structural analysis has successfully been used to generate variants of desired specificity[52,64]. Here, domains of perceived importance for a given interaction were selectively mutated; libraries of changes were iteratively generated and selected until the desired result was achieved. As such, random mutagenesis can function synergistically with rational design, typically producing the best results.

1.5 Nuclease selection and characterization

Once libraries have been generated by rational design, random mutagenesis, or a combination thereof, a screening and selection method must be chosen to recover LHEs with the desired specificity. Although there are many schemes that have been developed to select for enzymes with desired activities, I will briefly review a few that have been used on homing endonucleases in order to better put my experimental choices in context.

There are a number of simple screening platforms for homing endonucleases. Seligman, Sussman, and colleagues developed a highly sensitive bacterial screening method in which loss of Kanamycin resistance or lactose metabolism is associated with cleavage of the target embedded in a plasmid[65,66]. The method reveals toxicity, partial activities, and relative specificity of activity by replica plating the same samples on multiple media. Unfortunately, replica plating limits the technique to clonal analysis or screening of very small libraries (small enough to replica plate). Joung, Gimble, *et al*., use a bacterial two-hybrid selection system. Although this system allows the high throughput screening of millions of clones, its read-out is binding and not cleavage efficiency of the LHEs; thus, false positives are found in refined libraries[67]. A third bacterial system pioneered by Doyon and Chen enables both positive and negative high-throughput cleavage selection[68–70]. Here, the target is embedded in a plasmid harboring an inducible toxin, CcdB. Cleavage allows survival of the bacteria, and positive selection for functional nucleases. Counter-selection is achieved by embedding undesired target sites in the resistance-conferring (LHE expression) plasmid. Library sizes in the low millions can be processed by this method. In a more directly-applied selection method, Chames *et al*. have designed a yeast mating system to screen libraries that use homologous recombination rather than cleavage alone. This system relies on HR to restore an auxotrophic marker (for selection) or

reporter (for screening). This method can obtain 10^5-fold enrichment for HR-stimulating enzymes, similar to phage panning but with the ability to screen based on HR rather than on binding alone. That said, HR is significantly higher in yeast than most known organisms and results from this assay may not accurately reflect enzyme functionality in mammalian cells, for example. This method also utilizes an involved, time-consuming multistep process involving gridding and mating yeast. All of the above methods have successfully been used to screen for novel enzymes of desired specificity. Each method also typically allows for an indirect readout of nuclease activity via the number of surviving colonies, or level of activity of the reporter. However, these are only a few of the many possible methods for screening for enzymes with desirable properties.

Although the selection methods described above allow for some amount of assessment of nuclease activity, the yeast surface display (YSD) system is a complete engineering platform. In most other platforms, selections are followed by expression, purification, and biochemical analysis. With YSD, the generation of libraries, selection and characterization are integrated. Library creation can be performed using any of the methods described in section 1.4. Genetic manipulations of enzymes are simple in this system since yeast are readily transformed and recombine DNA fragments of libraries[71]. Once the enzyme libraries have been generated, they can be selected and characterized in the same system, as described next.

Flow cytometry allows rapid, detailed assessment and sorting based on stability, binding and cleavage activity of surface-displayed LHEs[64,72–74]. In the YSD system, the cloned or gene synthesized LHE open reading frame (ORF) is fused to an inducible surface displayed protein, Aga2p, which is anchored by two disulfide bonds[75] (**Figure 3a**). YSD can improve the

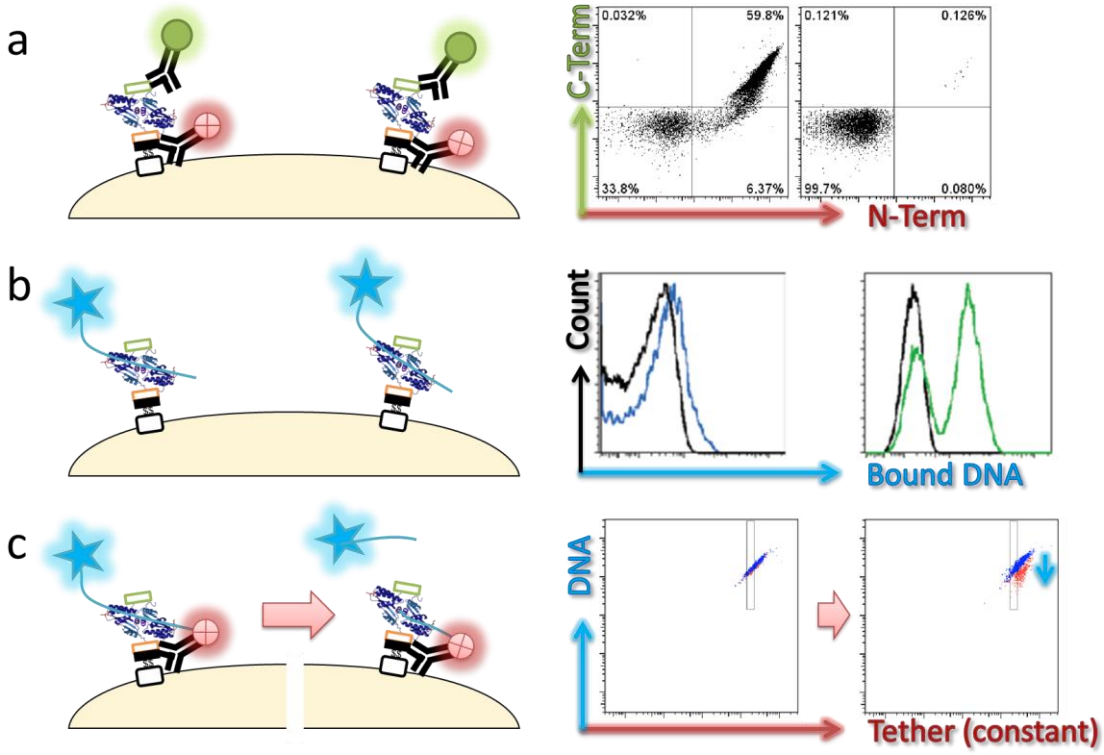


Figure 3. Yeast surface display platform. The homing endonucleases are fused to the Aga2p protein (solid black box), which is anchored to the surface-expressed Aga1p protein (white box). Nucleases are tagged at the N-terminus (orange box) and C-terminus (green box). (**a**) Immunofluorescent staining allows surface expression to be tracked by flow cytometry. A typical expression profile is shown for a well-expressed enzyme (left plot) and poorly-expressed enzyme (right plot). Full-length surface-expressed protein stains dual-positive for N- and C-terminal tags (upper right quadrant of the plots). (**b**) Binding can be measured by incubating surface-expressing yeast with fluorescently-tagged DNA (blue). By varying concentrations (left plot to right plot) and graphing the resulting change in fluorescence, binding affinity can be determined. (**c**) Cleavage activity can be measured by tethering DNA via an antibody-streptavidin bridge (black/red). Populations are split into and cleavage-inhibiting (left diagram, blue population in the plot) and cleavage-permissive conditions (right diagram, red population in the plot). Any relative decrease in tethered DNA due to cleavage is measured as a decrease in fluorescence. The left plot shows an example of no cleavage (identical fluorescence in both conditions), and the right plot shows an example of cleavage activity (drop in fluorescence of the red population relative to the blue, denoted by the blue, downward-pointing arrow).

throughput of analyzing uncharacterized LHEs in three important ways. First, transit through the ER and secretory quality control pathways helps ensure that only stably-folded LHEs are surface-displayed; dysfunctional variants that do not fold correctly are typically retained[76]. Flow cytometry using a C-terminal epitope tag can therefore reveal whether the candidate protein likely represents a functional LHE. Second, binding affinity can be assessed by incubating the surface-displaying yeast with fluorescent target DNA, washing, and looking for an increase in

fluorescence above background by flow cytometry (**Figure 3b**). Third, cleavage activity of the enzymes can be assessed in a detailed, high-throughput fashion. Fluorophore-conjugated oligonucleotides containing the putative target site are tethered to the yeast, and cleavage-associated loss of fluorescence is then quantified[73] (**Figure 3c**). However, this method of cleavage selection may negatively influence the results since some of the target-binding requirement of the enzyme may be artificially supplied by the tethering[77]. Overall, multiple parameters and iterations can be used to dynamically screen libraries in the tens of millions.

Modifying the above screening methods, the yeast surface display platform also enables high throughput characterization of LHEs. First, the same assays and parameters that are used for sorting libraries (e.g. expression, binding, and cleavage) can be used to characterize many individual enzymes in a parallel fashion. Since flow cytometry enables high-throughput analysis, the above characterizations can be extended to interrogate the LHE-nucleotide interaction at each base by assaying targets with base substitutions at each position. This method of elucidating specificity profiles is rapid because it does not rely on target cloning or sequential steps, as other methods do[78]. Second, crude protein can be isolated and used for alternative biochemical analysis by reducing the disulfide linkage that tethers the LHE to the yeast[73,75]. Overall, yeast surface display represents an integrated, robust platform for library generation, screening and characterization.

In conclusion, genome engineering is a powerful tool that can be used clinically, or for research or biotechnological purposes. One of the most promising methods for genome engineering relies on site-specific nucleases to generate breaks in DNA that can then be used to interrupt or replace the gene at the site of the break. Though there are many highly specific

nucleases, LAGLIDADG homing endonucleases have a number of characteristics that make them an ideal choice. However, one significant drawback to LHEs is the difficulty involved in re-directing their target sequence specificity. Yeast surface display is a powerful engineering platform that allows the creation of engineered libraries, and the selection and characterization of active enzymes. My work, described in the following chapters, aims to ameliorate some of the troubles of LHE engineering. With the engineering roadblock removed, I hope to bolster precision genome engineering, and help lay the foundation for further scientific advances that depend on it.

Chapter 2
Novel Homing Endonucleases Aid Engineering

2.1 Introduction

An important limitation to widespread application of LHEs in genome engineering is the requirement to modify a native LHE ("scaffold") to create variants that cleave at specific target sites. Although computational design methods and selection protocols for this purpose are now quite advanced[51,57,59–61], it remains challenging to consistently produce variants with high levels of *in vivo* activity. For example, multiple changes to an LHE often interact to cause unwanted effects, rendering it less active, or no longer specific[51]. At the outset of my thesis, a major constraint on engineering LHEs was the very limited number of characterized, native LHE scaffolds: I-SceI, I-CreI, I-DmoI, and I-AniI[77,79–81]. Furthermore, only I-AniI had been engineered to modify genomic targets[82,83]. I aimed to address this limitation in two ways. First, I hypothesized that because members of this small group were not identified based on biotechnologically useful properties, homologous proteins might represent a source of alternative but related scaffolds with more useful attributes (e.g. catalytic activity, ability to stimulate HR, etc.). Second, I aimed to diversify the set of target sequences that the base scaffolds recognized. By discovering new LHEs that cumulatively recognized a more diverse set of sequences, one could target a given locus with fewer changes to the parent scaffold's specificity. Data obtained while characterizing the new enzymes' protein-DNA interactions could also be used to train computational algorithms, further benefiting engineering effort.

I first addressed the question of whether relatives of already-characterized LHEs might have desirable biochemical properties. I searched public sequence databases to identify open

reading frames encoding proteins homologous to I-AniI, and surveyed the properties of a subset of these proteins. I-AniI was chosen because it was the most thoroughly characterized LHE available. Up to this point, yeast surface display had only been used to characterize individual proteins. However, YSD allows samples to be run in a parallel plate format using flow cytometry; this parallel workflow would be a useful when analyzing multiple enzymes. The yeast *in vitro* assays were used to detect expression, binding, and cleavage of fluorescently-labeled oligonucleotides in high throughput[73].

Next, a similar *modus operandi* was used to interrogate another sub-family of LHEs. The I-OnuI subfamily appeared be closely related at the protein level, but very divergent in the target sequences they recognized. Such a family would be ideal for diversifying our accessible target sequences.

**Note: Many of the figures and some of the text in this chapter (particularly in the methods section), are derivatives of or excerpts from my previously published work*[52,74].

2.2 Identifying new homing endonucleases and their targets

There are many putative homologs annotated in public databases, despite the lack of well-characterized LHEs. My first goal was to identify nucleases related to I-AniI. I identified multiple putative LHEs of varying similarity to I-AniI by using NCBI's tblastn function. To increase the chances of selecting functional LHEs, I constrained the search to homologs with conserved catalytic magnesium (Mg^{++})-coordinating residues within the LAGLIDADG motif. I chose six homologs, each from mitochondrial genomes of different fungi[84]. The homolog ORFs were trimmed to match the homologous functional sequence of I-AniI , codon optimized for

yeast expression, synthesized, and cloned into the yeast surface display vectors (see **Figure 5a** for the trimming schematic). The protein alignments are shown in **Figure 4a**, and their overall similarities are shown in panels **b** and **c**.

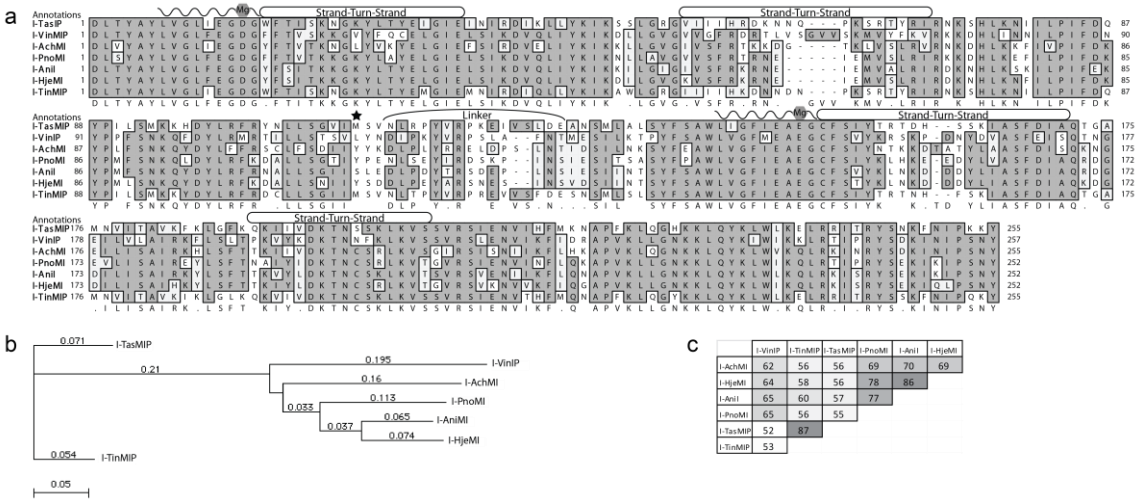
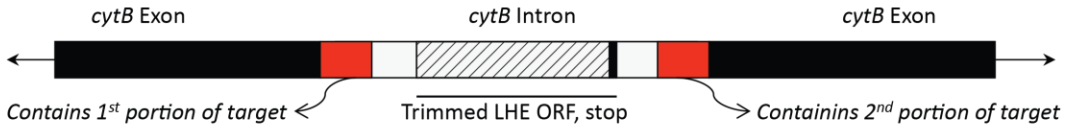


Figure 4. I-AniI homolog sequences. (**a**) The alignment of the I-AniI homologs with the residues shaded by chemical similarity. LAGLIDADG motifs are marked by waved lines. Conserved Mg^{++}-coordinating residues and DNA-contact rich strand-turn-strand regions are also annotated. The homologous serine 111, a residue important for increased catalytic activity in I-AniI, is starred. The map was generated by MacVector using Gonnet-weighted pairwise and multiple sequence alignments with residue-specific and hydrophilic penalties. Residue numbering was matched to I-AniI, based on the first LAGLIDADG motif. (**b**) The guide tree's distances are shown as uncorrected "p" values and gaps are distributed proportionally. The tree was generated by MacVector using the same parameters as the alignment. (**c**) Sequence identity comparison table for the I-AniI subfamily; higher identities are shaded darker. The identity table was generated after pairwise alignment of the protein sequences using Invitrogen's ClustalW-based AlignX program.

In order to test the functionality of the putative nucleases, I first had to predict their respective target sites. Since LHEs propagate by cleaving a site within their host gene, it is often possible to discern the original target site by carefully inspecting the host gene near the point of LHE gene insertion[85,86]. Since LHEs tend to avoid disrupting the host gene's sequence, the target is often found non-mutated, split on either side of the LHE insertion (**Figure 5a**). Since I-AniI propagates as an intron, the target site is split on each side of the intron/exon borders. These flanking sequences in the homologous LHEs differed from I-AniI, suggesting slightly altered cleavage specificities. **Figure 5b** shows the putative target sequences of the six homologs as

determined by comparison of the sequence flanking the LHE insertion where the I-AniI target sequence is found.

a



b

I-TasMIP	T	G	G	G	G	A	G	G	T	T	T	C	T	C	A	G	T	A	A	A
I-VinIP	T	G	G	G	G	A	G	G	C	T	T	T	T	C	A	G	T	G	A	A
I-AchMI	T	G	G	G	G	A	G	G	T	T	T	T	T	C	A	G	T	A	T	C
I-PnoMI	T	G	A	G	G	T	G	G	T	T	T	C	T	C	T	G	T	T	A	A
I-AniI	T	G	A	G	G	A	G	G	T	T	T	C	T	C	T	G	T	A	A	A
I-HjeMI	T	G	A	G	G	A	G	G	T	T	T	C	T	C	T	G	T	T	A	A
I-TinMIP	T	G	G	G	G	A	G	G	T	T	T	C	T	C	A	G	T	A	A	A
Pos.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10

Figure 5. Predicted I-AniI homolog targets. (**a**) A schematic of the original host gene (black) with intron insertion (white) from which the LHE ORF sequences were taken, and the exon/intron junctions used to predict target sequences (red). (**b**) The predicted targets for each homolog, derived by comparing flanking intron/exon regions for each intronic LHE with those from I-AniI; differences from I-AniI's target are shaded.

2.3 I-AniI homologs are active

Once the homologs sequences had been chosen and cloned into the yeast surface display vector, the first step was to assess the stability of the putative enzymes. As described in section 1.4, thermal stability associated with surface-expressed proteins can be measured in high throughput by flow cytometry by staining for the tagged protein. I assessed relative expression levels staining N-terminal hemagglutinin (HA) and C-terminal myc epitope tags (**Figure 6a**). αMyc antibodies were labeled with the fluorochrome, fluorescin isothiocyanate (FITC); αHA antibodies were conjugated to biotin (bio) which allowed counter-staining with streptavidin-phycoerythrin (SAV-PE). We detected expression of full-length protein for I-AchMI, I-HjeMI, and I-PnoMI, particularly the latter two, as determined by the level of C-terminal epitope tag expression (**Figure 6b**). I-TasMIP, I-TinMIP and I-VinIP showed minimal full-length protein

(dual-positive), indicating reduced thermostability and/or poor folding; presumably they were insufficiently stable at the 30 °C induction temperature.

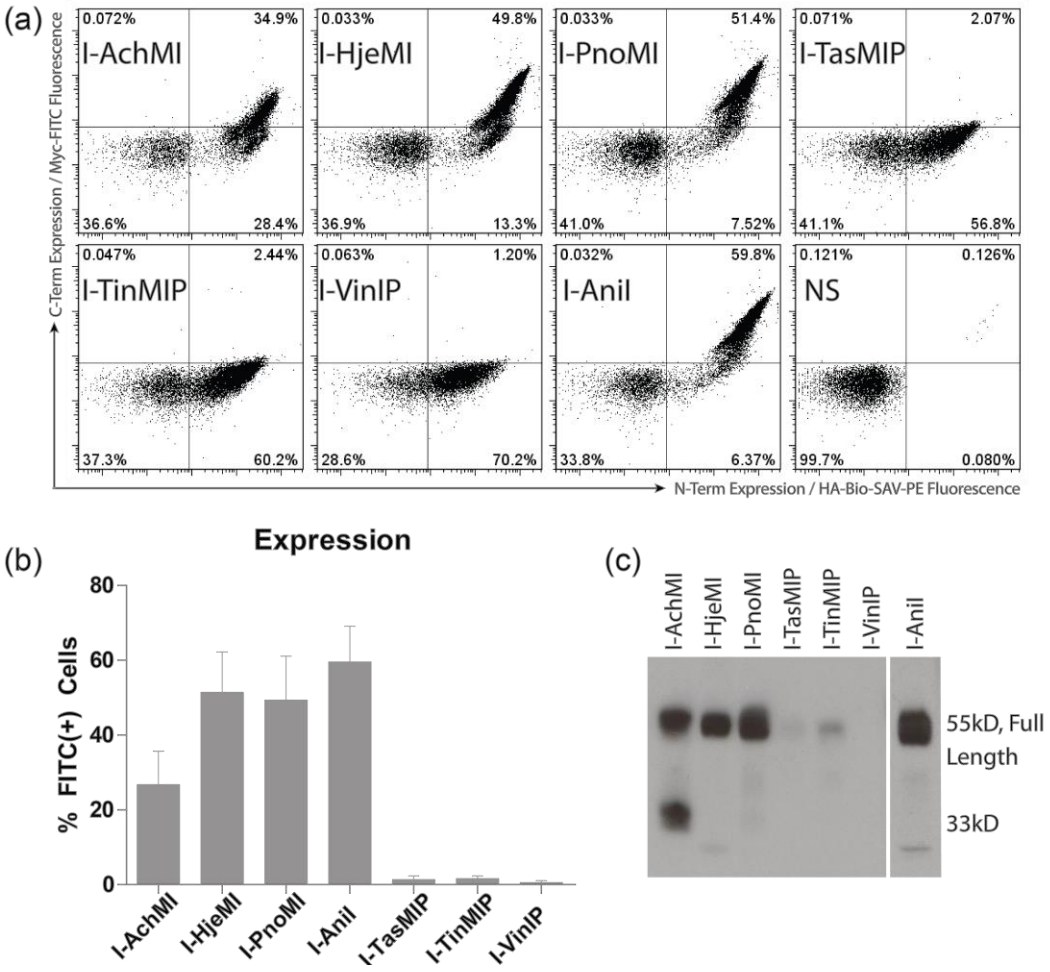


Figure 6. I-AniI homolog expression and stability. (**a**) Expression of full-length protein, determined by flow cytometry in a representative experiment. Fluorescence intensity resulting from staining against the C-terminal myc tag (*Y*-axis) and N-terminal HA tag (*X*-axis) allows quantitative assessment of surface expression. (**b**) Percentage of expressing (dual-positive) cells from (a), is summarized for five replicates (three for I-TasMIP, I-TinMIP and I-VinIP) with standard deviations. (**c**) A western blot probing the N-terminal epitope tag shows full length and truncated proteins.

To validate the flow cytometry data, I collected surface-expressed protein by reducing the disulfide linkage anchoring the nucleases to the yeast surface. As expected, poor surface expression correlated with the accumulation of heterogeneously truncated proteins containing only the N-terminal tag (**Figure 6c**); this correlation is consistent with previous observations of

surface-expressed proteins of low thermostability[87–90]. Notably, the level of surface expression correlated with the level of amino acid sequence homology to I-AniI (**Figure 4**). I-Hje, I-PnoMI and I-AniI were primarily full length and in great abundance while much of I-AchMI was expressed as a ~33 kilo-Dalton (kDa) protein fragment. Only minimal full-length protein and primarily heterogeneously truncated I-TasMIP, I-TinMIP and I-VinIP products were expressed.

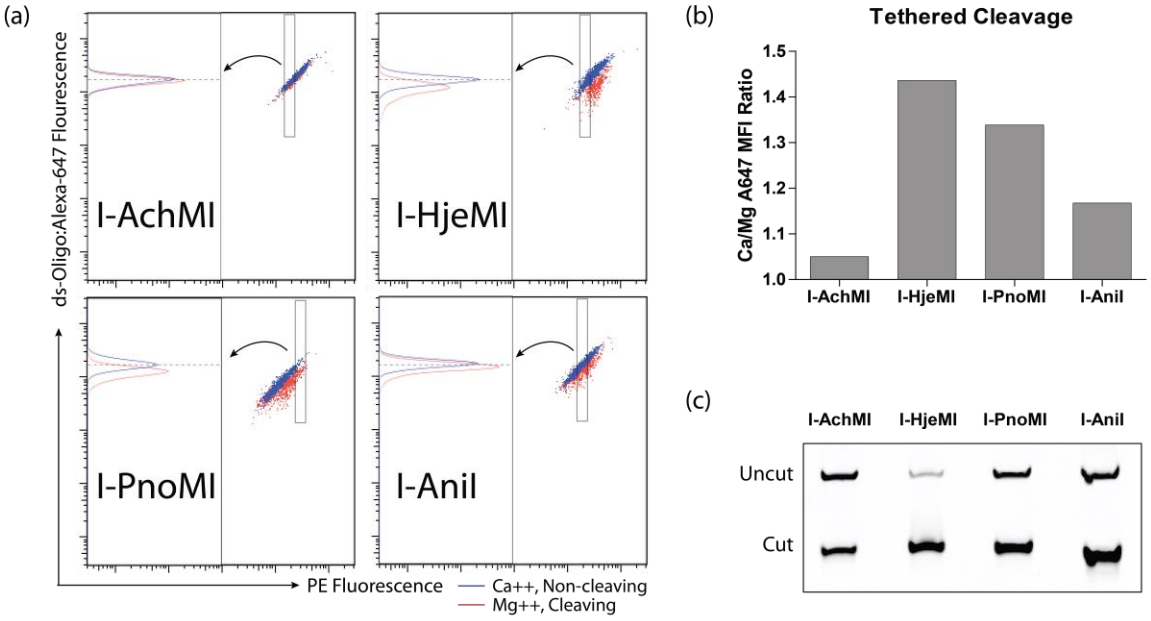


Figure 7. I-AniI cleavage activity. (**a**) Demonstration of the gating strategy used to normalize substrate for the flow cleavage assay. These displayed populations are already normalized for enzyme concentration by a uniform, narrow FITC (C-terminal epitope) gate (not shown). Equivalent amounts of tethered target across samples was selected by finding a streptavidin-PE level (rectangle) for each sample for which all DNA-Alexa647 median fluorescence intensities (MFI, dashed horizontal line) were equal in the calcium (Ca^{++}) sample (blue population). This gate was held constant for the matched pair Mg^{++} sample (red population), allowing quantification of magnesium-dependent loss of the DNA-conjugated fluorophore. The left half of the plot shows the population in the rectangular PE gate from the right plot (follow arrow). (**b**) Dividing the median Alexa647 fluorescence intensity of the calcium-containing sample (blue) by that of the magnesium-containing sample (red) yields a ratio proportional to the amount of enzymatic activity for a given LHE. (**c**) Cleavage in solution. After incubation of equal amounts of enzyme and substrate, target cleavage products were run on a polyacrylamide gel to identify the existence of specifically cut and uncut target DNA.

The three homologs with detectable surface expression (I-AchMI, I-PnoMI, and I-HjeMI) were further assayed for cleavage activity. To this end, I produced fluorescently-labeled oligonucleotide containing the predicted target site for each enzyme. Cleavage was assessed

using the previously described tethered oligonucleotide assay[73,91] depicted in **Figure 3c**. **Figure 7a** illustrates the gating strategy used to normalize the enzyme and substrate levels and quantify the amount of cleavage shown in **Figure 7b**. I-HjeMI and I-PnoMI demonstrated catalytic activity against their putative targets at levels comparable to, or slightly greater than, that of I-AniI; I-AchMI showed very low levels of activity. I subsequently validated the flow cytometry with a standard solution-based assay using protein released from the yeast surface (**Figure 7c**). First, duplication of results with cleavage in solution ensured that tethering the substrate near the enzyme did not affect the results. Second, running it on a gel verified that the cleavage event was specific by producing two products of precise sizes.

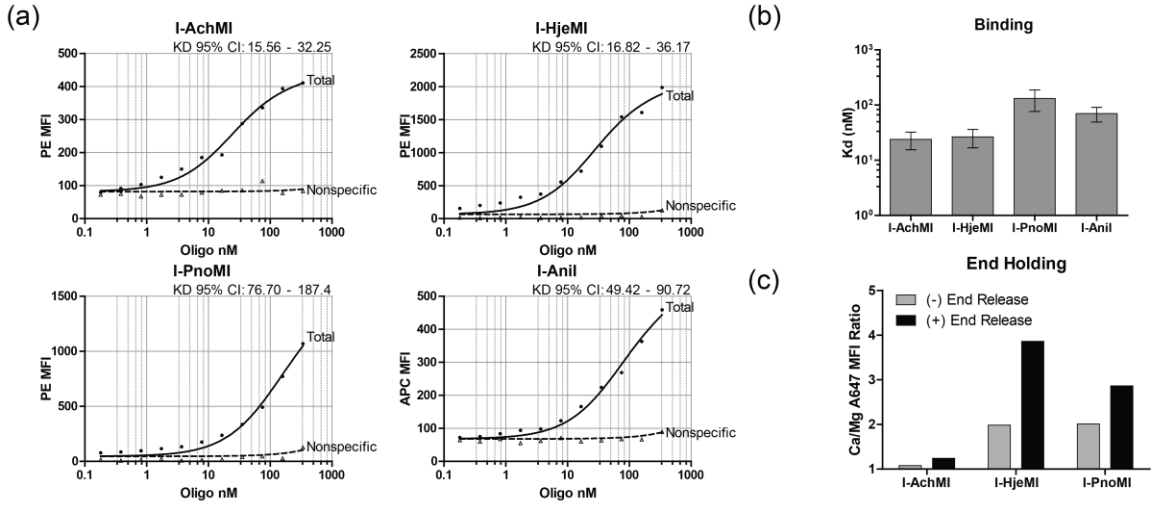


Figure 8. I-AniI homolog binding. (**a**) Binding curves. The mean fluorescence intensity (MFI) of Alexa-647 labled or streptavidin-PE counter-stained target DNA for total and non-specific DNA binding are plotted as a function of target concentration. Binding curves were fitted by non-linear regression analysis. Confidence intervals for the K_d in nanomolar (nM) target DNA. (**b**) Approximate binding affinities (Kd) for each enzyme, plotted with the 95% confidence intervals of the fitted curves from (a). (**c**) End release of fluorophore-conjugated DNA. DNA substrates, labeled either at the (+) or (-) end of the target were tethered and cleaved as in the standard flow cleavage assay. The lower Ca^{++}/Mg^{++} ratio for the (-) end-conjugated DNA in the I-AniI homologs is consistent with (-) end-holding.

Finally, I assayed each enzyme's binding affinity. Each homolog bound its predicted native target with similar affinity to I-AniI (**Figure 8a**, **b**). I also assessed their potential for "end holding," a property in which the LHE binds one DNA half-site (and retains it after cleavage)

with particularly high affinity as compared to the opposing half-site. This behavior is particularly notable for I-AniI, and has been exploited for computational design purposes [77]. Like I-AniI, both I-HjeMI and I-PnoMI were found to 'end-hold' the minus (or left) half of their DNA substrates (**Figure 8c**). This asymmetric pattern suggests that these homologs use a similar nucleotide discrimination mechanism as I-AniI[77], consistent with the high conservation of amino acid identity in the protein/DNA interface among the three enzymes in the beta sheets regions of the strand-turn-strand domains[91] (**Figure 4a**).

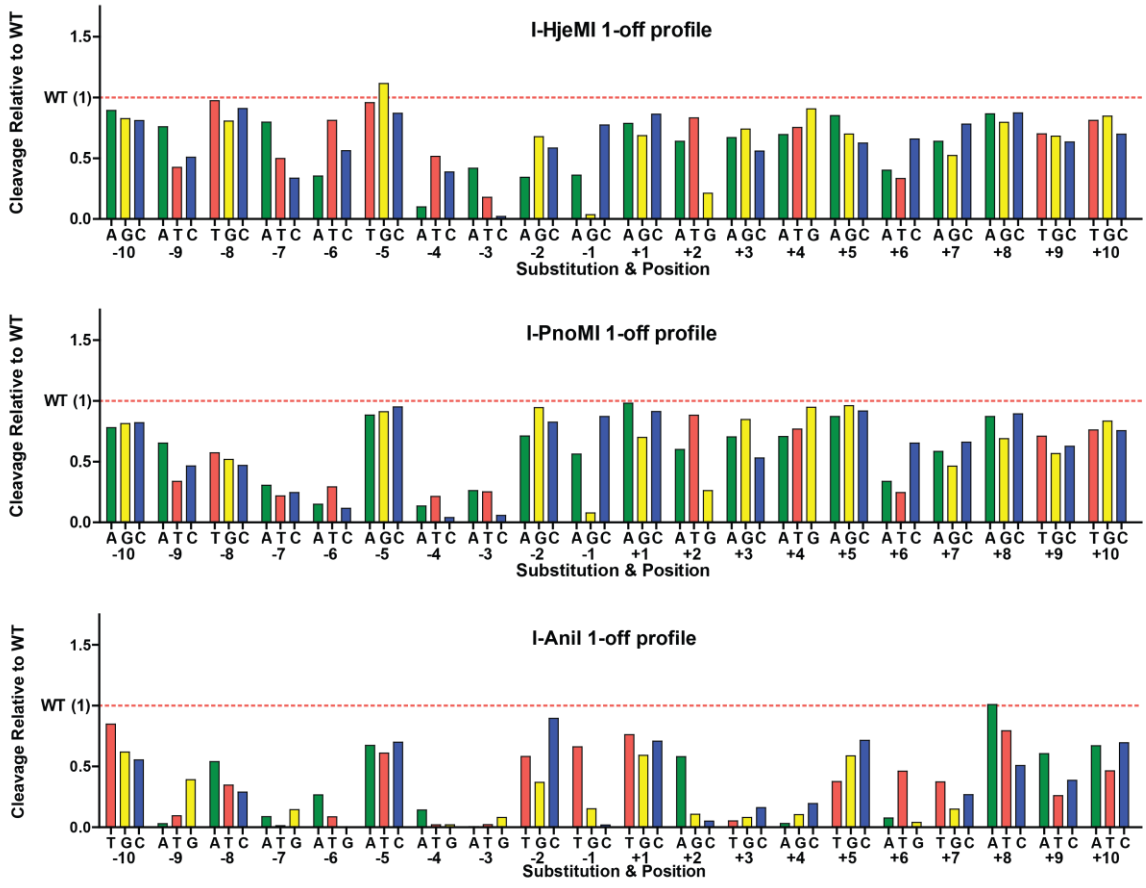


Figure 9. I-AniI homolog specificity profiles. The impact of each possible single-base pair substitution is shown relative to wild-type cleavage efficiency (red dashed line, wild-type base noted above). Values at or close to zero denote minimal tolerance of the mismatch and therefore minimal cleavage; values above 1 indicate a target is cleaved more efficiently than the predicted target.

To compare the biochemical properties of these enzymes in more detail, I generated "1-off" cleavage specificity profiles. Wild-type I-AniI and each of the two highly active enzymes, I-HjeMI and I-PnoMI, were characterized using the yeast tethered cleavage assay, with a panel of DNA substrates, each harboring a single bp mismatch relative to the LHE's physiological target is tested (**Figure 9**). This assessment revealed that, as expected, I-HjeMI and I-PnoMI exhibit I-AniI-like profiles with localized variances in positions where their predicted targets sites differ from that of I-AniI. For example, I-HjeMI demonstrated elevated specificity at position -2 compared to the other two enzymes, but reduced specificity at -8, and to a lesser extent, -7 and -6, while I-PnoMI preferred a "T" at -5, one of the two differences in its cognate target from I-AniI. Some small idiosyncratic differences were also observed, such as I-HjeMI preferring a "G" at the -5 position, despite the fact that its predicted native target site has an "A". Overall, the regions of high and low specificity are conserved across this sub-family of enzymes.

2.4 Homolog structural differences

Based on I-HjeMI's excellent *in vitro* and *in vivo* functional properties (the latter will be described in the next chapter), we were curious whether it might possess structural differences from I-AniI that could be correlated with its performance characteristics. Thus, in collaboration with the Stoddard lab, I-HjeMI was expressed in bacteria, purified it to homogeneity, and placed it into crystallization trials using a spectrum of standard conditions. In striking contrast to I-AniI, which we have found to be prone to chronic aggregation that required multiple solubilizing mutations to ameliorate, I-HjeMI was easily produced in large quantities, and remained soluble, even at high concentrations of the purified protein. The structure of the resulting complex of I-HjeI bound to its DNA target was determined by Stoddard *et al.* at 3.0 Å resolution. Although

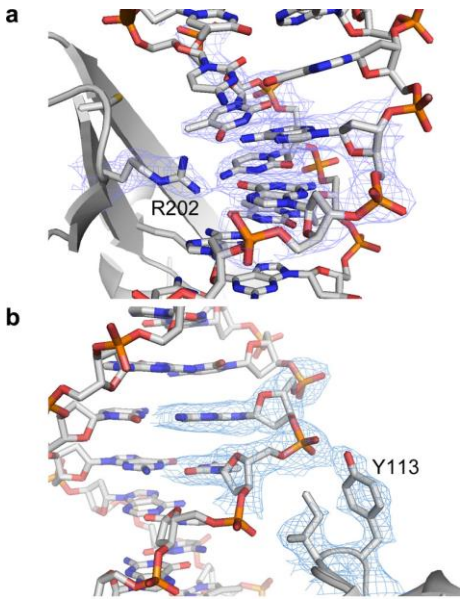


Figure 10. I-HjeMI model and electron density map. There is high quality density (blue mesh) around (**a**) R202 (K200 in Ani) and (**b**) Y113 (Y111 in Ani).

the crystal structure was not flawless[74], the high sequence identity (85%) of I-HjeMI to I-AniI (which had previously been solved and refined in multiple independent space groups to high resolution[91–93]) and the high resolution electron density for the well-ordered complex of I-HjeMI and its DNA target nevertheless allowed us to generate an unambiguous comparison of the structures of the two homologous homing endonucleases (**Figure 10**).

As expected, I-HjeMI displays a very similar overall structure to the structure of I-AniI (**Figure 11**), except for the few final residues of its C-terminus and a short region of extended peptide sequence (spanning residues 123 to 129 in I-HjeMI) that links the N-terminal and the C-terminal domains; these linker regions are typically very diverse[64]. The regions of folded secondary structure across the two enzymes, particularly the two central α-helices that contain the 'LAGLIDADG' sequence motifs, are closely superimposable (root-mean-squared deviation (rmsd) less than 1 Å between all α-carbons) while the overall rmsd for all α-carbons across the superimposed proteins is approximately 1.6 Å. The overall bend angles of the DNA and the geometry of individual base pairs (i.e. propeller twist, roll etc.) in the I-HjeMI and I-AniI complexes are also very similar.

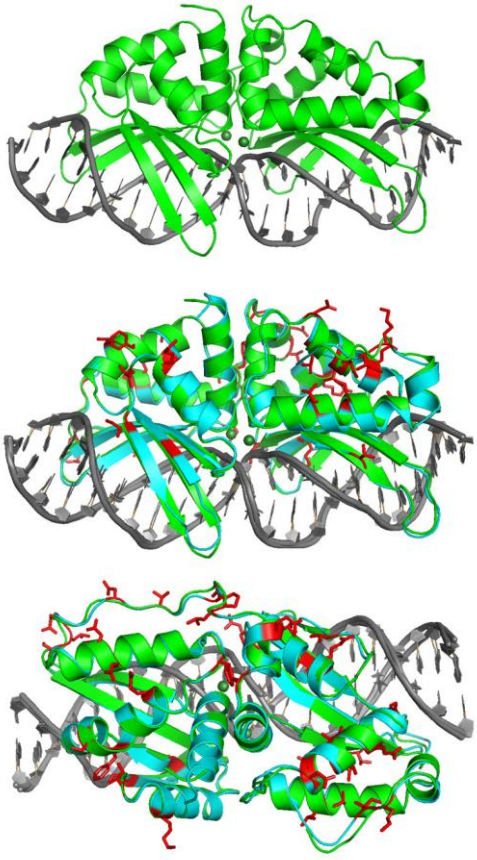


Figure 11. Structure of I-HjeMI. The solved structure of I-HjeMI (green) is shown bound to its target DNA (gray). This structure has been aligned to that of I-AniI (cyan) with differences highlighted red.

Of the 37 amino acid substitutions that distinguish I-HjeMI from I-AniI, eleven are in the N-terminal folded domain (residues 1 to 110), eighteen are in the C-terminal domain (residues 126 to 254), and eight are in the linker that connects the two (residues 111 to 125). Of those substitutions, none are located in the LAGLIDADG helices and very few are buried in the hydrophobic core (the exception being I212, I213, L215 and L235 in the core of the I-AniI C-terminal domain, which are V212, V213, I216 and I235 in I-HjeMI). The remaining amino acid differences involve residue positions that are partially or fully surface accessible. Four substitutions appear to involve residues that are involved in DNA contacts: I55, S111, R172 and K200 in I-AniI are K55, Y111, K172 and R200 in I-HjeMI. Of these substitutions, two result in additional nonspecific contacts to the DNA backbone (I55K and S111Y), one appears to have little effect on the structural mechanism of DNA recognition (R172K), and one involves a side chain that appears to contact nucleotide bases in the DNA target site (K200R, where the arginine in I-HjeMI is located within hydrogen-bonding distance of -4A and -5G on one strand of the DNA target). The substitution of a tyrosine for serine at residue 111 (S111Y) results in a nonspecific interaction with the DNA backbone outside of the 22 base pair target site. A corresponding mutation to in I-AniI was previously described during a selection experiment for improved cleavage of the wild-type target site[93]. It

therefore seems possible that I-HjeMI is a variant of I-AniI that was naturally selected for higher activity.

2.5 I-OnuI: An exciting subfamily of homing endonucleases

Shortly after searching for I-AniI homologs with enhanced biochemical properties against closely-related target sites, I began characterizing homologs that recognized a more diverse set of target sequences. I-OnuI was a recently described LHE that recognized a target site vastly different from any known enzyme[86]. Initial stages of characterization (by the same methods used above) met with overwhelming success. I-OnuI was found to cleave at comparable or better

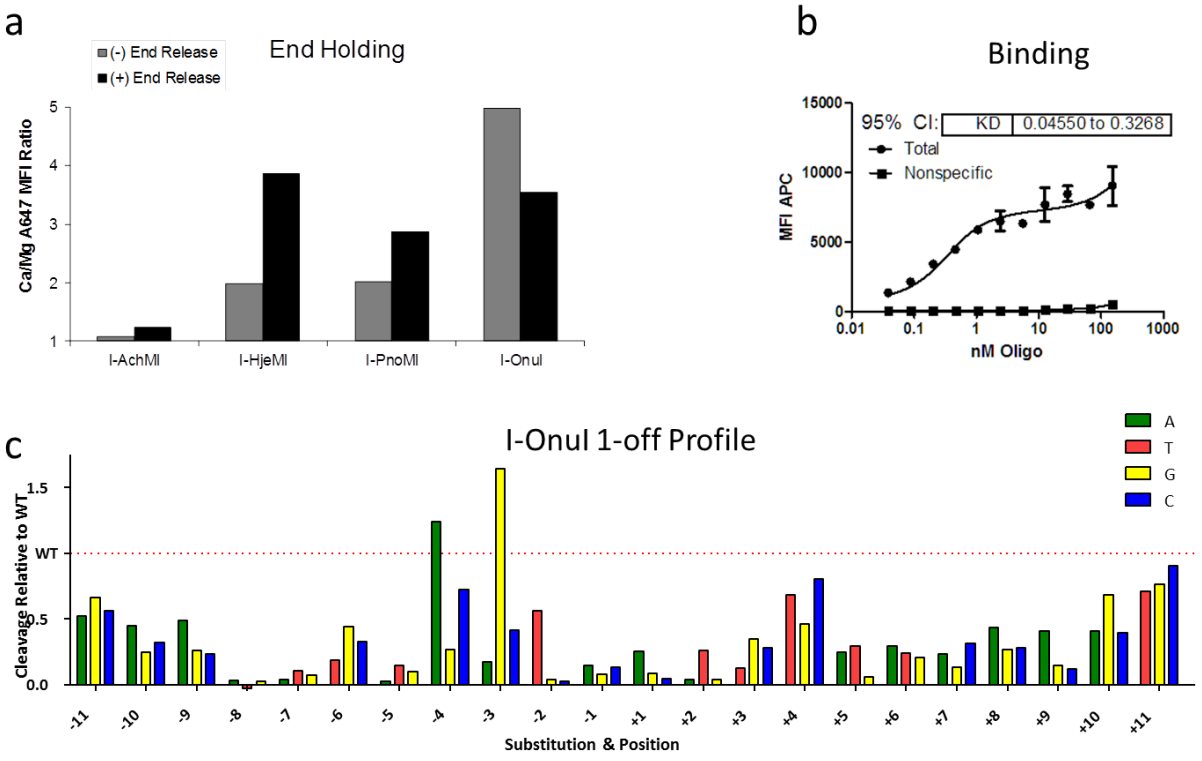


Figure 12. I-OnuI biochemical properties. (**a**) Cleavage activity and end-holding preference of I-OnuI compared to the I-AniI homologs. (**b**) Binding affinity of I-OnuI. K_d is given within a 95% confidence interval in nM target concentration. (**c**) Specificity profile for I-OnuI.

levels of efficiency compared to the I-AniI subfamily (**Figure 12a**). Furthermore, the end-holding properties were inverted compared to I-AniI, lending further diversity (**Figure 12a**). I-OnuI binds its target site with an affinity 10-100 times greater than that of I-AniI's best engineered variant, yet still retains high levels of specificity (**Figure 12b**, **c**). The tight binding and high levels of cleavage activity and specificity are all important properties for a starting scaffold since loss of these properties is a common "side effect" of engineering.

The spectacular performance of I-OnuI led our lab and the Edgell lab to analyze available genome sequence databases for homologs. Over 200 such single-chain LHEs were identified by structure-based alignments and subjected to a phylogenetic analysis using PhyML[94] that illustrated the evolutionary diversity of monomeric LHE scaffolds. In the subset of LHEs that included the previously characterized enzymes I-OnuI, I-LtrI, and I-LtrII[95], many additional putative endonucleases reside within a wide variety of different host genes (**Figure 13a-c**). Insertion at different sites suggests that members of the I-OnuI subfamily can target highly diverse DNA sequences. To test this hypothesis, and to determine how often LHE genes identified solely on the basis of sequence homology encode active endonucleases, we cloned and characterized I-OnuI, I-LtrI, and 11 putative endonucleases (most of the homolog characterization was carried out by Abigail Lambert in our lab). Eight of the putative enzymes were efficiently expressed on yeast cell surface, and six of those enzymes displayed robust cleavage of predictable DNA target sites (which correspond to the LHEs' intron insertion sites within their host genes, as described previously). Cleavage activity against each predicted site was verified using yeast surface-displayed enzyme in both in vitro and flow cytometric cleavage assays. Sequence analysis of cleaved products showed that all of the homologues generated 3′, 4-base overhangs by hydrolyzing the phosphodiester bonds between the base-pair positions ±(2

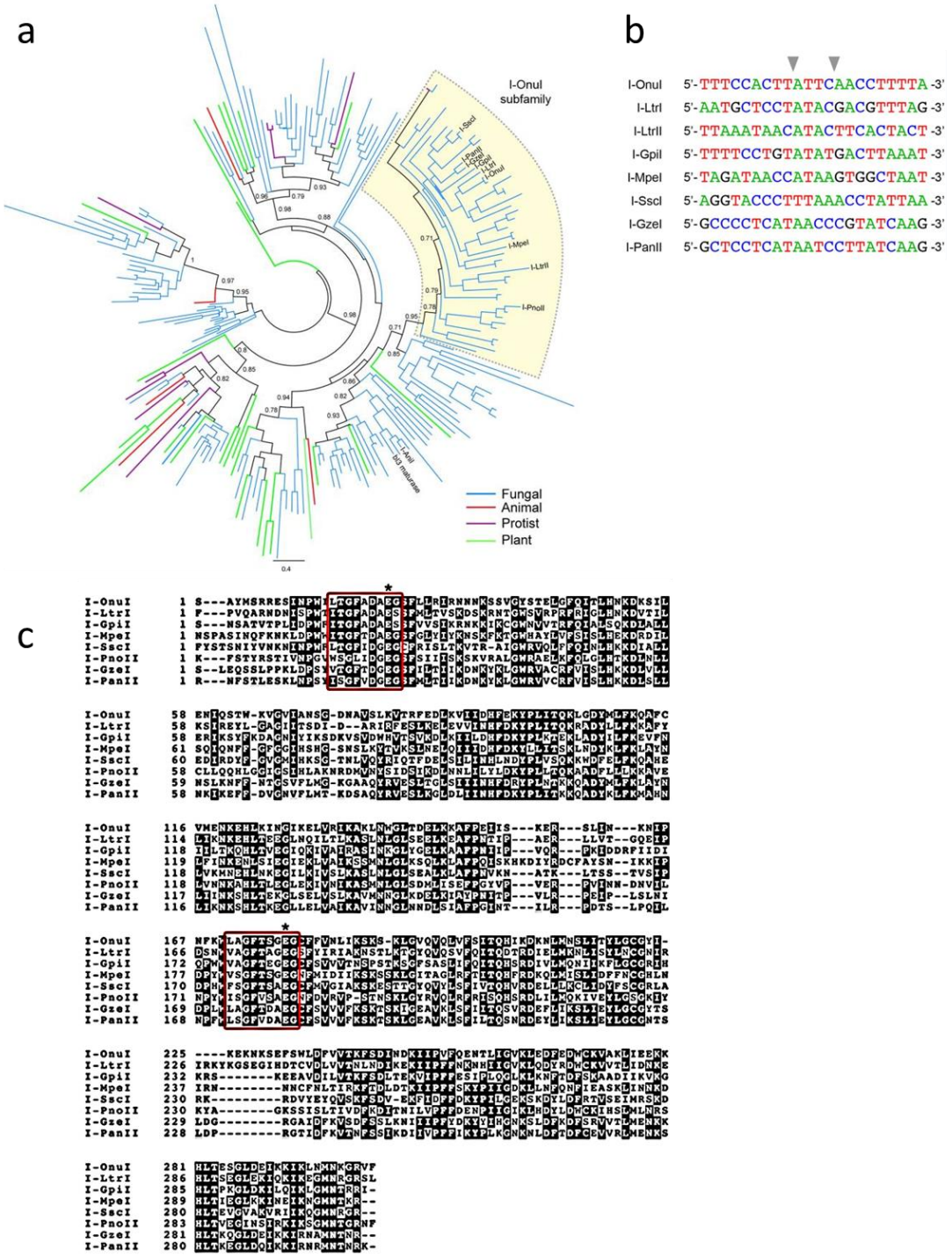


Figure 13. I-OnuI sub-family of enzymes. (**a**) Phylogenetic tree where groups of LHEs are colored by taxonomic group of the host. Black lines indicate that members of a subfamily are derived from more than one taxonomic grouping. The phylogenetic tree was generated by PhyML and approximate likelihood-ratio test values greater than 0.7 are indicated on major nodes. The relationship of the I-OnuI sub-family to other LHEs (including I-AniI) is shown. (**b**) Validated target sites for each I-OnuI homolog. Cleavage sites are indicated by gray arrows. (**c**) The sequence alignment of the I-OnuI subfamily was carried out using the MAFFT program. The conserved LAGLIDADG motifs are highlighted in red boxes. Asterisks (*) indicate the C-terminal acidic residues of each LAGLIDADG motif.

and 3). The variation in the central four residues is particularly desirable since LHE re-design in this area has been historically difficult to achieve.

To visualize the molecular contacts that facilitate recognition of different DNA sequences by otherwise very closely related proteins, the Stoddard lab solved the crystal structures of I-OnuI and I-LtrI bound to their DNA targets at 2.4-Å and 2.7-Å resolution, respectively. Although these enzymes have relatively low sequence identities compared to previously well-characterized LHEs, including I-AniI and I-CreI (< 25%), the structures revealed similar LAGLIDADG folds relative to both of those enzymes as well as to each other (**Figure 14a**). Rmsd values of I-AniI and I-CreI superimposed on I-OnuI/I-LtrI are 1.85⁄1.77 Å and 2.05⁄2.21 Å, respectively. Both proteins make contacts with approximately one-half of the nucleotide bases and backbone phosphate groups of their target sequences, via a mixture of direct and water-mediated contacts, as observed in previous LHE crystal structures[91,92,96] (**Figure 14b**). The rmsd across 284 superimposed α-carbons from the two structures is approximately 1.3 Å, and the DNA backbone conformations were similar, as well.

Other than at their active sites[52], the two protein–DNA interfaces are dissimilar. Although 12 of 22 base pairs are identical between the two target sites, only one contact between a side chain and a nucleotide contact (corresponding to glutamine 195 and the adenine ring at base-pair position +9 in I-OnuI) is observed in both structures (**Figure 14b**). These results suggest that even two closely related LHEs such as I-OnuI and I-LtrI rapidly evolve unique, divergent surfaces to recognize corresponding DNA target sites, while maintaining conserved protein folds and catalytic mechanisms. Indeed, the characterization of this sub-family of LHEs greatly diversified our targetable sequence library.

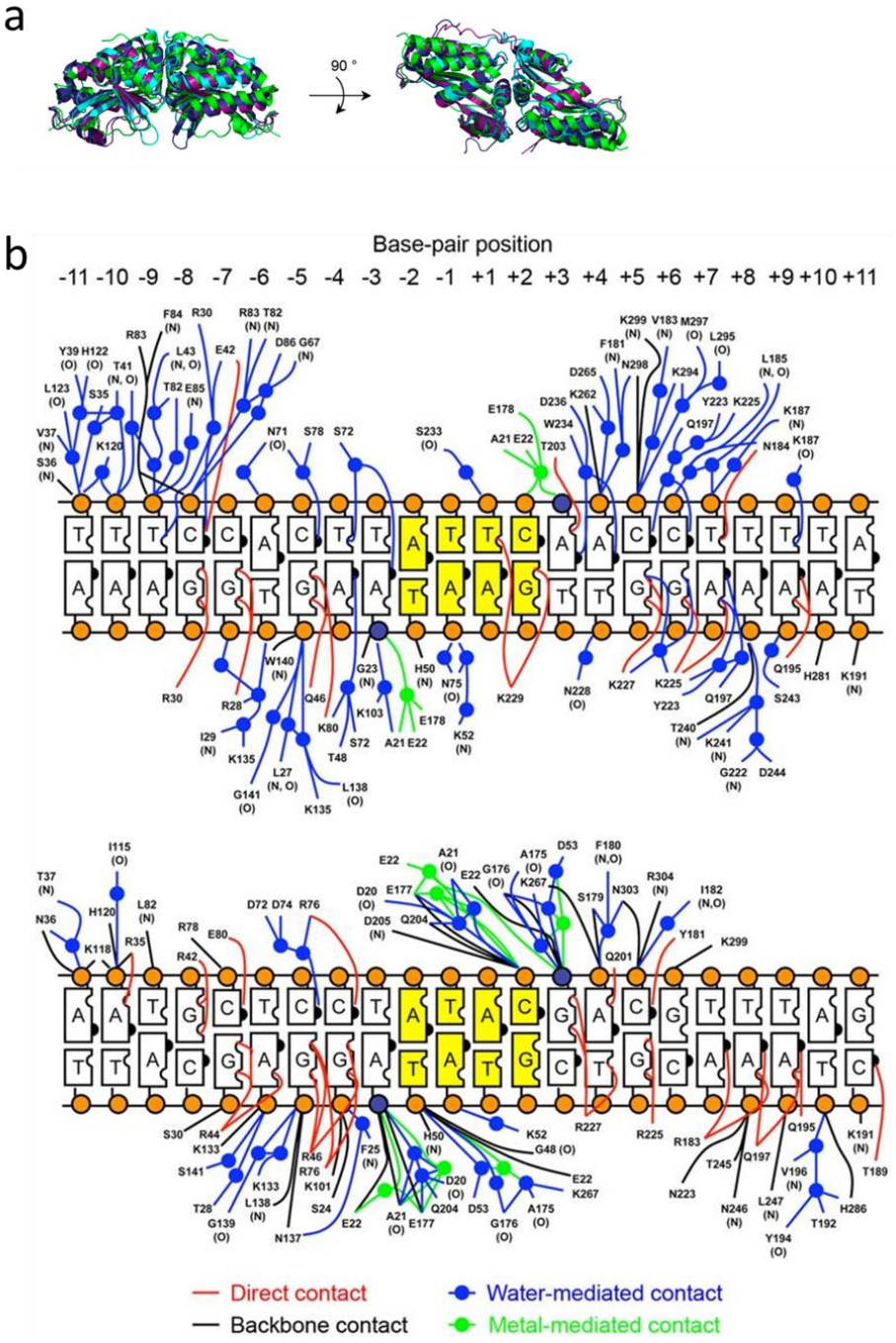


Figure 14. I-OnuI homolog structures. (**a**) Superposition of I-OnuI, I-LtrI, I-AniI, and I-CreI. I-OnuI, I-LtrI, I-AniI, and I-CreI are colored in dark blue, purple, cyan, and green, respectively. (**b**) Schematic of I-OnuI (*Upper*) and I-LtrI (*Lower*) DNA contacts. The two scissile phosphates and the other backbone phosphates are depicted as dark blue and orange spheres, respectively. The central four base pairs (positions ± 1 and ± 2) are colored in yellow. Residue numbers are adjusted to match (to I-OnuI) at the first residue of the LAGLIDADG motif.

2.6 Summary

The above data indicate that an appreciable fraction of LHE ORFs identified in public databases by sequence similarity possess potentially useful properties. I-AniI's close homologues exhibit a spectrum of biochemical properties, including increased stability, binding, and cleavage. As engineering attempts to modify an LHE's target specificity are often associated with reductions in these very properties, the availability of highly active scaffolds is a crucial component of engineering. Not only did the I-OnuI sub-family have biochemical properties superior to many of the previously characterized LHEs, they also recognized a diverse range of targets. As shown in the next chapter, the availability of high quality enzymes with alternative initial specificities facilitates rapid engineering. YSD also facilitated rapid, in-depth interrogation of LHE specificity via 1-off profiling. These profiles will 1) enable their informed use when genome engineering; 2) empower enzyme-target matching when choosing an initial LHE scaffold; and 3) train modeling software used for targeted LHE engineering[51]. Our experiments demonstrated the utility of the yeast surface display platform in rapidly expanding and thoroughly characterizing LHE scaffold diversity.

We anticipate that characterization of a wide range of monomeric LHEs will accelerate their continued development and application for genome editing, particularly when combined with protein chimerization[64] and DNA shuffling approaches, eventually allowing coverage of DNA sequence space by engineered homing endonucleases at a much increased density and success rate.

2.7 Methods

Yeast surface display expression constructs and flow cytometric expression analysis

The ability of an LHE to bind and cleave a broad panel of DNA target sequences can be readily assayed using enzyme constructs that are displayed on the surface of yeast, as described in Jarjour, *et al.*[73]. Yeast surface display of I-AniI homologs on EBY100 *S. cerevisiae* was achieved using the standard vector backbones and methods described previously [73]. Putative LHE ORF sequences were selected, corresponding to full-length I-AniI beginning 3–4 amino acids before the first LAGLIDADG helix. Corresponding DNA sequences were synthesized and cloned into the pETCON2 vector (map available on addgene.org) between N-terminal HA tag and C-terminal Myc tag coding sequences using NheI and XbaI; clones were verified by sequencing. Accession numbers for the protein sequences of I-AchMI, I-HjeMI, I-PnoMI, I-TasMIP, I-TinMIP, and I-VinIP are AAX34413, BK008014, ABU49435, BK008015, BK008016, and AAB95258, respectively. These homologs were named according to the conventions put forth by Roberts *et al* [97]; notably, a "P" suffix denotes a homolog of unverified enzymatic functionality. The additional suffix M was added to avoid redundancy in nomenclature relative to previously identified restriction or homing endonucleases, and to also denote that the host genome that harbored the LHE gene was mitochondrial.

To induce surface expression, strains harboring these vectors were grown in media containing 2% raffinose + 0.1% glucose at 30 °C for 1 day before induction in 2% galactose for 2–3 hours at 30 °C and 18–26 hours at 20 °C. To measure expression levels, 10^6 cells were washed in yeast staining buffer (YSB): 180 mM KCl, 10 mM NaCl, 0.2% BSA, 0.1% galactose, and 10 mM HEPES, pH 7.5. Cells were then stained with a 1:100 dilution of ICL Labs' αMyc-

FITC antibody and a 1:250 dilution of biotinylated αHA (Covance) antibody in YSB for 30 minutes at 4 °C. Cells were washed and counterstained with streptavidin-PE (BD Biosciences) in YSB for 15 minutes at 25 °C, washed again, and run on a BD LSRII™ cytometer (BD Biosciences). The output was analyzed using FloJo software (Tree Star) for the percent FITC-positive cells compared to an unstained population.

Immunoprecipitation and Western blot of surface-released protein

LHEs were liberated from the yeast surface by reducing the disulfide bond anchoring the Aga2P-LHE fusion to the surface expressed Aga1P protein (**Figure 3**). 250 million LHE-expressing yeast cells (induced as above) were harvested, washed twice in 1x phosphate buffered saline (PBS, Thermo Scientific), and incubated for one hour at 30 °C in 1 ml 2 mM dithiothreitol (DTT) in PBS with protease inhibitor (complete mini EDTA free, Roche). The release reaction was quenched with 10 mM iodoacetamide for 10 minutes at 25 °C to allow subsequent immunoprecipitation. The LHE-containing supernatant was incubated with 1:100 monoclonal rabbit αHA antibody (C29F4, Cell Signaling) for one hour at 4 °C and precipitated with protein A-conjugated sepharose (GE Healthcare) by incubating overnight at 4 °C. Samples were treated with PNGaseF (New England BioLabs) according to the manufacturer's protocol; this removes glycosyl residues and allows proper migration on a gel. Samples were prepared for loading by boiling in 1x Laemmli buffer (Bio-Rad).

Denaturing polyacrylamide gel electrophoresis and Western blot to a PVDF membrane were performed using standard protocols. The blot was stained with a 1:1000 dilution of rabbit αHA antibody (Cell Signaling), washed, and counter-stained with a 1:5000 dilution of donkey

αRabbit, horseradish peroxidase antibody (GE Healthcare) for imaging with the ECL system and Kodak Biomax light film.

Flow cytometric cleavage assay, end-holding, and specificity profiling

The catalytic activity of each LHE was measured by tethering Alexa647-fluorescent target DNA to the surface expressed LHE and measuring the decrease in fluorescence associated with DNA cleavage. Biotinylated fluorescent DNA is tethered to the HA epitope via an antibody-streptavidin bridge. Approximately $5x10^5$ cells were first stained with 1:250 dilution biotinylated αHA (Covance) and 1:100 fluorescin isothiocyanate (FITC)-conjugated αMyc (ICL Labs) for 30 minutes at 4 °C in the YSB. Pre-conjugated streptavidin-PE:Biotin-DNA-A467 was then bound to the yeast via the HA-biotin:streptavidin-PE interaction. This secondary stain was performed in the same buffer plus 400 mM KCl to allow biotin-streptavidin conjugation while disallowing the LHE to bind the DNA directly. Cells were washed in the cleavage solution: 10mM NaCl, 113mM K-Glutamate, 0.05% BSA, and 10mM HEPES, pH 8.2. Cells were resuspended in the cleavage buffer and split into two wells per sample. The plate was centrifuged and the pair of wells was resuspended in cleavage buffer plus 2 mM either $MgCl_2$ (cleavage permissive) or $CaCl_2$ (cleavage restrictive). Fluorescence loss due to magnesium-dependent cleavage of the DNA can subsequently be measured by comparing these otherwise identical samples. After cleaving for 20 minutes at 37 °C, cells were pelleted and resuspended in cold secondary stain buffer plus 4 mM ethylenediaminetetraacetic acid (EDTA) to aid release of cleaved substrate and mitigate any end-holding effects on DNA-fluorophore release. In end-holding experiments, this final wash was omitted. End-holding was determined by an increased

loss in fluorescence when the fluorophore was conjugated to the plus half of the DNA substrate compared to when it was conjugated to the minus half during the flow cleavage assay.

Sample fluorescence was measured on a BD LSRII™ cytometer and the resulting data was analyzed using Flowjo (TreeStar). Each sample was normalized for enzyme concentration by applying an identical narrow FITC gate. Cells were then controlled for initial substrate concentration by adjusting a narrow PE gate for each non-cleaving Ca^{++} sample until the median A647 fluorescence intensity was matched for all samples. Relative cleavage efficiencies were derived for this normalized population by dividing the median DNA-A647 fluorescence value of the Mg^{++} sample (reduced fluorescence due to cleavage) by the corresponding median fluorescence value of the Ca^{++} matched pair (no cleavage). Higher Ca^{++}/ Mg^{++} ratios indicate more cleavage.

Specificity profiles were produced by determining cleavage of each of the 60 possible target sequences wherein each base at each of the twenty positions was substituted with each of the alternate three bases, as in Jarjour's original description of this assay [73]. In these analyses, all Ca^{++}/Mg^{++} ratios were normalized to the Ca^{++}/Mg^{++} ratio of the native target site.

Flow cytometric binding analysis

K_d was measured by analyzing fluorescence of yeast surface expressing enzyme after a 2 hour incubation in serially diluted dsOligo. Alexa-647-fluorescent dsOligo containing the putative target and an unrelated target were generated & purified using the same PCR method described in the cleavage assay [73]. dsOligo dilutions between 250 nM and 0.25 nM were prepared for each dsOligo in cleavage buffer supplemented with 2 mM $CaCl_2$ (to prevent

cleavage but allow binding) and incubated with 100 pM enzyme, assuming 10^5 molecules per expressing yeast surface. Yeast were pre-stained with αMyc-FITC antibody to identify expressing populations (see methods). Samples were washed in the same buffer and fluorescence data was acquired using a BD LSRII™ cytometer.

The expressing (FITC-positive) population's A647 fluorescence was quantified using Flowjo and analyzed by non-linear regression, yielding each enzyme's affinity. To calculate specific binding, Median fluorescence intensity (MFI) of the fluorophore (A647/APC, or PE if counter-stained with streptavidin-PE) was plotted versus DNA concentration (GraphPad Prism); the resulting data was fit using iterative least-squares modeling to the equations for equilibrium specific and nonspecific binding:

$$Specific = \frac{B_{max}*[L]}{[L] + Kd} \qquad Nonspecific = NS*[L] + A \qquad Total = Specific + Nonspecific$$

where [*L*] is the concentration of substrate, *Total* is the measured median A647 value for the target-containing dsOligo, and *Nonspecific* is the median A647 value for the irrelevant target-containing dsOligo. Specific binding was calculated by subtracting nonspecific binding from the total binding to determine the maximum bound ligand, B_{max}, and the dissociation constant, K_d. The constant term *A*, and the slope of nonspecific binding, *NS*, were included to adjust for median fluorescence value of unstained cells for a given cytometer's photomultiplier tube voltage setting, and nonspecific DNA interactions, respectively.

In solution cleavage assay

The cleavage assay was performed in concert with the expression assay so that the number of expressing yeast, and therefore enzyme concentration, could be normalized. One

million expressing yeast (~280 pM enzyme, assuming 10^4 molecules per yeast surface), were incubated with 50 nM Alexa-647-labeled dsOligo for 30 minutes at 37°C in cleavage buffer supplemented with 5 mM DTT, and 5 mM $MgCl_2$. Supernatants were run on a 12% non-denaturing polyacrylamide gel and the fluorescent DNA bands were quantified using an Odyssey infrared imaging system (Li-Cor Biosciences).

I-HjeI Protein expression and purification

The I-HjeMI reading frame was ligated into a commercially available pET15b expression plasmid (Novagen, Inc) that incorporates an N-terminal 6-histidine affinity purification tag and subsequent thrombin cleavage site prior to the endonuclease reading frame. One point mutation was incorporated in the I-HjeMI (corresponding L232K), based on the knowledge that a similar mutation at that position increases the solubility of the homologous I-AniI [92]. The I-AniI construct used for parallel expression experiments under the same conditions was as previously described [91]. Both I-HjeMI and I-AniI constructs were expressed in *Escherichia coli* strain BL21-CodonPlus (DE3)-RIL (Stratagene Inc.), using a previously described method for automatic induction of protein expression [98].

Harvested cells were collected by centrifugation, resuspended in 500 mM NaCl, 50 mM Tris-HCl, pH 8.0, 5% glycerol, with 0.2 mM PMSF and benzonase and lysed by sonication. After a second centrifugation step, the clarified cell lysate was filtered (45 micron pore size), and subsequently purified using a single Heparin affinity purification chromatography step (HiTrap Heparin HP, GE Healthcare Life Sciences), eluted with an increasing gradient of 0.5 to 1.0 M NaCl. The resulting protein was exchanged into thrombin cleavage buffer and the N-terminal His-tag was proteolytically removed. The homing endonuclease was then purified further by

incubating the sample with nickel-NTA agarose resin (to remove the cleaved histidine tag and any remaining fusion protein), followed by size exclusion chromatography.

I-HjeI Crystallographic analysis

The DNA oligonucleotides used for co-crystallization (5'-GCG CTG AGG AGG TTT CTC TGT TAA GCG A-3' and 5'-CGC TTA ACA GAG AAA CCT CCT CAG CGC T-3') were synthesized by Eurofins MWG Operon Inc (desalted; 50 nanomole scale syntheses). The oligonucleotides were dissolved in 10 mM Tris-EDTA buffer pH 7.8, and the complementary DNA strands were annealed, to a final concentration of the resulting DNA duplex of 1 mM, by incubation at 95°C for 5 minutes and cooling to 25°C, over a 2 hour period. Purified I-HjeMI protein described above was mixed with 1.2-fold molar excess of the DNA substrate for a final concentration of 4.5 mg/mL protein, in the presence of 1 mM $CaCl_2$, 400 mM NaCl, and 50 mM Tris-HCl. The protein-DNA drops were mixed in a 1:1 volume ratio with a reservoir solution containing 0.2 M Ammonium sulfate, 0.1 M Bis-Tris pH 5.5, and 25% Polyethelylene Glycol 3350. Crystals grew within one week and were frozen by transfer for 1 to 2 minutes in the crystallization reservoir solution supplemented with 30% sucrose (w/v) followed by direct submersion into liquid nitrogen. The space group of the crystals corresponded to $P2_12_12$; a = 181.54; b = 73.58; c = 82.0 Å. The crystals diffracted up to approximately 2.3Å resolution at the ALS beamline 5.0.2 (Lawrence Berkeley National Laboratory). Data sets were processed using the HKL2000 software package[99]. The structure of the I-HjeMI/DNA complex was solved by molecular replacement using the PDB coordinates of the wild-type (WT) I-AniI/DNA complex (PDB: 2QOJ), and was modeled using COOT (Crystallography Object-Oriented Toolkit)[100] and refined using REFMAC/CCP4i[101].

Construction of the LAGLIDADG Multiple Sequence Alignment and Phylogenetic Analysis

Putative full-length LAGLIDADG homing endonuclease sequences (pf00961, pf03161, and pf05204) were collected from the Pfam database[102]. The structure-based multiple sequence alignment was built using Cn3D (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml). The structure of I-OnuI was aligned to the structure of I-AniI complexed with DNA substrate (PDB ID code 1P8K)[91]. Collected pfam sequences were aligned to the I-OnuI and I-AniI structures, at which point the homodimeric LAGLIDADG sequences were removed based on the length of the open reading frame and the occurrence of a single LAGLIDADG motif per sequence. Elimination of the homodimeric LAGLIDADG homing endonucleases (LHEs) aided in the alignment of structurally homologous regions of single chain, monomeric LHEs. The sequence alignment generated by Cn3D was subsequently validated using a modified version of Jalview[103] that calculates the MIp/Zp covariation statistic[104] in real time while the alignment is edited. Groups of misaligned sequences were realigned to minimize local covariation, as local covariation is a unique indicator of misalignment that is independent of methods used to build the multiple sequence alignment[105]. Local covariation was also used as a guide to reject partial and erroneous sequences. To prepare the sequence alignment for phylogenetic analysis, the alignment was trimmed such that only regions of putative structural homology were included; gapped positions and structurally divergent regions were removed from the alignment. A region of high local covariation was identified where no alternative (and putatively correct) alignment was found; this region was removed, as it is likely misleading. PhyML was used to calculate the tree[94]. The approximate likelihood-ratio test values were calculated to provide statistical support for the branching order, and values greater than 0.7 are indicated at major nodes on the tree (**Figure 13a**). The gene tree was rendered using FigTree. An enhanced version of the tree, with

branches labeled by accession numbers, can be found online[52] (http://tree/bio.ed.ac.uk/software/figtree/). Partial sequences of all the LHEs identified are provided at the same location.

I-OnuI Homolog Protein Expression and Purification

The open reading frames of I-OnuI and its variants were inserted between BamHI and NotI sites of pGEX 6P-3 vector (GE Healthcare Life Sciences), and the GST fusion recombinant proteins were expressed in *Escherichia coli* strain BL21-CodonPlus (DE3)-RIL (Agilent Technologies). Protein expression was induced in LB medium supplemented with 0.2% glucose, 1 mM MgSO4, and 100 μg/mL ampicillin at 16 °C for approximately 20 hours after the culture had achieved early log growth phase (OD_{600} ¼~0.6). The harvested cells were resuspended in TDG buffer [20 mM Tris-HCl (pH 7.5), 1 mM DTT, and 5% glycerol] supplemented with 0.5 M NaCl. After adding lysozyme to 0.5 mg/mL, the cells were sonicated for 30 seconds 6 times and stirred on ice for 30 min. The clarified cell lysate was obtained by centrifugation at 25,000×g for 30 minutes at 4 °C, and nucleic acids were precipitated by adding polyethylenimine (pH 7.9) to 0.25% (vol/vol). After centrifugation at 25,000 × g for 10 minutes at 4 °C again, the supernatant was filtered through a 0.45 μm PVDF membrane and mixed with glutathione sepharose 4B beads (GE Healthcare Life Sciences). The beads were extensively washed with TDG buffer supplemented with 2 M NaCl and equilibrated with Digestion buffer [50 mM Tris-HCl (pH 7.0), 0.5 M NaCl, 1 mM DTT, and 5% glycerol]. The intact I-OnuI and subsequent variant proteins were eluted by incubation with PreScission protease (GE Healthcare Life Sciences) for 16 hours at 4 °C. The collected proteins were concentrated and stored at −80 °C until use. I-LtrI protein was expressed and purified as previously described[86], with a slight modification: The N-terminal

tag was eliminated by digestion with thrombin before the protein was loaded on a Superdex 75 column (GE Healthcare Life Sciences). The purified protein was concentrated and stored at −80 °C until use

Crystallization of I-OnuI and I-LtrI Bound to Their Cognate Target Site

To crystallize I-OnuI, the DNA oligonucleotides (5′-CTT TCC ACT TAT TCA ACC TTT TAC CC-3′ and 5′-GGT AAA AGG TTG AAT AAG TGG AAA GG-3′) were purchased from Integrated DNA Technologies (1 μmol scale, HPLC purified). The oligonucleotides were dissolved in TE buffer [10 mM Tris-HCl (pH 8.0) and 1 mM EDTA], and the complementary DNA strands were annealed by incubation at 95 °C for 10 minutes and slowly cooling to 4 °C over a 6 hour period. Two hundred μM I-OnuI protein in 50 mM Hepes-NaOH (pH 7.5), 150 mM NaCl, 20 mM MgCl2, and 5% (v/v) glycerol was mixed with a 1.2-fold molar excess of the DNA substrate. The protein–DNA drops were mixed at a 1:1 ratio (v/v) with a reservoir solution containing 100 mM sodium acetate (pH 4.6), 100 mM ammonium sulfate, and 25% (v/v) polyethylene glycol 300 and equilibrated at 22 °C. Crystals were soaked in 100 mM sodium acetate (pH 4.6), 100 mM ammonium sulfate, and 30% (v/v) polyethylene glycol 300 at 4 °C overnight and frozen by looping and submersion into liquid nitrogen. The native crystals diffracted up to approximately 2.4-Å resolution at the Advanced Light Source (ALS) beamline 8.2.1.

I-OnuI protein incorporating selenium-substituted methionine (SeMet) (120 μM) in 50 mM Hepes-NaOH (pH 7.5), 150 mM NaCl, 20 mM MnCl2, and 5% (v/v) glycerol was mixed with a 1.2-fold molar excess of the DNA substrate. The protein–DNA drops were mixed in a 1:1 volume ratio with a reservoir solution containing 200 mM magnesium acetate, 30 mM ZnCl2,

and 16% (v/v) polyethylene glycol 3350 and equilibrated at 22 °C. Crystals were soaked in 200 mM magnesium acetate, 30 mM ZnCl2, 16% (v/v) polyethylene glycol 3350, and 12% glycerol, and frozen by looping and submersion into liquid nitrogen. The SeMet crystals diffracted up to approximately 3.6-Å resolution at the ALS beamline 5.0.2. The homolog sequence dataset was processed using HKL2000 package[99]. The polyalanine model of I-AniI/DNA complex (PDB ID code 1P8K) was used as a search model for molecular replacement. Two copies of the search model were found using the SeMet data by PHASER[106] and refined using REFMAC5[107]. The anomalous difference Fourier map was used to determine an orientation of the protein coordinates relative to the DNA target site. After the R free reached 44%, the refined model was used to search the protein/DNA complex in the native data by molecular replacement. The coordinate was further refined by REFMAC5 and Crystallography & NMR System[108]. The final model was deposited in Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank with RCSB ID code 3QQY. Statistics for the crystallographic data can be found online[52].

To obtain I-LtrI-DNA cocrystals, the DNA oligonucleotides (5′-GGT CTA AAC GTC GTA TAG GAG CAT TTG G-3′ and 5′-CAA ATG CTC CTA TAC GAC GTT TAG ACC C-3′) were purchased from Integrated DNA Technologies (1 μmol scale, standard desalting purification). The oligonucleotides were dissolved in TE buffer, and the complementary DNA strands were annealed by incubation at 95 °C for 10 minutes and slowly cooling to 4 °C over a 6 hour period. One hundred μM I-LtrI protein in 50 mM Hepes-NaOH (pH 7.5), 150 mM NaCl, 5 mM MnCl2 and 5% (v/v) glycerol was mixed with a 1.5-fold molar excess of the DNA substrate. The protein–DNA drops were mixed in a 1:1 volume ratio with a reservoir solution containing 100 mM BisTris (pH 6.5), 200 mM magnesium chloride, and 20% (vol/vol) polyethylene glycol

3500 and equilibrated at 22 °C. The crystals diffracted up to approximately 2.7-Å resolution at the ALS beamline 5.0.1. The homolog sequence dataset was processed using HKL2000 package. The polyalanine model of I-OnuI/DNA complex (PDB ID code 3QQY) was used as a search model for molecular replacement. One copy of the search model was found and refined using REFMAC5. The final model was deposited in RCSB Protein Data Bank with RCSB ID code 3R7P. Statistics for the crystallographic data are available online[52].

Chapter 3
Using New Homing Endonucleases *in vivo*

3.1 Introduction

Once a set of homing endonucleases had been obtained and characterized *in vitro*, the next task was to investigate their performance *in vivo*. Both mutagenic non-homologous end joining and homologous recombination can be desirable repair pathways for a DNA break, depending on what outcome is desired (see 1.2 Targeted genome engineering and **Figure 1**). This chapter discusses the experiments conducted to assess the potential of using the LHE homologs from Chapter 2 for genome engineering. First, I characterized the *in vivo* function of each I-AniI homolog against its native target site. Second, in collaboration with Michael Certo in our lab, we sought to modify these activities to obtain higher levels of targeted gene disruption[109]. Third, in collaboration with Ryo Takeuchi in the Stoddard lab, we engineered and characterized the activity of an I-OnuI variant against an endogenous genomic target[52].

In order to assess the potential to stimulate HR and NHEJ repair pathways, I created reporters for each I-AniI homolog that could read out these two types of events by flow cytometry. This "Traffic Light" reporter (TLR) was under development by Certo *et al*. at the time, and was later published[110]. The reporter contains a pre-programmed target sequence that is integrated into the genome (**Figure 15a**). The system can then provide quantitative single cell tracking of nuclease delivery and associated DNA break repair at the genomic target (**Figure 15b**). The reporter works because HR-corrected breaks repair an interrupted green fluorescence protein (GFP) ORF. These repair events are made possible by delivering a truncated GFP repair template along with the nuclease. Alternatively, DNA breaks that are repaired in a mutagenic

fashion and shift the reading frame to +3 (a 1 in 3 chance for insertions or deletions) will place an mCherry fluorescent protein ORF in frame. Thus, cells that repair a nuclease-induced break by HR will fluoresce green, while about one third of those that use mutagenic NHEJ will fluoresce red. Cells in which breaks do not occur, or in which breaks are repaired faithfully, will not fluoresce red or green. The TLR therefore provided a way to quantify nuclease actions and cellular repair outcomes on a single-cell basis for each of the I-AniI homologs.

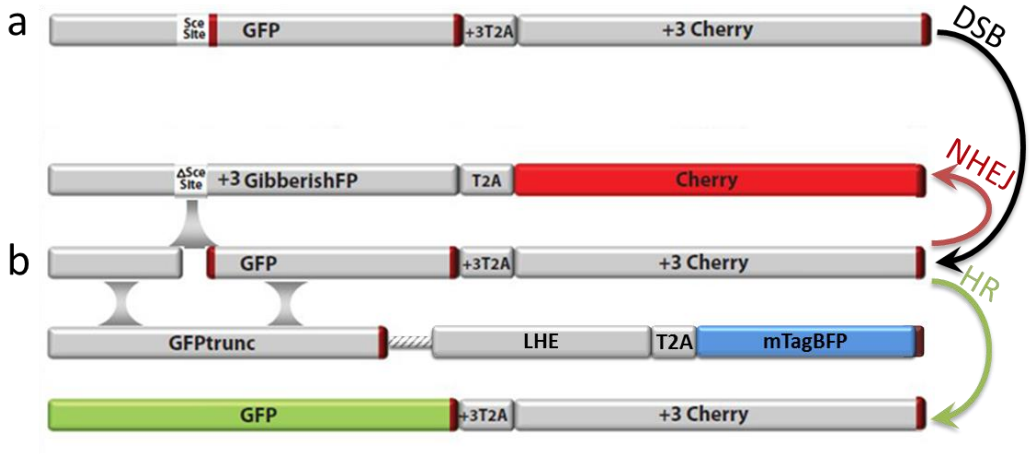


Figure 15. Traffic Light reporter schematic. (**a**) Schematic of the TLR locus harboring the I-SceI target site embedded within the +1 GFP reading frame, followed by the T2A.mCherry ORFs in the +3 reading frame. (**b**) Diagram of the TLR outcomes following double strand break repair pathway choice. A fluorescent GFP protein will be translated if a homology-direct repair (HDR) event occurs, while a fluorescent mCherry protein will be expressed if a mutagenic NHEJ event occurs that results in a +3 frameshift. The repair template also carries an LHE expression cassette whose expression is tracked by mTagBFP, making this a 2-part system.

We wanted not only to determine native levels of LHE-induced modifications, but also to increase the abundance of certain outcomes. Gene disruption is a very commonly used tool in biological experiments. In fact, site-specific endonucleases have been used to disrupt genes in many different organisms including zebrafish[82,111], drosophila[112], rats[113], and humans[114]. From a gene disruption standpoint, it is unfortunate that ends produced by nucleases can be precisely ligated by the classical NHEJ pathway[115]. This faithful repair leads to recreation of the target, and a break/repair cycle that is not resolved in a mutagenic way during the lifespan of the

nuclease[30], and is therefore unproductive from an engineering standpoint. Recent research showed that a non-processive exonuclease, Trex2, was able to prevent persistent breaks by removing bases at the break before faithful repair can be undertaken[115]. In collaboration with Michael Certo *et al*., we tested the ability of Trex2 to increase gene disruption in a panel of LHE homologs from the previous chapter.

Finally, we tested our hypothesis that "a diverse set of endonucleases facilitates engineering." Ryo Takeuchi engineered the I-OnuI homolog to target a locus in the human genome, and we assessed gene disruption at that locus. To determine the safety of the engineered variant, I generated specificity profiles (as in section 2.3 and 2.5), and we used them to search for off-target cleavage events.

3.2 I-AniI homologs are active in vivo

The three I-AniI homologs that exhibited detectable surface expression and cleavage activity (I-AchMI, I-PnoMI, and I-HjeMI) were also assayed for their potential for endogenous DNA targeting and genome engineering using the Traffic Light reporter integrated into a human cell line. The reporter system was implemented in two parts: the chromosome-embedded reporter and an endonuclease expression and repair template vector (**Figure 15a**). Nuclease expression and donor delivery was tracked by a blue fluorescent protein linked in translation via a T2A self-cleaving peptide.

I created reporters for each enzyme, each harboring the respective enzyme's target site. Human cells lines were then transduced with the reporters to create polyclonal populations with each cell having a single integrated copy of the reporter. Next, each of these cell lines were

transfected with equal amounts of a donor template plasmid carrying an expression cassette for the respective homing endonuclease, linked by a self-cleaving peptide to a monomeric blue fluorescent protein, mTagBFP (**Figure 15b**). This linkage resulted in approximately equivalent amounts of nuclease expression and repair template, as tracked by the expression of mTagBFP (**Figure 16a**). 72 hours post-transfection, I analyzed the TLR cells by flow cytometry for HR and mutagenic NHEJ events.

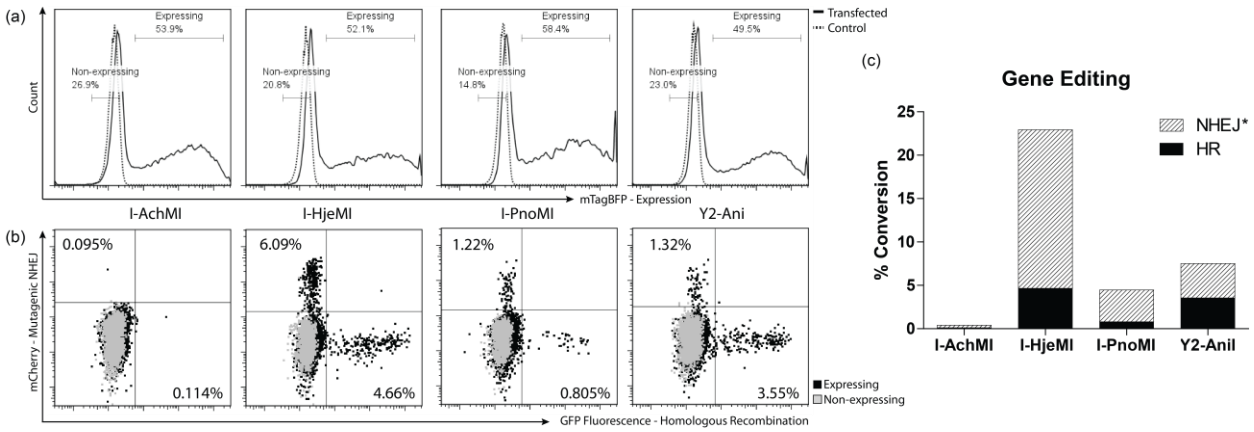


Figure 16. I-AniI homolog activity *in vivo*. (**a**) Nuclease-expression histogram. Number of cells (Y-axis) of a given mTagBFP fluorescence (X-axis) are shown to be uniform for all transfected cells (solid line), and are compared to an untransfected control (dashed line). Gates used for comparison of expressing and non-expressing populations in panel B are shown. (**b**) Mutagenic non-homologous end-joining (NHEJ) and homologous recombination (HR) repair events are shown for each nuclease-expressing population (black) compared to the non-expressing (gray). NHEJ events are mCherry-positive (Y-axis), and HR events, GFP-positive (X-axis). (**c**) Since each mCherry(+) cell represents approximately one third of the actual mutagenic NHEJ events[110] (**Figure 15a**), a corrected value is plotted for NHEJ events, calculated by multiplying the number of mCherry(+) cells by three. Cells with converted loci, by event type, are shown as a percent of the total expressing population.

I-AchMI exhibited little to no *in vivo* activity, consistent with its poor performance in the yeast tethered flow cleavage assay; this low activity may reflect either an actual reduced catalytic efficiency, or that impaired protein folding and/or thermal stability limits accumulation of active enzyme in cells cultured at 37 °C. For these reasons, I-AchMI should be considered a poor engineering scaffold for *in vivo* applications. In contrast, I-PnoMI and especially I-HjeMI, demonstrated repair of the GFP reporter by homologous recombination at frequencies much

higher than that of native I-AniI (which is nearly undetectable[109]) and comparable to the engineered, high activity variant Y2-Ani[93] (**Figure 16b**,**c**). Furthermore, remarkably high levels of mutagenic NHEJ were observed for I-HjeMI in the traffic light reporter assay: about three-fold higher than those stimulated by Y2-Ani and I-PnoMI. Thus, biotechnologically relevant activities appear to vary substantially among this group of closely related proteins.

3.3 Enhancing gene disruption efficiency

Although the LHE homologs achieved high levels of targeted gene disruption, even the best-performing enzyme (I-HjeMI) in its non-mutated form could not achieve complete knockout. Furthermore, re-designing enzymes to recognize alternative targets further reduces their mutagenic capacities. We therefore wanted to determine whether Trex2 could elevate rates of targeted gene disruption in a general fashion.

First, Dr. Certo validated our preliminary Trex2 assumptions using I-SceI. Expression vectors were created with the exonuclease, Trex2, and mTagBFP for tracking transfection efficiency (**Figure 17a**). To measure the rate of targeted disruption, these constructs were transfected into a human HEK293T cell line containing a Traffic Light Reporter (TLR) harboring the cognate I-SceI target[110]. Neither I-SceI^{D44A} (catalytically inactive) nor I-SceI^{D44A} plus Trex2 induced any measurable gene disruption, but I-SceI expressed with Trex2 produced a nearly-maximal increase in mCherry-positive cells (up to almost 33%) compared to I-SceI alone (**Figure 17b**; see manuscript for underlying flow cytometry data[109]).

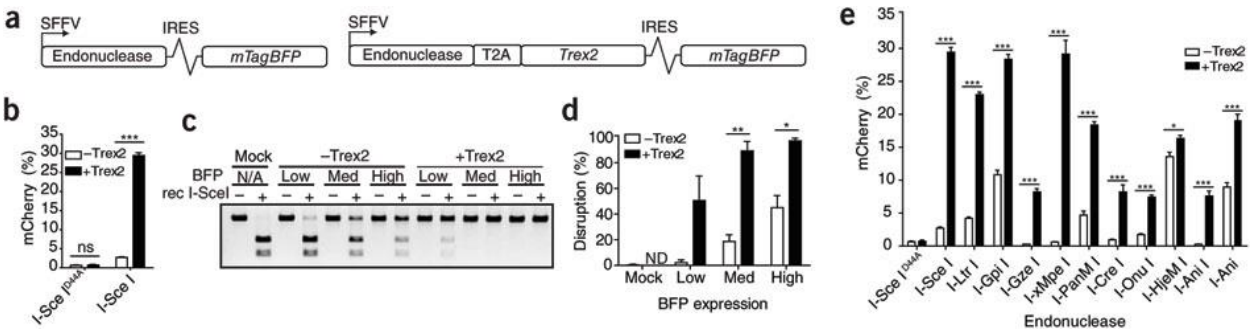


Figure 17. Gene disruption with Trex2. (**a**) Schematic of Trex2 exonuclease expression vectors, driven by the spleen focus-forming virus (SFFV) promoter/enhancer, and linked to BFP expression by an internal ribosomal entry site (IRES). (**b**) Quantification of mCherry expression in BFP-positive HEK293T cells transfected with the indicated vectors and analyzed 72 hours after transfection ($n = 3$). (**c**) Recombinant (rec) I-SceI-digestion gene disruption assay in cells sorted based on nuclease/BFP expression. (**d**) Quantification of band intensity (undigested out of total) of three independent experiments of the Sce digestion assay as performed in **c**. ND, not determined. (**e**) Quantification of mCherry expression in BFP-positive HEK293T cells harboring the respective TLR target for each of the indicated homing endonucleases ($n = 3$). Error bars, s.e.m. P values (*$P < 0.05$, **$P < 0.005$ and ***$P < 0.0005$) were calculated using the Student's two-tailed unpaired t-test to compare the samples indicated.

Next, we wanted to determine how Trex2-facilitated gene disruption was influenced as a function of nuclease expression. Cells in the above assay were gated on BFP (nuclease) expression and re-analyzed, revealing that near-maximal gene disruption rates occurred at even low endonuclease expression levels. To validate these data, Certo *et al.* sorted cells from varying nuclease expression levels (using BFP as a proxy), and the sequence flanking the I-SceI target from each of the populations was PCR-amplified. These amplicons were digested to enable us to assay mutagenesis of the I-SceI target site (target mutagenesis prevents cleavage, **Figure 17c**). At low endonuclease expression levels, we observed a 25-fold increase in total gene disruption between I-SceI and I-SceI+Trex2 (2.2% to 50.2%, respectively), and nearly 100% of targets were disrupted in the medium and high expression gates of I-SceI+Trex2 (90.3% and 97.1%, respectively) (**Figure 17c**,**d**). In contrast to I-SceI alone, which exhibits a dose-dependent increase in gene disruption, I-SceI+Trex2 quickly saturates disruption. Sequence analysis of the I-SceI target site in cells highly expressing I-SceI confirmed that 100% of cells were modified in the I-SceI+Trex2-treated cells. These data indicate that even at low levels of endonuclease

activity, high amounts of targeted disruption can be achieved. Trex2 may therefore effectively rescue homologs that may be otherwise useful in their ability to recognize new targets, but which natively have lower levels of activity.

Finally, we evaluated the general applicability of Trex2-enhanced gene disruption for breaks induced by a panel of LHEs. To this end, we made HEK293T cell lines containing TLRs harboring the cognate target sites for each LHE. We transfected each cell line with its respective enzyme with and without Trex2. This experiment produced results consistent with the increase in disruption rates associated with Trex2 coupling, even with endonucleases that exhibit nearly undetectable activity when expressed alone (e.g. I-AniI, **Figure 17e**). Although Trex2 consistently elevated the rates of gene disruption, not all LHEs were affected equally. Of note, I-HjeMI alone produced considerable levels of mutagenic NHEJ, and addition of Trex2 only increased disruption slightly (~15%) compared to other enzymes (200% ~ 1000%). The cause of these differential effects remains speculative (discussed briefly in this chapter's summary), and warrants further investigation into additional homologs with varied *in vivo* properties.

As a final note, these findings were also validated in more physiological models by Certo *et al*. Native I-SceI targets were similarly modified by Trex2 in murine embryonic fibroblasts (MEFs). These experiments demonstrated that Sce+Trex2 expression produced a six-fold increase in disruption at an unexpressed locus over that of I-SceI alone (I-SceI = 15.8%, I-SceI+ Trex2 = 88.7%)[109]. Additionally, the knockout efficiency of an engineered I-CreI variant targeted to CCR5 was approximately seven-fold in a human cell line, compared to endonuclease alone[109]. For all of the above experiments, we observed no toxicity as a result of overexpressing Trex2 in conjunction with any endonuclease, nor did we observe any effects of Trex2 on cell cycle or

growth kinetics in the presence or absence of active endonucleases. Although we might have expected that coupling Trex2 would result in increased mutagenesis at off-target sites, we observed minimal activity both with and without Trex2 at two previously characterized off-targets for the *VegF* zinc finger nuclease[37] despite clearly detectable activity at the endogenous locus[109]. Further, we and others[115] have found that the expression of Trex2 decreases the frequency of distal end joining, which suggests that it may also decrease propensity for deleterious translocations.

3.4 Homologs in vivo: re-design of I-OnuI

The most important aspects of preparing a homing endonuclease for use *in vivo* are: choosing a specific genomic target within a gene of interest based on likeness to existing homolog targets, modifying the chosen homolog's specificity to match that target, and evaluating the induced modifications *ex vivo*. We evaluated this process using I-OnuI, which recognizes a target site that is closely related to an endogenous human gene target. We engineered I-OnuI to cleave a DNA sequence that is found in the third exon of the human MAO-B gene and that differs from the WT I-OnuI target site at only five base-pair positions (**Figure 18a**). MAO-B is one of two monoamine oxidases localized on the mitochondrial outer membrane, where it oxidizes neurotransmitters and dietary amines and produces hydrogen peroxide as a byproduct (a known oxidative cytotoxin). MAO-B is a potential therapeutic target for a wide variety of neurodegenerative disorders including Parkinson disease (PD)[116]. Pharmacological MAO-B inhibitors appear to slow the progress of PD symptoms via a neuroprotective activity, but the disease-modifying effect and mechanism of action of the inhibitors have been controversial[117–121]. The ability to generate tissue-specific disruption or modifications of the MAO-B gene might

therefore be a valuable tool for future clinical research. The target sequence in MAO-B is completely conserved among the primates shown, and only slightly diverged in other mammalians (**Figure 18a**).

The Stoddard lab used directed evolution of the I-OnuI LHE to produce an enzyme that targeted the MAO-B gene in a human cell line[52]. The enzyme's *in vivo* activity was initially validated using an episomal system similar to the TLR that reports only homologous recombination by repair of a GFP gene[122]. Although the ability of the modified enzyme (E2 I-OnuI) to induce targeted repair was reduced compared to wild-type enzymes, there was still a pronounced effect specific to the MAO-B target (**Figure 18b**). This specificity was also

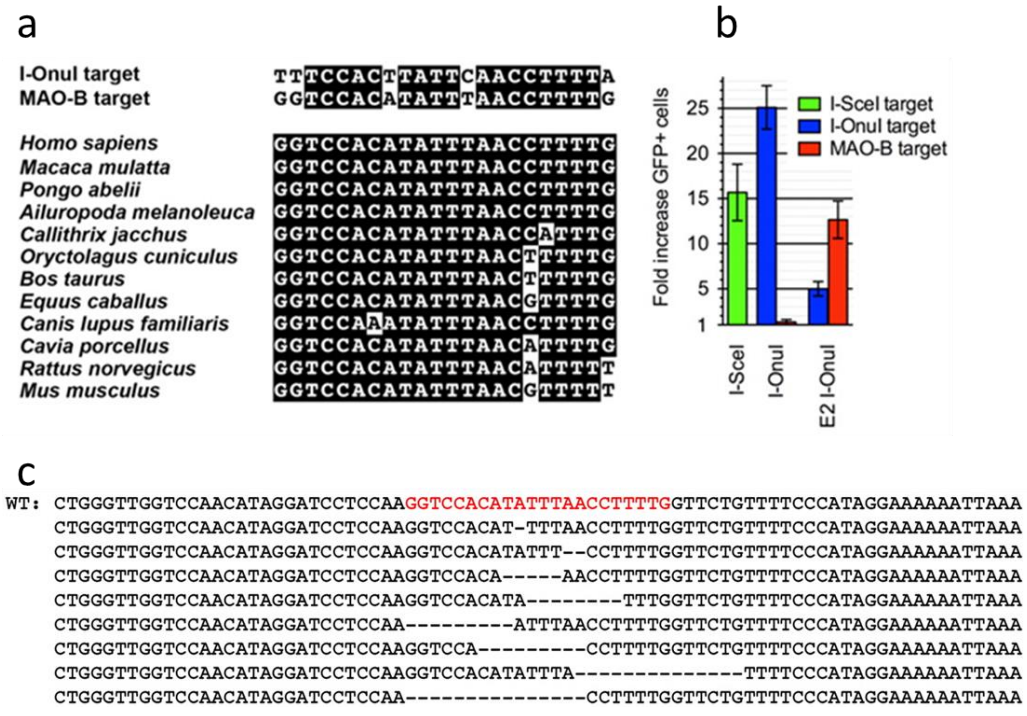


Figure 18. I-OnuI redesign. (**a**) MAO-B target Sequence alignment of the I-OnuI and MAO-B target sites (*Upper*) or the human MAO-B target and the corresponding sequences coded in other mammalian genomes (*Lower*). (**b**) An episomal GFP gene HR assay was carried out as described previously[122]. Error bars refer to ± SD of three independent experiments. (**c**) Putatively mutated (cleavage-resistant) amplicons from E2 I-OnuI-transfected cells (sorted for expression) were analyzed by sequencing. Small deletions were found within the endogenous MAO-B target site. The intact genome sequence is shown on the top (WT), and the MAO-B target site is in red.

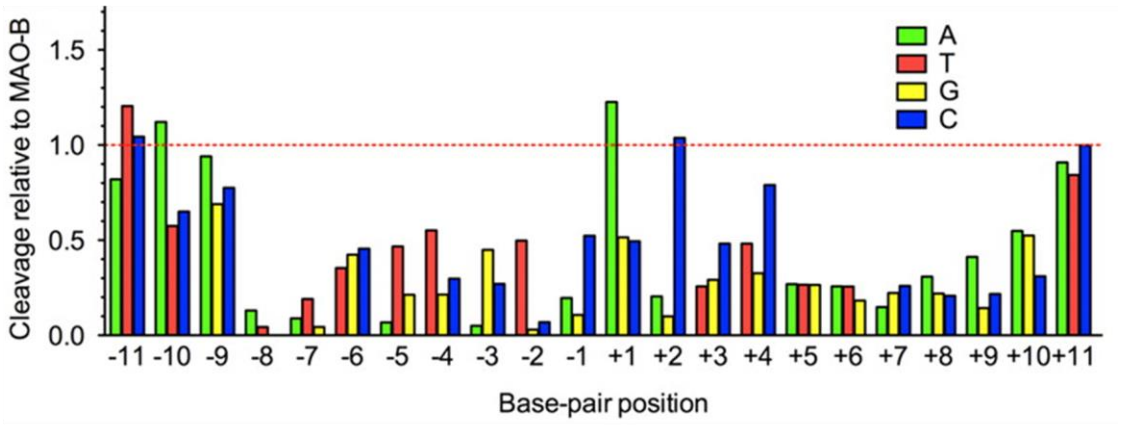
validated in an *in vitro* cleavage assay. The engineered enzyme's binding affinity as well as its stability (as assayed by western blot) were also maintained[52]. We verified targeted mutagenesis *in vivo* by transfecting human HEK 293T cells, sorting for the top two highest quartiles of transfected cells, and PCR amplifying around the target locus. The locus was assayed for target mutagenesis by cleavage by a recombinant enzyme, as previously described[109,123] (see section 3.3 Enhancing gene disruption efficiency and **Figure 17c**), and subsequently sequenced (**Figure 18c**).

We also determined the DNA sequence specificity profile of E2 I-OnuI across the MAO-B target. Overall, the profile of E2 I-OnuI is very similar to that of the native I-OnuI (**Figure 12c**), though specificity is slightly reduced at positions -11, -10, -9, -5, +1, +2, and +11. However, specificity appears to have increased relative to that of native I-OnuI at position -3 (**Figure 19a**). The positions of attenuated specificity correlate well with the positions of mismatches between the native I-OnuI target and the MAO-B target (-11, -10, -4, +2, and +11). This correlation suggests that altered specificity as a result of amino acid substitution at the protein–DNA interface is confined to the regions of the altered residues.

Undesired, off-target cleavage events induced by engineered endonucleases could lead to gene disruption and/or a variety of additional undesirable mutagenic events and carcinogenesis. To test whether off-target cleavage by the WT or E2 I-OnuI enzyme was predictable on the basis of sequence homology to a desired chromosomal target site, we conducted a BLAST search for DNA sequences in the human genome that are similar to the central 18 base-pair sequence of the MAO-B target site. We concentrated on these positions because these endonucleases (particularly E2 I-OnuI) were already known to tolerate single base-pair mismatches at the most

distal base-pair positions in their respective targets (base-pair positions ± 10 and ± 11; see **Figure 12c** and **Figure 19a**). Four potentially cleavable chromosomal loci (**Figure 19b**) were investigated by the same cleavage assays used to detect mutagenesis at the endogenous MAO-B locus. None of these sites were located within protein-coding regions. E2 I-OnuI appeared to induce mutations at off-target sites #1 and #2, while WT I-OnuI accumulated apparent mutations at the off-target site #2 (data not shown). In contrast, no mutagenesis induced by either enzyme was significantly detected at the off-target sites #3 and #4 over the slightly high background in the single round of digestion with E2 I-OnuI protein. However, a gel of an additional *in vitro* digestion of the re-amplified cleavage-resistant fragments showed infrequent indels. These results suggest that potential off-target cleavage sites can be predicted by a sequence homology

a



b

	Sequence	Chromosome
MAO-B target	GGTCCACATATTTAACCTTTTG	X
I-OnuI target	TT.....T....C.....A	
Off-target #1	TA.....	13
Off-target #2	.T.....C.....T	4
Off-target #3	T..A.....A	6
Off-target #4	TC.....CA	4

Figure 19. E2 I-OnuI specificity. (**a**) Specificity profile. Tolerance of each base substitution at each position was measured, as was done for wild-type I-OnuI in **Figure 12c**. (**b**) List of the sequences closely matched to the MAOB target in human genome.

search of genome sequence, and off-target effects of an engineered endonuclease can be detected.

3.5 Summary

The *in vivo* experiments presented here demonstrate the ability of newly characterized homing endonucleases to assist engineering efforts. First, closely-related LHEs that recognize similar targets can have a diverse range of *in vitro* and *in vivo* properties. I-PnoI and I-HjeI both displayed enhanced *in vivo* activity compared to the original homolog I-AniI, which was able to elicit genomic changes only after directed evolution to increase its cleavage activity. I-HjeI also demonstrated an altered ratio of induced mutagenic NHEJ to HR, as assayed by the traffic light reporter. This result, combined with I-HjeI's comparatively minimal change in this ratio when co-expressed with Trex2, begs the question of whether an LHE's only role is to induce a break. It would appear that different LHEs can affect the type of repair performed at their targets, though the mechanism for this remains a mystery. One possibility may be that differences in binding affinity to one or both of the broken DNA ends affects the type of repair performed. Alternatively, LHEs might also have protein domains to interact with host break repair machinery given their natural mode of DNA break repair-dependent propagation. These preliminary *in vivo* data combined with the ability to rapidly survey natural LHEs described in Chapter 2 will empower future studies in this area.

We also demonstrated the advantages of co-expressing an exonuclease with LHEs. First, Trex2 lowers the minimum activity required of LHEs for activity *in vivo*. The exonuclease allows a greater range of LHE homologs to act as scaffolds for engineering, and reduces the

quality requirements of engineered enzymes. This reduced quality requirement will translate to a reduction in time and effort spent engineering enzymes, and should also translate to a greater number of targetable loci. Second, endo-exo coupling increases the potential for making multiple changes in a single round of mutagenesis via multi-allelic knockouts and multiplexing. Third, elimination of 'persistent' breaks, owing to reduced cleavage cycling, is a potential safety benefit, limiting translocations and other deleterious repair events[115]. Overall, implementing an endo-exo experimental design will facilitate use of LHEs in genome engineering.

Finally, we demonstrated the power of using a newly-characterized LHE for genome engineering. Using I-OnuI, we were able to access a target that would have been difficult or impossible to engineer using I-AniI. Furthermore, the reduction in catalytic efficiency due to the mutations required to recognize the new target site was still acceptable, likely owing to the initial robustness of the homolog. This study further highlights the utility of characterizing new LHEs with altered specificities.

3.6 Methods

Assessment of *in vivo* gene modification activity using the Traffic Light reporter

Each LHE's target site was ligated into the truncated green fluorescent protein (GFP) of the Traffic Light Reporter[110] using annealed, phosphorylated dsOligo (**Figure 15a**). Lentivirus containing this construct was used to transduce HEK 293T cells at limiting dilution in order to obtain a population of cells with single copy chromosomal integration events. Cells were sorted against GFP and mCherry fluorescence to ensure that the reported started in the "off" state. Endonuclease expression/GFP repair template vectors were generated by cloning each LHE from

the yeast surface display vectors into the Lentiviral backbone containing the GFP repair fragment (**Figure 15b**). ORFs were ligated in frame with a self-cleaving T2A peptide sequence, followed by a blue fluorescent protein, mTagBFP, to allow expression levels to be measured. On day 0, $1x10^5$ HEK cells of each reporter cell line were plated. On day 1, each reporter cell line was transfected with 400 ng of LHE expression/repair plasmid with polyethylenimine at a wt/wt ratio of 4:1 in a pH 7 150 mM NaCl, 5 mM HEPES buffer. Cell media was replaced on day two, and cells were allowed to accumulate conversion events until day 4, when they were analyzed by flow cytometry on a BD LSRII™. Using FloJo (TreeStar), each expressing population was defined by mTagBFP fluorescence. GFP+ and mCherry+ statistics, representing homologous recombination and mutagenic non-homologous end joining events respectively, were tabulated. mTagBFP positivity was determined by comparison to non-transfected cells for each cell line; GFP and mCherry positivity by comparison to non-expressing populations in the transfected cells. In order to ensure that the non-expressing population was truly not expressing the construct, a small number of the highest mTagBFP-low cells were excluded from the non-expressing population.

Mutagenic NHEJ detection by digestion

Genomic DNA was isolated from bulk or sorted cells as indicated using Qiagen's DNeasy kit (Qiagen), and TLR target sites were PCR amplified as previously described[110]. 100 ng of each PCR product was digested *in vitro* with LHE for 6 hours at 37 °C. DNA was separated using a 1% agarose gel stained with ethidium bromide. Percent disruption was calculated by quantifying band intensity using ImageJ (NIH) and dividing the intensity of the undigested band by the total.

Mutagenic NHEJ detection by sequencing

Genomic DNA was isolated from cells using Qiagen DNeasy kit (Qiagen) and the LHE target region was amplified as previously described[110]. PCR products were cloned using a CloneJET PCR cloning kit (Fermentas) according to the manufacturer's protocol, followed by transformation into chemically competent DH5α *Escherichia coli* bacteria. Bacterial colonies were directly sequenced using a standard colony-sequencing protocol. Sequences were analyzed using the Contig Express software provided in Invitrogen's Vector NTI software suite.

Episomal GFP recombination reporter assay

To express a LHE, the LHE gene including the Nterminal HA tag followed by a nuclear localization signal was linked to mCherry gene by the 2A peptide sequence from *Thosea asigna* virus in the pExodus plasmid. The two-gene expression was driven by CMV promoter and the cotranslated proteins were separated by ribosomal skipping[124]. DR-GFP reporter codes a GFP gene sequence interrupted by a LHE target site and an in-frame stop codon, followed by the truncated gene sequence[122]. HEK 293T cells were grown in DMEM supplemented with 10% fetal bovine serum, 10 units/mL penicillin and 10 μg/mL streptomycin at 37 °C in 5% CO_2 atmosphere. HEK 293T cells (6×10^4) were plated 24 hours prior to transfection in 12-well plates, and transfected with 0.25 μg each of DR-GFP reporter and pExodus plasmid using Fugene 6 transfection reagent (Roche Applied Science). The GFP-positive cells were detected by flow cytometry at 48 hours post transfection. Western blotting was carried out using a rabbit polyclonal antibody against HA-epitope tag and a mouse monoclonal antibody against β-actin.

Targeted mutagenesis and detection at the MAO-B Locus

HEK 293T cells (1.3×10^5) were plated 24 hours prior to transfection in six-well plates and transfected with 1 μg of pExodus plasmid. The top 25% and the following 25% of mCherry positive cells (fluorescent marker for a LHE gene expression) were separately collected using BD FACSAria cell sorter (BD Biosciences) 48 hours post transfection. To extract genomic DNA, the sorted cells (approximately 1×10^5) were washed with cold PBS buffer, resuspended in TNES buffer [50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 10 mM EDTA, 1% SDS, 0.25 mg/mL proteinase K], and incubated at 50 °C for 30 minutes. RNase A was added to 0.25 mg/mL, and the reaction mixture was further incubated at the same temperature for 30 minutes. The genomic DNA was recovered by phenol/chloroform/isoamyl alchohol extraction followed by ethanol precipitation. Both on-target (i.e., MAO-B gene) and off-target loci were amplified from 50–80 ng of the extracted genomes using Phusion DNA polymerase (Finnzymes). The DNA products resulting from two rounds of PCR amplification were cleaned using PCR purification kit (Qiagen), and 150 ng of the fragments were incubated with 1.5–3.0 pmol of E2 I-OnuI recombinant protein in 20 mM Tris-acetate (pH 7.5), 100 mM potassium acetate (pH 7.5), 1 mM DTT, and 10 mM $MgCl_2$ at 37 °C for 2 hours. The cleavage reactions were terminated by adding 4× Stop solution. After incubation at 37 °C for 30 minutes, each sample was separated on a 1.8% agarose gel containing ethidium bromide in TBE. The DNA bands were quantified using ImageJ software.

Chapter 4
First Principles Approach to Target Determination

4.1 Introduction

The diverse collection targets associated with homolog discovery has proven to be a powerful aid to LHEs' use as genome engineering tools, but LHE discovery is still limited by one dominant factor: target prediction. Careful examination of host intron-exon borders or comparison with an insert-less allele can reveal the HE's native target site, given the mode of LHE propagation[125,126]. The insert-free allele must possess the site that the LHE originally cut, or "homed" to, when creating the initial double-strand break that led to its introduction into the host allele[127]. While this method of target determination can work, it explicitly requires precise intron-exon borders or sequence of the insert-less allele for comparison; something frequently unavailable, especially in the case of LHEs identified in partial or metagenomic collections. Furthermore, in the case of engineered variants (e.g. chimeras), it may not be possible to predict the target site despite the possible functionality of the new enzyme. Consequently, although we may predict the existence of a novel homing endonuclease– or more generally, any DNA-binding protein – by homology, the frequent lack of a predictable DNA target precludes analysis of the protein's fundamental characteristics.

To lessen our reliance on frequently unavailable sequence data, and thereby better exploit the vast number of putative LHEs being identified, I devised a first principles method for determining a DNA-binding protein's target specificity. This approach is a combination of two existing technologies: yeast surface display, which is already the foundation of our characterization pipeline, and Systematic Evolution of Ligands by Exponential Enrichment

Selection (SELEX)[128,129]. In essence, the protein of interest is expressed on the surface of yeast, which acts as a solid support for the binding of an initially randomized library of potential DNA targets (under non-cleaving conditions); the unbound DNAs are washed away, and bound targets are released, amplified, and subjected to further selection in an iterative fashion known as SELEX (**Figure 20**). Like YSD-based assays, SELEX can be carried out in a parallel fashion (i.e. many enzymes at once in a 96-well plate). The output of a successful SELEX experiment is a collection of target sequences that are best bound by the protein (LHE) of interest.

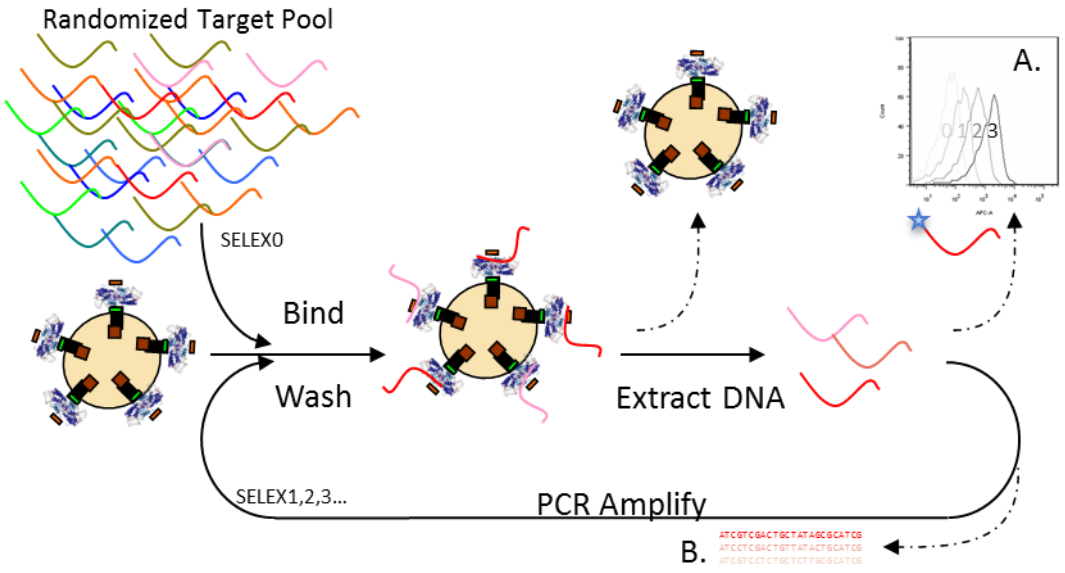


Figure 20. Schematic of SELEX using yeast surface displayed protein. The homing endonuclease of interest is expressed on the surface of yeast. The yeast is used as a solid support for the protein to allow binding of randomized oligos and subsequent wash steps. The best-bound sequences are extracted by heating and collecting the supernatant, amplified, and subjected to iterative rounds of selection. (**a**) The initial optimization and later, the success of the experiment are assayed using fluorescently labeled pools and flow cytometry. (**b**) After sufficient enrichment for high affinity targets, the oligos are sequenced and aligned.

4.2 Adapting SELEX for Yeast Surface-Displayed LHEs

I first hypothesized that the SELEX protocol could be adapted to use yeast surface-displayed LHE as the protein expression system and the solid support. SELEX is typically

carried out with protein synthesized *in vitro* and coupling it to a solid support (by biotinylating the protein and using streptavidin magnetic beads, for example)[129,130]. This complex allows the protein to be incubated with the randomized oligo pools and washed, and the bound oligo to be retrieved. Using yeast as both the method of expression and as the solid support would cut down on experimental complexity and cost, and provides a suite of built-in tools. Yeast surface display produces large amounts of inexpensive solid support-coupled protein, removing the need for specialized protein production and labeling kits, magnetic beads, or antibodies. YSD would improve SELEX by providing a means of expression quality control (by flow cytometry), and binding and cleavage analysis throughout the experiment as described in Chapter 2; SELEX would improve YSD by providing an additional tool for the engineering and characterization suite. Second, I hypothesized that we could capture and amplify cleavable substrate with LHEs, while *not* cleaving them during selection. Cleavage was prevented by replacing the catalytically required Mg^{++} with Ca^{++} during the binding selection, as is done for the binding experiments and cleavage negative controls in Chapter 2. This substitution has been found to only minimally alter binding specificity, if at all[73].

Still, care needs to be taken when considering the use of binding selection to determine LHE target sites. SELEX will provide the investigator with a set of sequences best bound by a given protein. Although binding discrimination is integral to achieving cleavage specificity, binding and cleavage specificity are not synonymous[77]. That said, we hypothesized that the correlation between binding and cleavage specificity would be tight enough that cleavable substrates could be produced from SELEX. Tight correlation of binding and cleavage specificity observed in I-OnuI was key preliminary evidence. Here, poorly-cleaved sequences are not bound tightly by the enzyme, and the best-bound sequences also tend to be the best-cleaved sequences

(**Figure 21**). In fact, all of the best-bound sequences are cleaved, even if at a lower efficiency. These data suggest that a binding-selection based approach, as with SELEX, should yield target sequences that could be cleaved.

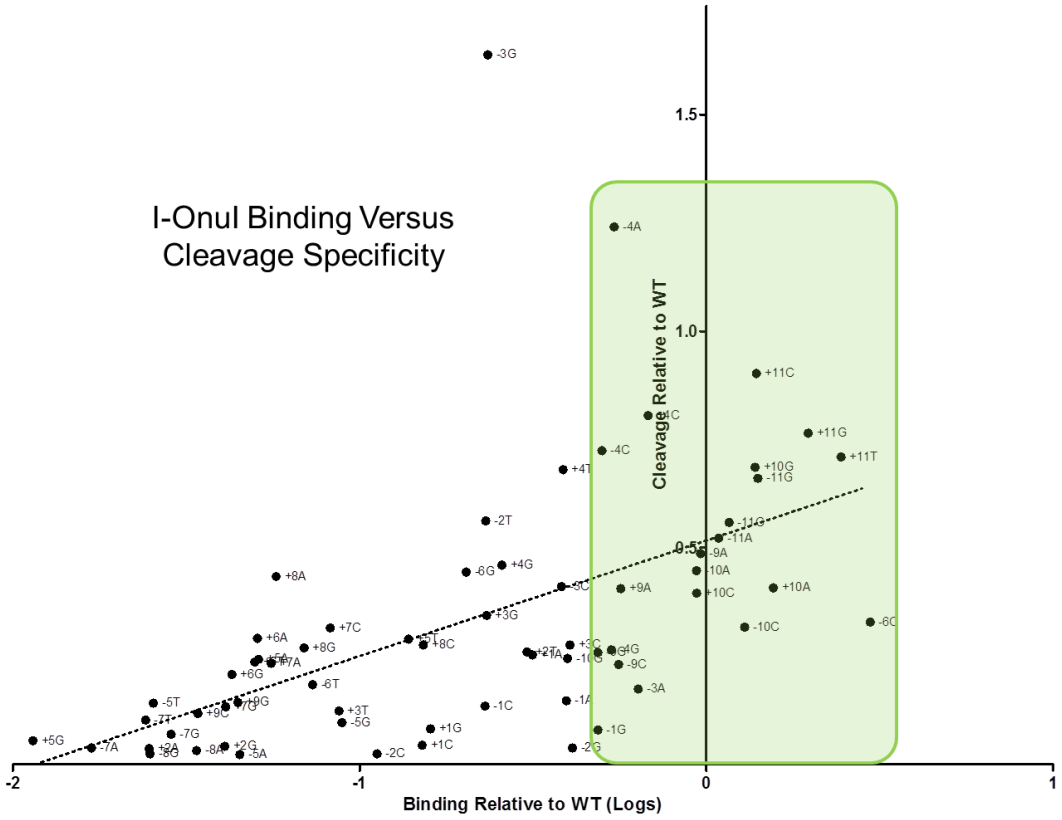
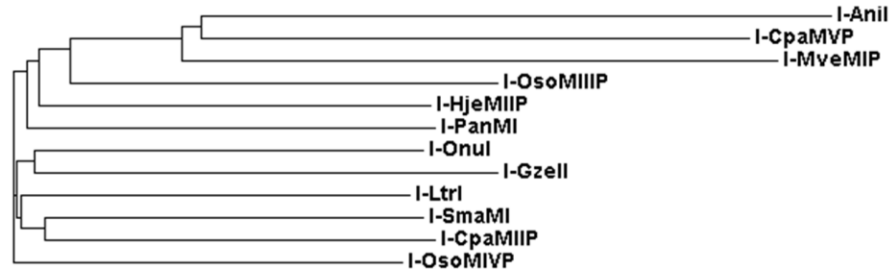


Figure 21. Binding versus cleavage specificity. This data was obtained by assaying the ability of each 1-off target to bound or cleaved by I-OnuI compared to the wild-type sequence using the binding and cleavage profiling protocols outlined in Chapter 2. Importantly, the best cleaved sequences are also the best bound (green box). In fact, the binding of poorly-cleaved targets (even targets altered at only one base) drops off rapidly.

To test our hypotheses, we selected several enzymes with known targets, as well as putative LHEs with no predictable target sites. Six characterized enzymes of the I-OnuI subfamily and I-AniI were included from Chapter 2 as a benchmark for the protocol: I-AniI, I-GzeII, I-LtrI, I-OnuI, I-PanMI, and I-SmaMI. Six putative nucleases with good surface expression, but which had insufficient genomic data to predict a target site, were included as the

test group: I-CpaMIIP, I-CpaMVP, I-HjeMIIP, I-MveMIP, I-OsoMIIIP, and I-OsoMIVP. These enzymes varied in the level of homology to each other (**Figure 22a**) and which genes they were inserted in (**Figure 22b**), though they were all found in yeast mitochondrial genomes. These putative I-OnuI homologs would be some of the most distant relatives characterized to date – the closest characterized relatives being 48% identical at the amino acid level (I-SmaMI and I-CpaMIIP), and the most distant being 26% identical (I-CpaMVP and I-PanMI). Most of these homologs would also be in new insertion points in the same or different host genes, and likely significantly different target sites.

a



b

Enzyme	Gene Inserted Into	Insertion type
I-AniI	Cytochrome b (cobA)	Intron
I-CpaMIIP	Ribosomal protein subunit 3 (RPS3) in LSU intron	Fusion (type C)
I-CpaMVP	Small subunit ribosomal RNA (rns)	Intron
I-GzeII	NADH:ubiquinone oxidoreductase subunit 2 (nad2)	Intron
I-HjeMIIP	Undetermined (possibly cytochrome c oxidase subunit 2)	Undetermined
I-LtrI	Ribosomal protein subunit 3 (RPS3) in LSU intron	Fusion (type B)
I-MveMIP	Small subunit ribosomal RNA (rns)	Intron
I-OnuI	Ribosomal protein subunit 3 (RPS3) in LSU intron	Fusion (type C)
I-OsoMIIIP	Small subunit ribosomal RNA (rns)	Intron
I-OsoMIVP	Small subunit ribosomal RNA (rns)	Intron
I-PanMI	NADH:ubiquinone oxidoreductase subunit 3 (nad3)	Intron
I-SmaMI	Cytochrome oxidase subunit 1 (COX1)	Intron

Figure 22. SELEX homing endonuclease set. (**a**) The phylogram resulting from a ClustalW2 multiple alignment of the protein sequences of the set of LHEs in the SELEX experiments. (**b**) The gene into which each enzyme is inserted into gene is given, as well as the predicted type of insertion.

Initial experiments focused on optimizing the binding conditions for SELEX for LHEs. We chose a concentration for the randomized library that was close to the approximate K_d of our

enzymes (~0.3-30 nM, see **Figure 8** and **Figure 12**), but below the onset of significant non-specific binding (~50-100+ nM). Minimizing the level of non-specific binding is key to a successful SELEX experiment[131]. When the proportion of bound enzyme to unbound enzyme ([ES]/[E]) is too high, selection of high-affinity targets is unable to occur; we saw no increase in binding affinity to the library as a function of SELEX round number (data not shown). We increased the salt concentration of the bind/wash buffer, which has been found to decrease non-specific LHE-DNA interactions[73], until only a small amount of the naive library could bind (~5%). We found that nonspecific competitor nucleic acids (poly[dI-dC]) as used by others in similar experiments[114,132] had an insignificant effect on total binding (data not shown). Under salt-optimized conditions, binding affinity increases as expected in each subsequent round of iteratively selected library (**Figure 23**). Similarly, we found it necessary to carry out binding at room temperature, rather than 4 °C (data not shown). Since the proportion of bound to unbound enzyme increases with each round of SELEX – especially after round 3 – we found it necessary to increase the stringency of selection by increasing the salt concentration for later rounds. This modification is similar to fixed-stringency SELEX[129]. Under these conditions, each LHE, except I-CpaMVP, showed increased affinity for targets in each subsequent round of SELEX. Possible explanations for I-CpaMVP's failure include: the initial conditions did not allow any oligo binding at all; the portion of the enzyme that was cloned was too short or long; or the enzyme is simply no longer functional. The salt and/or nonspecific competitor concentration (described in the methods) would likely need to be readjusted depending on the class of proteins being studied. This process would likewise be aided by the ability to perform high-throughput binding analysis using YSD, as shown here. The hallmark of a successfully optimized SELEX reaction should be

evidenced by an increase in binding to the oligo pool for each subsequent round of SELEX, as in **Figure 23**.

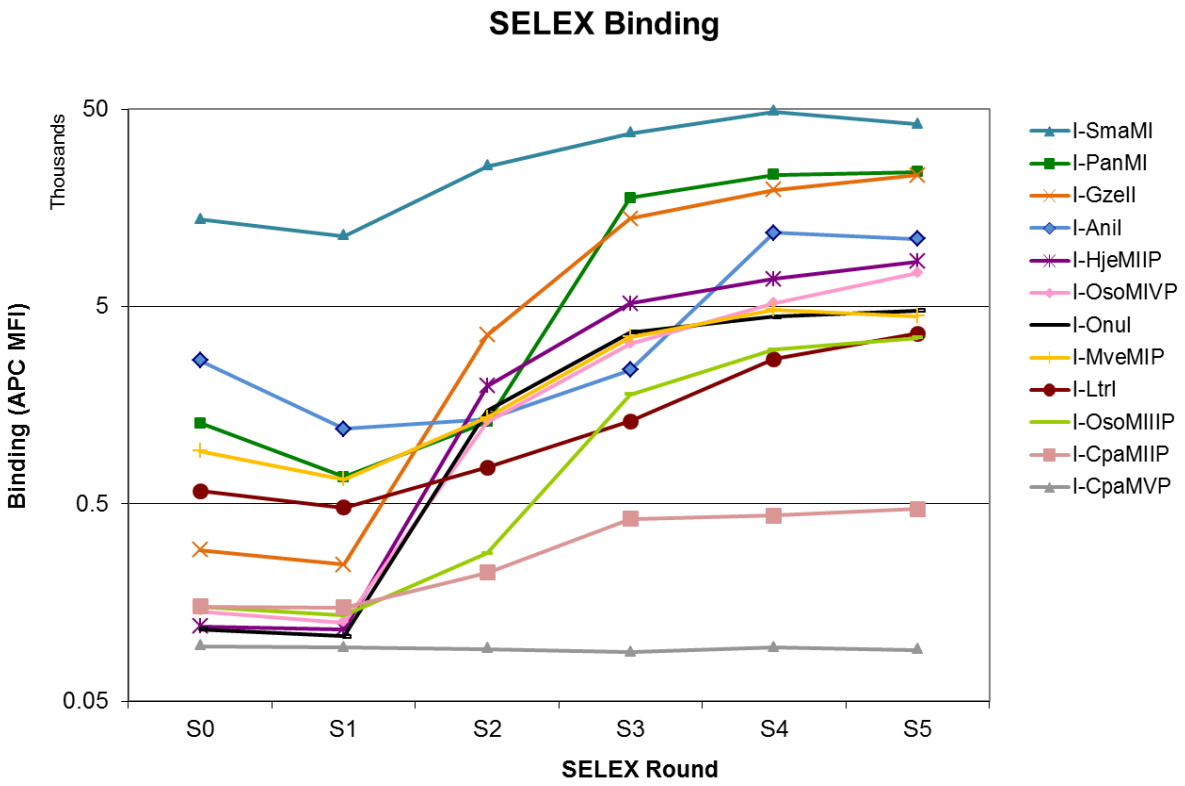


Figure 23. SELEX pool binding versus round number. These binding experiments were carried out by fluorescently labeling the SELEX pools from each round and measuring the fluorescence of yeast surface displayed enzyme via flow cytometry. Each enzyme's affinity for the subsequent SELEX pool (y-axis) increased with each round (x-axis) with the exception of I-CpaMVP. S0, or round 0 is the initial randomized pool.

A number of other protocol parameters required optimization as well. First, the PCR conditions used to amplify the randomized pool between rounds had to be adjusted to minimize the background. We found that 20 rounds using the primers described in the methods section was sufficient to obtain amplification with low background (data not shown). Next, the number of randomized bases was adjusted. We found that fewer randomized bases impaired (but did not prevent) selection. One explanation may be that the lower number of randomized bases per oligo translates to a smaller number of possible 20-mers per mol of library. The total concentration of

library used, and therefore the total complexity of the pool, is limited based on the maximum allowed oligo concentration used (to prevent nonspecific binding). For example, a single 60-mer oligo with 20 randomized bases (plus 2x 20 bp flanking PCR primers) has 2 possible 20-mer targets (including the antisense target). An oligo with 30 randomized bases would have 22 possible 20-mers – 10-fold as many, albeit related to each other – while only increasing in length by <1.2-fold (60 bp to 70 bp). Finally, we had to optimize our method of target release after the binding and washing steps. We found that heating the yeast to 70 °C was sufficient to promote release of the oligo, as assayed by a flow cytometric binding assay (data not shown). The oligo is released at this temperature because the LHEs used likely had melting temperatures below 70 °C[93]. For enzymes with melting temperatures significantly higher than this, it may be necessary to capture the oligo by releasing the enzyme itself (by reducing the disulfide bonds that anchor it to the yeast). Once these factors were optimized, surface-displayed LHEs were able to select high-affinity targets, as shown in **Figure 23** and in the next section.

Preliminary analysis of the SELEX pools from rounds 3-5 showed that an appreciable fraction of the selected targets were cleavable. The same fluorescent oligo made for the yeast flow binding analysis above was used for cleavage analysis. Pools from rounds 2-5 were assayed for their corresponding enzyme's ability to cleave them (**Figure 24**). Since no targets should be present in high numbers in round two[133], cleavage of this pool should serve as a conservative estimate of background noise. Three of the six enzymes with known functionality (I-GzeII, I-OnuI, and I-PanMII) were able to select cleavable targets, signifying the success of the SELEX experiments with these enzymes. Two previously uncharacterized enzymes (I-HjeMIIP and to a lesser extent I-OsoMIIIP) also demonstrated activity against an appreciable fraction of the selected oligos. While this assay unequivocally indicates the functionality of enzymes that

showed cleavage, and the success of the selection, those enzymes that were not able to cleave a significant fraction of the pool may still have meaningful SELEX output. It is possible that the enzymes selected cleavable oligos, but they did not represent enough of the total population to be detected; or alternatively, the enzymes selected for positions that would allow cleavage, but those positions were not all present in a single, highly abundant oligo. In either case, enzymes that failed this preliminary test may yet reveal their binding specificities through sequence analysis in the next section.

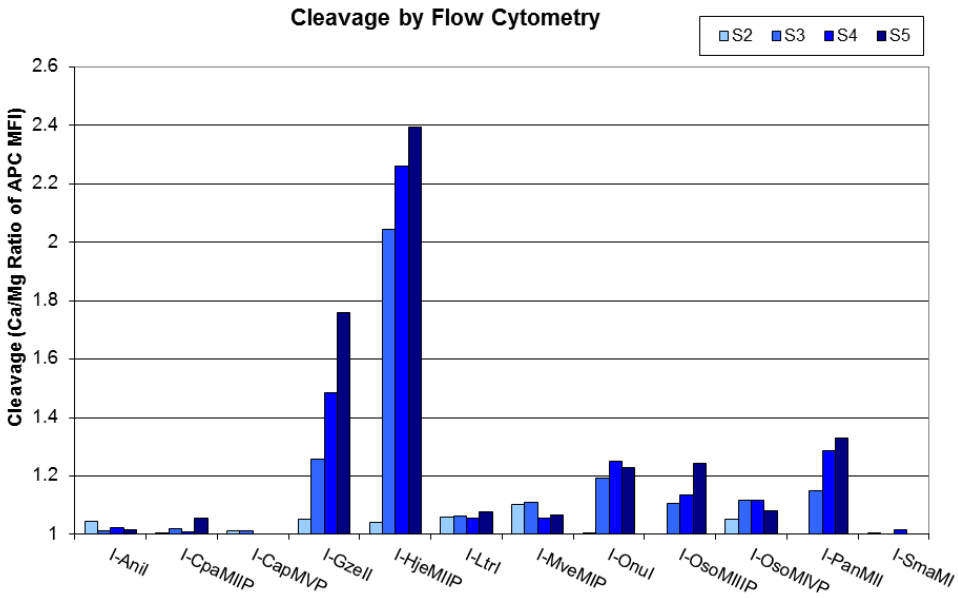


Figure 24. SELEX pool cleavage versus round number. The pools of selected oligo from round 2, 3, 4, and 5 were subjected to flow cleavage analysis, as in Chapter 2. Ca/Mg ratios above 1 indicate a loss of fluorescence in the catalytically active samples (containing Mg^{++}) due to cleavage of the target.

4.3 SELEX sequence analysis

Once the binding selection was complete, we needed to analyze the oligo sequences. Given the possible differences between binding and cleavage specificities, we also wanted to incorporate a cleavage selection step prior to sequencing. In an effort to maintain a parallel

experimental flow and avoid a lossy gel extraction step of an existing method[92], we opted to try a new, PCR-based cleavage selection protocol. SELEX pools would be subjected to cleavage, and cleavage products would be ligated via the newly exposed phosphate to a reverse primer. PCR using this reverse primer would allow specific amplification of cleaved products. However, since half of the target site would be lost in this process, we barcoded each sequence so that once the cleaved target was recovered, the barcode could reveal its original full-length sequence. To this end, each pool from each round (3-5) and each enzyme was barcoded with a unique 4-base barcode in addition to a unique 7-base randomized barcode. Sequencing the un-cleaved population (in depth) should allow us to match each 11-base barcode from the cleaved, partial sequences back to their full-length "parents." Barcoding in this fashion would allow the entire experiment to be sequenced as one sample using next generation sequencing.

Although sorting barcoded sequences is a trivial matter, matching barcodes and exporting corresponding uncut sequences would require specialized software. I developed a flexible sequence-handling framework that would allow import, parsing, matching, binning, and export of large numbers of sequences. Given the large dataset storage and searching capabilities of SQL the powerful string-handling of PHP, and the availability (and therefore portability) of both, I chose these software tools to build the framework. I created various sequence modules that could parse and handle multiple types of sequences (FASTA, FASTQ, etc.); data input/output modules that could read and write sequences from various sources (files, forms, databases); and various tools to work with these sequences (parsing, searching, trimming, etc.). More of these modules can be created and implemented or used for a different purpose at any time. This system is valuable because the functions can be run locally or remotely on a more powerful computer. Furthermore, the input/output modules and use of SQL minimize the required memory by

requesting or searching sequences one at a time, rather than all at once. This low memory footprint is ideal for dealing with large datasets, such as those produced by next generation sequencing. Once the foundation was in place, I created scripts to use these functions to execute the previously described process of sorting and matching sequences for each enzyme and round. The output was a collection of FASTA sequences for each enzyme (12) for each round (3-5), uncut and cut/matched. Unfortunately, fewer than 5% of cut sequences were able to be matched back to a parent uncut sequence. This lack of matches was primarily due to a lack of depth in sequencing of the uncut population (the barcode was missing), but was also complicated by mismatches (non-unique barcode collisions). To prevent this type of error in the future, more uncut sequences should be obtained and a smaller number of sequences should be barcoded. For this reason, subsequent analysis is primarily focused on the binding motifs made from the uncut sequences.

Once the sequences were sorted into their appropriate groups, they were analyzed for the presence of sequence motifs. Expectation maximization was used to discover motifs by using the Multiple EM for Motif Elicitation (MEME) tool[134], as described in the methods (**Figure 25**). Of the twelve enzymes analyzed, only I-CpaMVP failed to produce a motif, as expected from the SELEX binding analysis (**Figure 23**). Failure here is defined as not producing a strong motif that is similar in all three rounds. I-AniI produced a weak motif that was only moderately conserved across rounds 3-5, and did not match the known specificity. Of the remaining five characterized enzymes, I-GzeII, I-OnuI, I-PanMI produced motifs that closely mirrored their known specificities. I-SmaMI's motif was congruent with its known motif, except that only about half of it was strong. I-LtrI's motif included a 'GAC,' found in the wild-type motif, but the rest of the SELEX motif was palindromic rather than following the complete wild-type motif. Of the

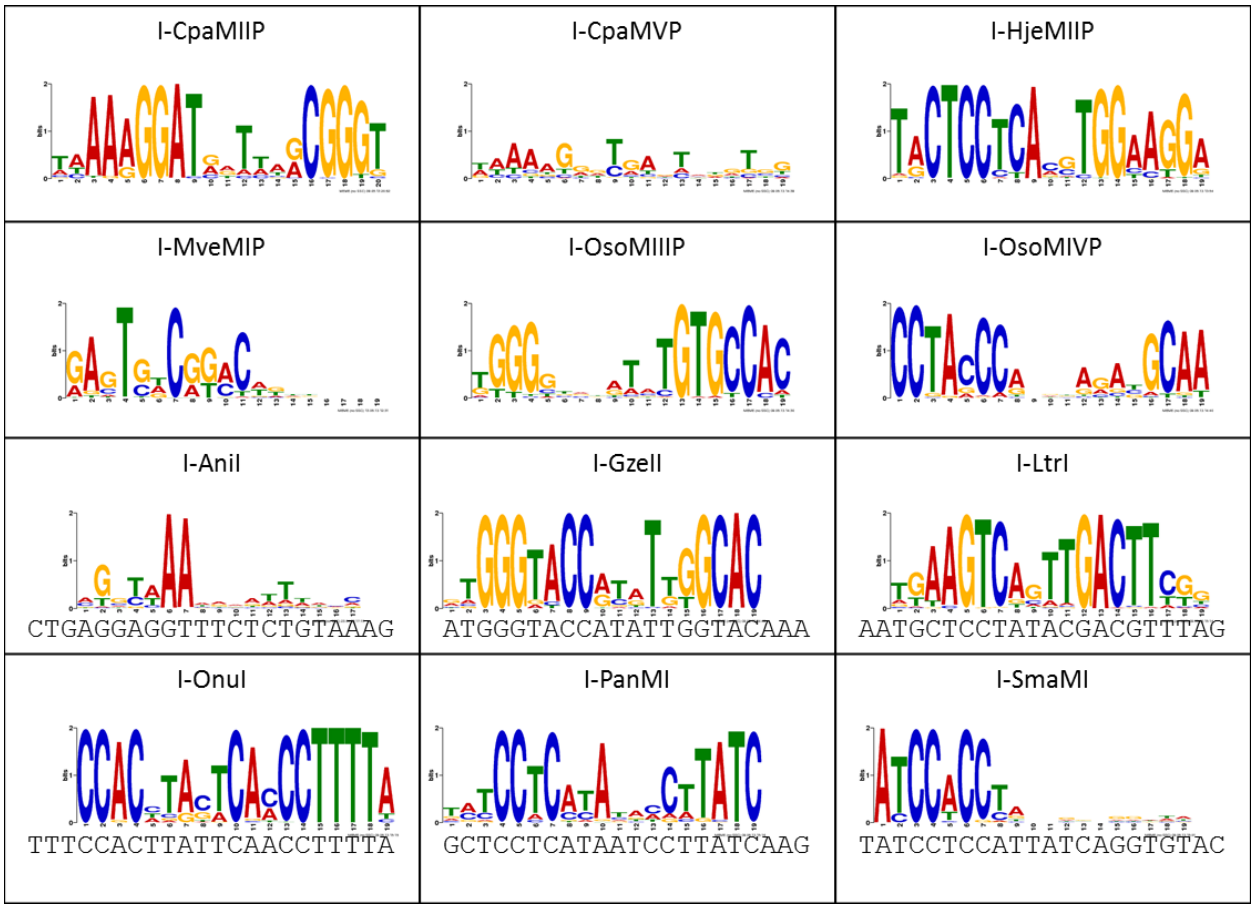


Figure 25. SELEX sequence motifs. The motifs found by analyzing the SELEX sequences for each enzyme using the expect maximization tool, MEME. The wild-type targets are shown below the motifs for the previously characterized enzymes.

remaining five uncharacterized enzymes, each produced strong motifs, consistent across the rounds analyzed. I-MveI's motif, however, was shorter than the canonical ~20 bp LHE motif. One feature consistent across many of these motifs was a lack of specificity in the central region. This feature is likely due to the lack of direct base contacts by LHEs to the central four bases[52,74,91]. The relative specificities implied by the motif for I-OnuI closely mirrored the relative specificities determined by the cleavage profiling carried out in Chapter 2 (**Figure 12c**). It should be noted that the sequence flanking the randomized region of the SELEX oligo was "TTT" 5' to the randomized sequence, and "AAA" 3' to it. Motifs that begin or end with these sequences may therefore be missing a 5' TTT or 3' AAA, as is likely the case for I-GzeII

(…AAA-3'), I-OnuI (5'-TTT...), and others. These bases were originally selected to flank the randomized region to avoid high-affinity guanine or cytosine protein-base contacts, though LHE target sequences that begin or end with these sequences occur with surprising frequency and may not have been an ideal choice in retrospect.

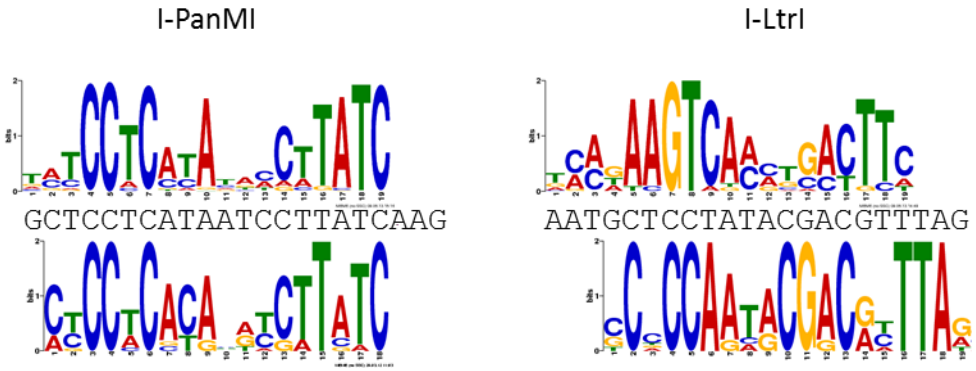


Figure 26. Comparison of SELEX motifs from different experiments. The motifs for the main experiment (top) and a preliminary, independent selection (bottom) are shown compared to their wild-type sequence. Most enzymes produced highly similar results, with differences akin to those shown for I-PanMI. Although I-LtrI's motif was very similar to its known specificity in the preliminary round of experiments, the main run failed to reproduce those results.

Analysis of a preliminary SELEX experiment run under similar conditions produced very similar results. Motifs for I-OnuI, I-PanMI, I-SmaMI, I-HjeMIIP, I-OsoMIIIP were recapitulated with similar specific preferences and stringencies; I-PanMI is shown as a representative example in **Figure 26**. One exception was that I-LtrI's motif was much more closely matched to its known sequence in the preliminary experiment (**Figure 26**). This disparity was a result of experimental error, specifically sample contamination. The yeast used for selection in this experiment were later confirmed to not express I-LtrI (described later). These results suggest that the conditions used here provide good randomization in the initial pool, and a high level of precision in SELEX.

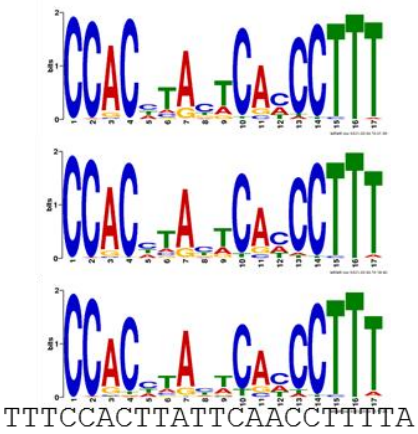


Figure 28. Comparison of SELEX motifs generated from different rounds of SELEX. Motifs generated from the sequences from round 5 (top), 4 (middle), or 3 (bottom) are shown for I-OnuI.

The number of iterations of SELEX, and the number of sequences required to generate a high-quality motif are few. Only three rounds of SELEX are required to generate a high quality motif for the LHEs presented here. This requirement is consistent with SELEX analyses by others, which found that motifs generated from three rounds of SELEX will typically better reflect the specificity profiles of the given protein[133]. Indeed, motifs from subsequent rounds have diminished abilities to show subtle differences in base preferences. For example, the motifs show an increased apparent stringency for the cytosine at position 4 of the I-OnuI motif in **Figure 28** as the number of rounds of SELEX is increased. However, the known specificity of I-OnuI (**Figure 12**) shows that, indeed, that position is less stringent compared to the preference at position 1 of the above motif. This subtle difference is therefore lost as more rounds of SELEX are performed. As such, when a motif that is more akin to the specificity profile of an enzyme is desired, fewer rounds of SELEX are needed; if only a small number of the best-bound sequences are required, more rounds of SELEX should be used. Surprisingly, only on the order of 10 sequences are required to create a good motif. The accuracy of the motif is increased only small amounts for orders of magnitude greater

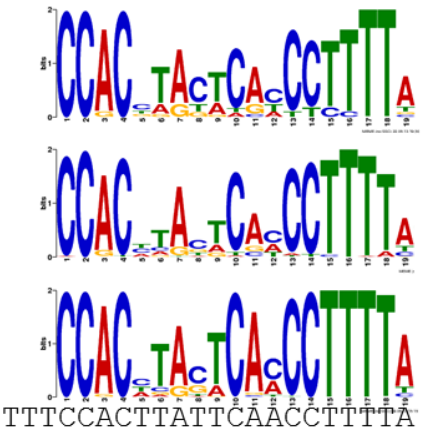


Figure 27. Comparison of SELEX motifs generated using different numbers of sequences. Motifs generated using 15 (top), 150 (middle), or 1500 (bottom) randomly selected sequences is shown for I-OnuI.

numbers of sequences (**Figure 27**). Large numbers of sequences may be required only when attempting to generate motifs that closely mimic the specificity profile of an enzyme, as would be obtained by using a more diverse set of sequences from an earlier round of SELEX (e.g. round 3). If only the few best-bound sequences are desired, three to five rounds of SELEX and low-throughput sequencing may suffice.

4.4 Validation of SELEX targets

Once a motif is generated from SELEX, it can be used to search genomic data for the wild-type LHE target site. Tools such as Motif Alignment & Search Tool (MAST, part of the MEME suite)[134] can use the motif's underlying position weight matrix generated by MEME to search genomic sequences for a best-fit match. These matrices are particularly useful if one has genomic data of insert-less homologs. Although comparing insert-less homologs to homologs with the LHE insertion can reveal the target site, sometimes the homolog is not a close enough relative, and the junctions of the insert, and therefore sequence of the target are not always clear. Although the investigator may be within a single base of the actual target, exact spacing must be maintained between the half-sites or no activity will be seen. Although the motifs generated by SELEX eliminate the spacing question, the genomic data help validate the motif, especially in the central four bp where binding discrimination is weak; these methods can therefore complement each other when homolog genomic data are available. When homolog data are not available, motifs can still be used to search the sequence of the host (with the LHE insert) by splitting the motif roughly in half. This search works because most LHEs leave the original target intact, but split on either side of the insert. Again, the full SELEX motif will aid in keeping

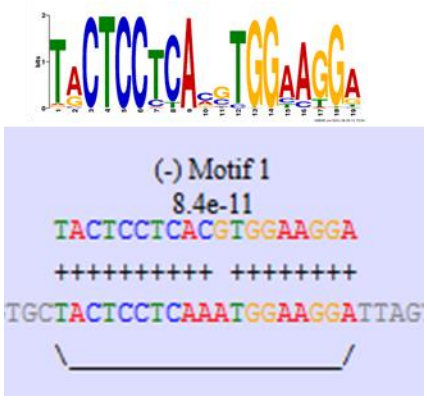


Figure 29. Alignment of I-HjeMIIP's SELEX motif to a genomic target. A number of close insert-less homologs of I-HjeMIIP's host gene were searched using the motif generated by SELEX. This was the best match; the single mismatch occurs at a position of low specificity in the motif.

the half-site spacing at the required length. This tactic of using SELEX in combination with available genomic data can therefore improve the likelihood of identifying a cleavable substrate.

Using this motif search tactic, I used the uncharacterized LHEs' motifs were used to search for genomic matches. Possible genomic matches were found for many of the putative LHEs. I-HjeMIIP matched an insert-less homolog almost exactly (**Figure 29**). The remainder of the motifs made only partial matches to genomic sequence, likely due to a lack of highly homologous insert-less host genes.

I used these sequences, motifs, and genomic matches to validate the ability of SELEX to find correct LHE substrates. Each MEME consensus sequence and two of the top ranking sequences from the expect maximization were chosen for the uncut and the cut pools (matched back to the full-length sequence). I also selected any intact possible targets identified in insert-less homologs, and targets predicted from half-sites in the original host found using the motifs as described above – or in the case of the characterized LHEs, I used the known wild-type genomic target. If the uncharacterized LHEs had a close relative, I also added the LHE homolog's target to the list. A close match to the I-CpaMIIP target was found in the *atp6* gene and was included, as well as the consensus preliminary SELEX sequence for I-LtrI. I synthesized these oligos and used them as a template for making fluorescently labeled dsDNA. The fluorescent targets were then used as binding and cleavage substrates in the flow cytometry-based assays described in

section 1.4 and 2.3. Targets which showed promise in the high throughput flow assay were validated in the in-solution cleavage and gel-based assay.

The binding and cleavage of the individually selected targets correlated well with the preliminary SELEX binding data and quality of the corresponding motifs. Those enzymes that showed the most promise in the preliminary binding assay (at least a 1-log increase in fluorescence between the initial randomized pool and rounds 3-5, **Figure 23**: I-GzeII, I-HjeMIIP, I-PanMI, I-OnuI, I-OsoMIIIP, and I-OsoMIVP) were also the enzymes that produced high-quality motifs of the expected length (**Figure 25**). These enzymes were also able to best bind and cleave targets predicted by SELEX (**Figure 30**). Those enzymes that showed less than a 1-log increase in fluorescence (I-AniI, I-CpaMIIP, I-LtrI, I-MveMIP, and I-SmaMI) showed the poorest binding and cleavage of their targets, though I-SmaMI was able to bind the predicted targets moderately well. Although I-MveMIP's cleavage was barely detectable in the flow-based assay, the gel cleavage showed that the enzyme's activity was specific, generating a well-defined band. All of the previously characterized enzymes bound and cleaved their wild-type targets as expected, with the exception of I-LtrI, supporting the hypothesis of experimental error – probably sample contamination. This was validated when the targets were cleaved in solution and run on a gel (the flow binding/cleavage for I-LtrI used the same contaminated culture as was used in the selection). In fact, the consensus sequence for I-LtrI from the preliminary SELEX was cleaved in this assay, verifying the functionality of SELEX for I-LtrI as well. Surprisingly, although I-AniI cleaved its wild-type target with high efficiency, binding was lower than many of the selected I-AniI targets. In fact, binding of the SELEX targets was at least as high as the wild-type target in some cases, indicating the success of the SELEX protocol in the strictest sense. However, which properties these targets were selected on (e.g. binding to the maturase

	Description	Binding	Cleavage	Validated	Target
I-AniI	wild-type	0.40	6.27	++	CTGAGGAGGTTTCTCTGTAAAG
	uncut motif	0.27	1.18		GGGTTAATTTATTTACC
	top uncut	0.50	1.10		TTTTCTTTATGGGGTAAATGCATTTACCGTTTTAAAA
	top uncut	0.64	1.11		TTTTTCGGTAAATTAATTTACCAGACTATAGT
	cut match motif	-0.22	1.19		TTTTGGTTAATTTTTTAACCAAA
	top cut match	-0.43	1.25		AGCAAGTATGGTTAATTCTTTAACTAGC
	top cut match	0.13	1.21		AGGGCAAGTTGGTTAATGTATTAACCTTC
I-CpaMIIP	host gene	0.05	1.14		TTAAAGGATGAATGAGTGGTAAAA
	host gene	-0.01	1.14		TTAAAGGATGAATACAGGGTGAAAT
	host gene	0.00	1.03		TTAAAGGATGATTTTAGGGTAAC
	homolog gene	0.01	1.40	-	GTAAAGGACAACCAGCGGGTACT
	best guess	0.01	1.06		TAAAAGGTTGAATAGCGGGTTAAA
	uncut motif	0.36	1.06		TAAAAGGATGATTAGCGGGTAAA
	top uncut	0.30	1.05		CAAAAGGATAGTTAGCGGGTAAA
	top uncut	0.15	1.03		TAAAAGGATAGTTAACGGGTAA
	cut match motif	0.20	1.05		TAAAAGGATATATTGCGGGTAAA
	top cut match	0.29	1.15	-	CAAAGGGATATATGGCGGGGTAA
	top cut match	0.29	1.17	-	TAAAAGGATGTATTGCGGTTAAA
	atp6 genomic	-0.02	1.05		TAAAGGACAACCAGCGGGT
I-GzeII	wild-type	1.37	10.20	+++	ATGGGTACCATATTGGTACAAA
	uncut motif	1.93	10.38	+++	GTGGGTACCATATTGGCACAAA
	top uncut	1.81	10.87		GTGGGTACCATATTGGCACTAG
	top uncut	1.56	8.76		GTGGGGACCATATTGGCACCAA
I-HjeMII	host gene	1.56	8.93	+++	TGCTACTCCTCAAATGGAAGGACTAG
	homolog gene	1.06	9.17	+++	GCTACTCCTCAAATGGAAGGA
	uncut motif	1.73	3.77		TTATACTCCTCACGTGGAAGGAAA
	top uncut	1.70	5.22	++	ATACTCCTCACGTGGAAGGATCTT
	top uncut	1.63	4.11		TTGGTACTCCTCACGTGGAAGGACATC
	cut match motif	0.38	2.79		TTTCTCCTCACGTGGAAGGGTTAA
	top cut match	1.62	3.80		TTTGCTCCTCACGTGGAAGGACTGTC
	top cut match	1.10	2.72		AATGCTCCTTACGTGGAAGGGATTGT
I-LtrI	wild-type	0.00	1.44	+++	AATGCTCCTATACGACGTTTAG
	uncut motif	0.98	1.19		TGAAGTCAGTTGACTTCGGTT
	top uncut	1.06	1.24		ACGTGAAGTCAGTTGACTTCGGAGG
	top uncut	0.97	1.32	-	GCATGAAGTCACTTGACTTTTTCGG
	cut match motif	0.91	1.10		AGAAGTCAGTTGACTTCTGATT
	top cut match	1.03	1.14		ACGTGAAGTCAGTTGACTTCGGAGGC
	old motif	-0.02	1.42	+++	TTCCCCCAATACGACGTTTAGA

Figure 30. Binding and cleavage of selected SELEX targets. Binding and cleavage is shown for the targets selected from the SELEX analysis. "Wild-type" targets are validated targets for pre-characterized enzymes; "host gene" targets are the equivalent targets, as predicted by half-site analysis for uncharacterized LHEs. "homolog gene" targets are targets which were found intact in insert-less host homologs. The "uncut motif" is the MEME-derived consensus sequence from the uncut pool analysis, and the "top uncut" are the top ranking originally selected sequences from the pool. Likewise, the "cut match motif" and "top cut match" are the corresponding consensus and top-ranked sequences from the original uncut parent sequences matched from the cut pool analysis. "Closest relative" targets are wild-type targets of the closest characterized LHE homolog. For binding, the log increase in fluorescence over a non-specific target is given; numbers above 0 indicate an increase in affinity above background. For cleavage, the Ca^{++}/Mg^{++} ratio is given (as previously described); numbers above 1 indicate cleavage. Red represents high binding or cleavage, and blue, low. Target cleavage was also validated in solution and run on a gel, and shown as no (-) or low, medium, or high levels (+, ++, +++) of cleavage *Continued on the next page*.

	Description	Binding	Cleavage	Validated	Target
I-MveMI	homolog gene	0.03	1.70		CTAATGTTTTGAGTTTCAGCCTTGCGACA
	homolog gene	-0.07	1.64		GGCTTTAGGAGTTTCGGGCATATAGTG
	homolog gene	0.07	1.83		CATAGTGTGACGGGCAGTGTGTA
	homolog gene	0.05	1.59	++	ATAGTCGAGTTACAGACTCTAATTC
	homolog gene	0.00	1.60		TCCATAGTGTGACGGGCAGTGTGTA
	homolog gene	0.11	1.67	+++	AATTCCCAGGGTGTGACGGGCGGTGTGTA
	uncut motif	0.36	1.67		TGTACGTTTGAGTGTCGGACAGTTTGTA
	top uncut	0.76	1.65	+	TGTACGTTTGAGTGTCGGACAGTTTATTGGGCCTATTT
	top uncut	0.73	1.73		TGTACGTTTGAGTGTCGGACAGTTGATAGAAAATTAGC
	cut match motif	0.56	1.63		TGTACGTTTGAGTGTCGTACCGCATAGC
	top cut match	0.58	1.75		TGTACGTTTGAGTGTCGTACCGTATGCCAAGCTTGGGG
	top cut match	0.39	1.78	-	TGTACGTTTGAGTGTCGTATAGTTTATCCTTTGTGGTG
I-OnuI	wild-type	1.73	6.58	++	TTTCCACTTATTCAACCTTTTA
	uncut motif	2.02	3.96	+	TTTCCACCTACTCACCCTTTTA
	top uncut	1.80	3.43		ATTCCACCTACTCACCCTTTTA
	top uncut	1.27	2.91		GTACCACCTACTCACCCTTTTA
	cut match motif	2.02	4.77		TTTCCACATACTCACCCTTTTAT
	top cut match	1.91	2.44		TTTCCACATGCTCACCCTTTTAA
I-OsoMIII	host gene	0.75	2.49		TATTGTGGGCTATAGACGTGCCACATAT
	homolog gene	0.77	2.55	++	TATTGTGGGCTATAGACGTGCCACATAT
	uncut motif	0.64	2.29		TTTTGGGGTATATATGTGCCACAAA
	top uncut	0.62	2.81	++	ACAATGGGGTATATATGTGCCACAA
	top uncut	0.64	2.62	++	GCTCTGGGGTATATATGTGCCACAAA
	closest relative	0.24	1.52		TGAACTGGGCTATAGACGTGCT
I-OsoMIV	host gene	0.79	3.27	++	ATGCCTAGACAAAGAGATGCAAAAA
	homolog gene	0.69	6.45	++	ATGCCTACACAAAGAGATGCAAAAA
	homolog gene	0.76	5.90		TTACCTAGACAAAGAGATGCAAAAA
	uncut motif	1.43	1.24	-	TTTCCTACCCAAGTAGATGCAAAAA
	top uncut	1.36	1.20		TTTCCTACCCAAATAGACGCAAAAG
	top uncut	1.29	1.51	-	CACCCTACCCAGGGAGATGCAAAAA
	cut match motif	1.39	1.18		TTTCCTACCCAATTAGATGCAAAAA
	top cut match	1.23	1.20		TTTCCTACCCAGGTAAGTGCAAAAA
	top cut match	1.10	1.23		TTTCCTACCCAAGTAAAGACAAAAA
	closest relative	0.55	1.28		TACACCTGATAAAGGAGGGTAAATAGTT
I-PanMI	wild-type	1.83	13.29	+++	GCTCCTCATAATCCTTATCAAG
	uncut motif	2.09	20.81	+++	TATCCTCATATACCTTATCAAA
	top uncut	1.62	15.19		GCTCCTCATATACCTTATCATA
	cut match motif	2.12	19.94		ATCCTCATATACCTTATCAAAA
	top cut match	1.73	20.79		GTCCTCATATAACATATCAAAA
	top cut match	1.77	12.52		TCCCCCATATACCTTATCAAAA
I-SmaMI	wild-type	1.51	8.39	+++	TATCCTCCATTATCAGGTGTAC
	uncut motif	1.37	1.25	-	TATCCTCCCTTAGGTGGATAAA
	top uncut	1.35	1.26		TATCCACCCATAGGTGGATACA
	top uncut	1.34	1.19		TATCCCCCCATAGGTGGATAAA
	cut match motif	0.11	1.13		ATCCACCTACCCTTGAAAC
	top cut match	1.25	1.18		TTATCCACCTAATCATGTAAC
	top cut match	0.34	1.21		ATCCACCTACATATGTTCT

Figure 30. ***Continued from previous page****.*

domain, binding at only a few positions, etc.) remain unknown. The apparent slight cleavage activity of I-CapMIIP against one target (from a homolog host gene), and of I-MveMIP against a number of varied putative targets, requires further validation. In all, three putative nucleases – I-HjeMIIP, I-OsoMIIIP, and I-OsoMIVP – were validated as functional, novel LHEs using targets identified by SELEX.

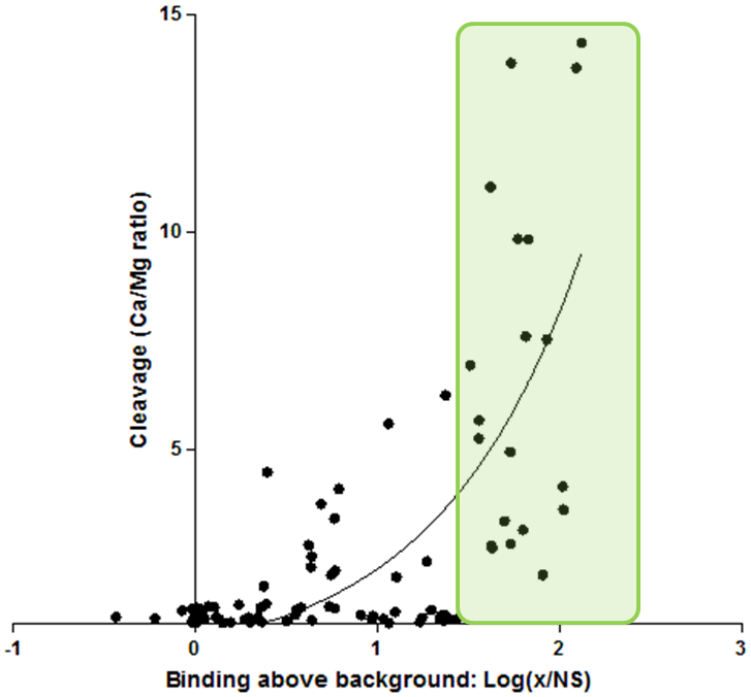


Figure 31. Binding versus cleavage of selected SELEX targets. The binding and cleavage values from **Error! Reference source not found.** are plotted with a semi-log linear regression line plotted for reference. The green box indicates targets with high binding (more than a 1.5-log increase in fluorescence over the non-specific target).

The correlation between binding and cleavage continued in the set of individually selected targets. All targets that demonstrated at least 1.5-log increases in fluorescence above background in the binding assay also demonstrated high levels of cleavage activity (Ca^{++}/Mg^{++} ratio > 2) against the same targets (**Figure 31**). These data continue to support the hypothesis that the best-bound targets will yield cleavable substrates, and that preliminary binding data are a strong predictor of SELEX success.

4.5 Summary

SELEX synergizes with yeast surface display to yield a powerful new tool that can be easily integrated into existing YSD experimental designs. Vectors made for yeast surface display can be used to characterize proteins in high throughput, and subsequently shunted into SELEX or directed evolution pipelines and vice versa[73,74]. For example, a panel of putative HEs can be assayed for proper folding, its targets determined by SELEX, its binding and cleavage properties interrogated in detail, all in a multi-well, high throughput manner[73,135]; enzyme specificities can then be altered by directed evolution using the same platform. Our approach to SELEX also benefits from the ability to test oligo binding in high throughput using yeast surface display and flow cytometry. SELEX conditions can be optimized and tuned to the investigator's precise needs, and can serve as a strong indication of the protocol's success. Selection conditions can subsequently be modulated to yield more diverse or narrow target pools depending on whether the investigator desires a binding profile, or simply a few best-bound targets[136]. Lastly, SELEX using yeast surface display is an improvement upon traditional SELEX insofar as it allows quick, easy, and inexpensive expression of protein that does not need further purification or modification. It also removes the explicit need for expensive consumables such as magnetic beads or antibodies. This method should be regarded as a practical alternative to traditional SELEX, as well as an additional tool for the yeast surface display toolbox.

SELEX can be used in a variety of instances. To name a few, it can be used to obtain targets of DNA-binding proteins (not only LHEs) where no prediction methodology exists; where engineering methods yield enzymes with unpredictable specificity; and where their target prediction rely on missing, partial, or improperly annotated sequence data. In fact, SELEX would likely perform more reliably when target cleavage is not required since using binding as a proxy

for cleavage may not always work. YSD-SELEX can also be leveraged to take advantage of existing SELEX modifications. For example, the protocol can be modified to use genomic sequence as the input pool[137], providing as output a set of best-bound sequences *in that genome* for the protein. This SELEX output would be valuable in determining any off-targets, which may be of keen interest in genome engineering[138].

Although SELEX can be used to determine some LHE target sites, our results also indicate its effectiveness may be reduced in some instances. SELEX was able to reveal only half of I-SmaMI's target, and the failure to focus the pool on I-AniI's known target site suggests that the SELEX conditions may need to be individually optimized for some LHEs. Even with optimization, SELEX may not be able to find a cleavable substrate for some LHEs with semi-repetitive target sites (as with I-SmaMI), or with DNA-binding concentrated in only one part of the enzyme (as with I-AniI[77]). It may also be possible that the secondary nucleic acid binding (maturase) domain of I-AniI interfered with the binding selection. Since multiple or alternative nucleic acid binding domains may obfuscate SELEX results, this protocol should be *carefully* used with such proteins. Ideally, any such domains should be separated or eliminated before testing. Still, the effect of these factors on the success of SELEX remains speculative and warrants further investigation.

It is still possible that SELEX may be able to generate high quality motifs, even for proteins with less-than-ideal properties such as those described above. Optimizing selection conditions for each individual protein (as opposed to using uniform conditions as was done for this set of proteins) may increase the signal/noise, resulting in high quality binding motifs. I-OnuI was primarily used for the binding selection condition (salt and DNA concentration)

optimization since this set of enzymes was found based on homology to that protein. However, I-SmaMI showed high levels of binding to the initial randomized pool (**Figure 23**), indicating that the initial selection conditions were not stringent enough. These suboptimal conditions may have led to poor selection and subsequently, the incomplete motif. Likewise, I-AniI is known to have substantially lower binding affinity compared to the I-OnuI subfamily, suggesting that its optimal selection conditions may have been different. It may also be necessary to alter the number of rounds of SELEX for the lower affinity enzymes, or fewer rounds for enzymes with skewed half-site affinity where one half-site would quickly dominate the selection. In any case, the best indication of a successful binding selection remains a high increase in affinity, as determined via yeast surface display flow-based binding assay (at least ~one log). The ability to quickly determine the binding of a protein to its selected DNA pool during condition optimization and analysis is one of the most useful aspects of using YSD with SELEX.

Cleavage selection may be an important final step after binding selection. Cleavage selection after binding selection would rectify any binding/cleavage discrepancy of SELEX. Although our attempts to incorporate a cleavage selection step were met with limited success, other established protocols exist[92]. Furthermore, since cleavage selection is unable to select cleavable substrate from a fully randomized pool as SELEX is, pre-selection using SELEX may empower cleavage selection by reducing the need for knowledge of the target site beforehand. Thus, these protocols would likely complement each other and be used to obtain otherwise unobtainable nuclease target specificities.

Lastly, SELEX may prove to be a valuable method for obtaining rough specificity profiles for LHEs as it has been for other proteins[129,135,136]. Specificity profiles are invaluable in

assisting engineering efforts. First, specificity profiles can guide investigators to the most accessible areas of the gene. Second, when re-shaping the specificity of LHEs it is important to know which base substitutions are tolerated and which require mutagenesis to alter. Third, more specificity profiles for LHEs will diversify the dataset used to train programs aimed at understanding and predicting modifications that govern protein-base specificities. Last, specificity profiles are important in the application phase to determine possible alternative targets for the LHE. Aside from aiding engineering, more specificity profiles may elucidate differences between LHEs from RNA host genes compared to protein-coding host genes. Because LHEs inserted in protein-coding regions of their host gene tend to have regions of reduced specificity[52,139], it would be valuable see if we could find LHEs of increased specificity in genes that code for RNA products. Since the profile predicted by SELEX closely mimicked the previously described profile for I-OnuI, it is likely that these profiles may be useful, especially with protocol modifications to that end. Using SELEX for preliminary characterization becomes even more inviting given that these profiles are obtained during initial enzyme characterization, and in a highly parallel fashion.

My goal is for yeast surface display and SELEX to further facilitate LHE-based genome engineering efforts. I envision the integrated methodologies and tools described here being used to expand our collection of, and help redesign well-characterized, highly active and widely diverse group of LHEs. When targeting a given locus, this library will be queried for the best possible starting enzyme with a native target site closest to a sequence in that locus. The investigator will then make minimal modifications to the enzyme, guided by available specificity data, altering the specificity toward the new target site. New specificity data can be generated, and the enzyme can be used to carry out the desired modifications. In all, the framework

provided here can advance many areas of research and biotechnology by assisting our genome engineering efforts.

4.6 Methods

SELEX library preparation and amplification

We ordered SELEX single stranded randomized oligo pool template with flanking SELEX primer sites (underlined), from integrated DNA technologies: CAG GGA TCC ATG CAC TGT ACG TTT (N30) AAA CCA CTT GAC TGC GGA TCC T, along with forward and reverse primers (unlabeled primer for selection experiments, A647 and biotin labeled for flow cytometry experiments). We created a dsDNA library from this oligo by running a single round of PCR with the reverse primer using Platinum Taq DNA polymerase High Fidelity (Invitrogen).

After each round of selection (below) the selected oligo was amplified using 20 rounds of PCR. A secondary PCR of 2 rounds, seeded by the first was then used to ensure each oligo is double stranded and properly paired for the next round of selection. Fluorescent, biotinylated oligo was also made and purified using the protocols described in chapter 2.

SELEX binding selection and analysis

Three million induced yeast (prepared as in Chapter 2) per protein per round were washed twice (2000 x g for 1 minute) with 200 µL bind and wash buffer (BWB: 0.15 M KCl, 0.002 M $CaCl_2$, 0.01 M NaCl, 0.01 M HEPES, 0.005 M L-Gluatmic Acid Potassium Salt Monohydrate, 0.05% BSA, adjusted to pH 7.5 with KOH), and resuspend to a final concentration of 500,000 yeast/µL. Each round of selection was carried out in a 96-well plate with the

following mixture: 5 µL SELEX0 dsDNA with 89 µL of BWB and 6 µL (3 million) yeast bearing each protein. The plate was sealed and incubated for 30 minutes at room temperature with agitation. Each sample was washed 6 times in 150 µL BWB. After the wash steps, the samples were resuspended in 40 µL 10% buffer EB, the plate was sealed, and the oligo was released by heating the protein past its melting temperature (70° C) for 10 minutes. Note: it may be necessary to release LHE (and extract the oligo with phenol chloroform if necessary) for enzymes that melt above ~70 °C. After release, the yeast were *immediately* spun down, and the supernatant was quickly transferred to a new plate for storage using a multichannel pipette. The oligo was amplified as described above and used to seed the next round of selection.

For binding analysis, A647-labeled oligo was used in place of unlabeled oligo. Following the 6 x washes, the samples were analyzed on a flow cytometer as in Chapter 2.

Once all rounds of selection were complete, the oligos from each round of SELEX to be analyzed and for each enzyme were amplified using the oligos which also had the next generation sequencing forward primer, a 4-base barcode for that round and enzyme, and a unique (randomized) 7-base barcode. Samples to be directly sequenced were then amplified with primers that also had the next generation sequencing reverse primer; samples to undergo cleavage selection will have their next generation sequencing primer added separately.

Cleavage selection

Singly-biotinylated oligo was made for each pool of oligos to be selected against, following the protocol for generating labeled oligo. Blunt adapters were made by annealing the sense and antisense next generation sequencing reverse primer. Oligos were cleaved in 50 µL

using the "in solution" yeast cleavage protocol outlined in Chapter 2, and the supernatant containing the oligo was collected after heating the samples to 70 °C, as above. The oligo was captured from 25 µL of the supernatant by incubation for 30 minutes with 15 µL settled streptavidin-agarose beads in buffer EB (Qiagen). In order to blunt the overhangs left by LHE cleavage, the beads were washed 2x in buffer EB (spinning at 500 x g for 1 minute) and resuspended in Klenow reaction buffer with enzyme and incubated at 37 °C for 30 minutes according to the manufacturer's instructions (New England Biolabs). The beads were washed twice more in buffer EB, resuspended in ligation mix with the annealed adapters, and incubated for 1 hour at room temperature according to the manufacturer's instructions (New England Biolabs); only cleaved oligo should have an exposed phosphate to ligate to. Finally, the beads were washed twice more in buffer EB, residual ligase was heat-killed, and the beads were resuspended in PCR mix. The oligo was amplified using the same protocol for oligo amplification as during the binding selection, and sequenced as described below. Cleaved and selected oligo should be missing part of the 30N randomized sequence and the SELEX reverse primer, but should have the next generation sequencing reverse primer.

Sequencing & analysis

Once the forward and reverse next generation sequencing adapters were added to the pools (at the end of the binding selection and during the cleavage selection), the oligos were run on a 3% agarose gel, purified, and sent for sequencing per the providers instructions (Edge Bio).

The sequences returned from Edge Bio were sorted into three types of collections for each round/enzyme for analysis. Each sequence was parsed for the enzyme/round 4-base barcode, 7-base unique barcode, and the randomized 30N region (±2 bp) or the truncated 30N (7-

23bp); this information was stored in an SQL database. Collections of uncut 30N sequences for each round/enzyme were output to their own FASTA-formatted files (uncut). Cut sequences were also output to their own files (cut). Cut sequences were also matched back to their parent uncut form, and the corresponding parents were output to FASTA files as well (cut/matched). Each round/enzyme analyzed therefore had three associated sequence collections.

Each collection of sequences was then analyzed by expect maximization using MEME in search of a sequence motif. The following parameters were used for full-length randomized regions (uncut and cut/matched) with MEME: -dna -mod zoops -noendgaps -minw 19 -maxw 22 -nmotifs 1 -maxsites *x* -minsites *y* -revcomp. For cut sequences, -nmotifs was changed to 2, -revcomp was eliminated, -minw was set to 9, and maxw to 11. Here, *x* was the lesser of 1500 or the total number of sequences, and *y* was the lesser of *x* or one third of the total number of sequences for uncut sequences, or one sixth for cut sequences (since there are twice as many motifs to find). The resulting motifs for round 5 were shown in **Figure 25**.

The motifs found by MEME for uncharacterized LHEs were also used to search LHE host genes and their homologs for the original LHE target sites. Close homologs of the host gene identified by nucleotide BLAST searching, along with the original LHE-inserted host were compiled into a single FASTA file. These sequences were passed to MAST online tool (http://meme.nbcr.net/meme/cgi-bin/mast.cgi) and the default parameters were used to find possible matches to the motifs. Half-motifs were also used in the MAST searches since target sites in the host gene are often split on either side of the LHE insertion. These hits were used while selecting targets to validate the SELEX results (**Figure 30**).

References

1. Bosma, G. C., Custer, R. P. & Bosma, M. J. A severe combined immunodeficiency mutation in the mouse. *Nature* **301,** 527–530 (1983).

2. Fuchs, H. *et al.* Mouse phenotyping. *Methods San Diego Calif* **53,** 120–135 (2011).

3. Guan, C., Ye, C., Yang, X. & Gao, J. A review of current large-scale mouse knockout efforts. *Genes. New York N 2000* **48,** 73–85 (2010).

4. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460,** 894–898 (2009).

5. Johnson, I. S. Human insulin from recombinant DNA technology. *Science* **219,** 632–637 (1983).

6. Shen, Y. *et al.* An efficient xylose-fermenting recombinant Saccharomyces cerevisiae strain obtained through adaptive evolution and its global transcription profile. *Appl. Microbiol. Biotechnol.* **96,** 1079–1091 (2012).

7. Baek, S.-H., Kim, S., Lee, K., Lee, J.-K. & Hahn, J.-S. Cellulosic ethanol production by combination of cellulase-displaying yeast cells. *Enzyme Microb. Technol.* **51,** 366–372 (2012).

8. Zhou, H., Cheng, J.-S., Wang, B. L., Fink, G. R. & Stephanopoulos, G. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by Saccharomyces cerevisiae. *Metab. Eng.* **14,** 611–622 (2012).

9. Khattab, S. M. R., Saimura, M. & Kodaki, T. Boost in bioethanol production using recombinant Saccharomyces cerevisiae with mutated strictly NADPH-dependent xylose reductase and NADP(+)-dependent xylitol dehydrogenase. *J. Biotechnol.* (2013). doi:10.1016/j.jbiotec.2013.03.009

10. Hinnen, A., Hicks, J. B. & Fink, G. R. Transformation of yeast. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 1929–1933 (1978).

11. Orr-Weaver, T. L., Szostak, J. W. & Rothstein, R. J. Yeast transformation: a model system for the study of recombination. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 6354–6358 (1981).

12. Sorrell, D. A. & Kolb, A. F. Targeted modification of mammalian genomes. *Biotechnol. Adv.* **23,** 431–469 (2005).

13. Pâques, F. & Duchateau, P. Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Curr. Gene Ther.* **7,** 49–66 (2007).

14. Rogers, S. Gene therapy: a potentially invaluable aid to medicine and mankind. *Res. Commun. Chem. Pathol. Pharmacol.* **2,** 587–600 (1971).

15. Munyon, W., Kraiselburd, E., Davis, D. & Mann, J. Transfer of thymidine kinase to thymidine kinaseless L cells by infection with ultraviolet-irradiated herpes simplex virus. *J. Virol.* **7,** 813–820 (1971).

16. Polmar, S. H., Wetzler, E. M., Stern, R. C. & Hirschhorn, R. Restoration of in-vitro lymphocyte responses with exogenous adenosine deaminase in a patient with severe combined immunodeficiency. *Lancet* **2,** 743–746 (1975).

17. Friedman, R. L. Expression of human adenosine deaminase using a transmissable murine retrovirus vector system. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 703 (1985).

18. Miller, A. D. Retrovirus packaging cells. *Hum. Gene Ther.* **1,** 5–14 (1990).

19. Miyake, K. & Shimada, T. Gene transfer into non-dividing cells by a lentiviral vector. *Uirusu* **47,** 213–219 (1997).

20. Blaese, R. M. *et al.* T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years. *Science* **270,** 475–480 (1995).

21. Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.* **118,** 3132–3142 (2008).

22. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.* **118,** 3143–3150 (2008).

23. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302,** 415–419 (2003).

24. Akopian, A. & Marshall Stark, W. Site-specific DNA recombinases as instruments for genomic surgery. *Adv. Genet.* **55,** 1–23 (2005).

25. Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L. & Barbas, C. F., 3rd. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res.* **41,** 3937–3946 (2013).

26. Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F., 3rd. Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res.* **40,** 11163–11172 (2012).

27. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* **14,** 8096–106 (1994).

28. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79,** 181–211 (2010).

29. Bennardo, N., Cheng, A., Huang, N. & Stark, J. M. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *Plos Genet.* **4,** e1000110 (2008).

30. Shrivastav, M., De Haro, L. P. & Nickoloff, J. A. Regulation of DNA double-strand break repair pathway choice. *Cell Res.* **18,** 134–147 (2008).

31. McConnell Smith, A. *et al.* Generation of a nicking enzyme that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 5099–5104 (2009).

32. Kim, E. *et al.* Precision genome engineering with programmable DNA-nicking enzymes. *Genome Res.* **22,** 1327–1333 (2012).

33. Liu, J. & Stormo, G. D. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24,** 1850–1857 (2008).

34. Sera, T. & Uranga, C. Rational Design of Artificial Zinc-Finger Proteins Using a Nondegenerate Recognition Code Table. *Biochemistry (Mosc.)* **41,** 7074–7081 (2002).

35. Cornu, T. I. & Cathomen, T. Quantification of zinc finger nuclease-associated toxicity. *Methods Mol. Biol. Clifton Nj* **649,** 237–245 (2010).

36. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.* **29,** 816–823 (2011).

37. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods* **8,** 765–770 (2011).

38. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkr218

39. Moore, F. E. *et al.* Improved somatic mutagenesis in zebrafish using transcription activator-like effector nucleases (TALENs). *Plos One* **7,** e37877 (2012).

40. Wilhelm, M. & Wilhelm, F. X. Reverse transcription of retroviruses and LTR retrotransposons. *Cell. Mol. Life Sci. Cmls* **58,** 1246–1262 (2001).

41. Basu, V. P. *et al.* Strand transfer events during HIV-1 reverse transcription. *Virus Res.* **134,** 19–38 (2008).

42. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11,** 181–190 (2010).

43. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).

44. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31,** 233–239 (2013).

45. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).

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

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

48. Choulika, A., Perrin, A., Dujon, B. & Nicolas, J. F. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol Cell Biol* **15,** 1968–73 (1995).

49. Puchta, H., Dujon, B. & Hohn, B. Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 5055–5060 (1996).

50. Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441,** 656–9 (2006).

51. Ashworth, J. *et al.* Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38,** 5601 –5608 (2010).

52. Takeuchi, R. *et al.* Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 13077–13082 (2011).

53. Antunes, M. S., Smith, J. J., Jantz, D. & Medford, J. I. Targeted DNA excision in Arabidopsis by a re-engineered homing endonuclease. *Bmc Biotechnol.* **12,** 86 (2012).

54. Muñoz, I. G. *et al.* Molecular basis of engineered meganuclease targeting of the endogenous human RAG1 locus. *Nucleic Acids Res.* **39,** 729–743 (2011).

55. Arnould, S. *et al.* Engineered I-CreI Derivatives Cleaving Sequences from the Human XPC Gene can Induce Highly Efficient Gene Correction in Mammalian Cells. *J Mol Biol* **371,** 49–65 (2007).

56. Redondo, P. *et al.* Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature* **456,** 107–111 (2008).

57. Grizot, S. *et al.* Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. *Nucleic Acids Res.* **38,** 2006–2018 (2010).

58. Joshi, R. *et al.* Evolution of I-SceI homing endonucleases with increased DNA recognition site specificity. *J. Mol. Biol.* **405,** 185–200 (2011).

59. Chames, P. *et al.* In vivo selection of engineered homing endonucleases using double-strand break induced homologous recombination. *Nucleic Acids Res* **33,** e178 (2005).

60. Li, H., Pellenz, S., Ulge, U., Stoddard, B. L. & Monnat, R. J., Jr. Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. *Nucleic Acids Res.* **37,** 1650–1662 (2009).

61. Smith, J. *et al.* A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res.* **34,** e149 (2006).

62. Li, H., Pellenz, S., Ulge, U., Stoddard, B. L. & Monnat, R. J. Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. *Nucleic Acids Res.* **37,** 1650–1662 (2009).

63. Siggers, T. W., Silkov, A. & Honig, B. Structural alignment of protein--DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.* **345,** 1027–1045 (2005).

64. Baxter, S. *et al.* Engineering domain fusion chimeras from I-OnuI family LAGLIDADG homing endonucleases. *Nucleic Acids Res.* **40,** 7985–8000 (2012).

65. Sussman, D. *et al.* Isolation and Characterization of New Homing Endonuclease Specificities at Individual Target Site Positions. *J. Mol. Biol.* **342,** 31–41 (2004).

66. Seligman, L. M. *et al.* Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic Acids Res.* **30,** 3870–3879 (2002).

67. Gimble, F. S., Moure, C. M. & Posey, K. L. Assessing the Plasticity of DNA Target Site Recognition of the PI-SceI Homing Endonuclease Using a Bacterial Two-hybrid Selection System. *J. Mol. Biol.* **334,** 993–1008 (2003).

68. Doyon, J. B., Pattanayak, V., Meyer, C. B. & Liu, D. R. Directed Evolution and Substrate Specificity Profile of Homing Endonuclease I-SceI. *J. Am. Chem. Soc.* **128,** 2477–2484 (2006).

69. Chen, Z., Wen, F., Sun, N. & Zhao, H. Directed evolution of homing endonuclease I-SceI with altered sequence specificity. *Protein Eng. Des. Sel.* **22,** 249–256 (2009).

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

71. Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 38–41 (2007).

72. Volná, P. *et al.* Flow cytometric analysis of DNA binding and cleavage by cell surface-displayed homing endonucleases. *Nucleic Acids Res.* **35,** 2748–2758 (2007).

73. Jarjour, J. *et al.* High-resolution profiling of homing endonuclease binding and catalytic specificity using yeast surface display. *Nucleic Acids Res.* **37,** 6871–6880 (2009).

74. Jacoby, K. *et al.* Expanding LAGLIDADG endonuclease scaffold diversity by rapidly surveying evolutionary sequence space. *Nucleic Acids Res.* **40,** 4954–4964 (2012).

75. Watzele, M., Klis, F. & Tanner, W. Purification and characterization of the inducible a agglutinin of Saccharomyces cerevisiae. *Embo J.* **7,** 1483–1488 (1988).

76. Shusta, E. V., Kieke, M. C., Parke, E., Kranz, D. M. & Wittrup, K. D. Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. *J. Mol. Biol.* **292,** 949–956 (1999).

77. Thyme, S. B. *et al.* Exploitation of binding energy for catalysis and design. *Nature* **461,** 1300–1304 (2009).

78. Argast, G. M., Stephens, K. M., Emond, M. J. & Monnat Jr, R. J. I-PpoI and I-CreI homing site sequence degeneracy determined by random mutagenesis and sequential in vitro enrichment. *J. Mol. Biol.* **280,** 345–353 (1998).

79. Perrin, A., Buckle, M. & Dujon, B. Asymmetrical recognition and activity of the I-SceI endonuclease on its site and on intron-exon junctions. *Embo J* **12,** 2939–47 (1993).

80. Jurica, M. S., Monnat, R. J. & Stoddard, B. L. DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. *Mol Cell* **2,** 469–76 (1998).

81. Silva, G. H., Dalgaard, J. Z., Belfort, M. & Van Roey, P. Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. *J Mol Biol* **286,** 1123–36 (1999).

82. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat. Biotechnol.* **26,** 702–708 (2008).

83. Grizot, S. *et al.* Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res.* **37,** 5405–5419 (2009).

84. Ho, Y., Kim, S. J. & Waring, R. B. A protein encoded by a group I intron in Aspergillus nidulans directly assists RNA splicing and is a DNA endonuclease. *Proc Natl Acad Sci U S* **94,** 8994–9 (1997).

85. Lucas, P., Otis, C., Mercier, J. P., Turmel, M. & Lemieux, C. Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. *Nucleic Acids Res* **29,** 960–9 (2001).

86. Sethuraman, J., Majer, A., Friedrich, N. C., Edgell, D. R. & Hausner, G. Genes within genes: multiple LAGLIDADG homing endonucleases target the ribosomal protein S3 gene

encoded within an rnl group I intron of Ophiostoma and related taxa. *Mol. Biol. Evol.* **26,** 2299–2315 (2009).

87. Pepper, L. R., Cho, Y. K., Boder, E. T. & Shusta, E. V. A decade of yeast surface display technology: Where are we now? *Comb. Chem. High Throughput Screen.* **11,** 127–134 (2008).

88. Jiang, W. & Boder, E. T. High-throughput engineering and analysis of peptide binding to class II MHC. *Proc Natl Acad Sci U S A* **107,** 13258–13263 (2010).

89. Vembar, S. S. & Brodsky, J. L. One step at a time: endoplasmic reticulum-associated degradation. *Nat. Rev. Mol. Cell Biol.* **9,** 944–957 (2008).

90. Wen, F., Esteban, O. & Zhao, H. Rapid identification of CD4+ T-cell epitopes using yeast displaying pathogen-derived peptide library. *J. Immunol. Methods* **336,** 37–44 (2008).

91. Bolduc, J. M. *et al.* Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev* **17,** 2875–88 (2003).

92. Scalley-Kim, M., McConnell-Smith, A. & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372,** 1305–1319 (2007).

93. Takeuchi, R., Certo, M., Caprara, M. G., Scharenberg, A. M. & Stoddard, B. L. Optimization of in vivo activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation. *Nucleic Acids Res.* **37,** 877–890 (2009).

94. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).

95. Mullineux, S.-T., Costa, M., Bassi, G. S., Michel, F. & Hausner, G. A group II intron encodes a functional LAGLIDADG homing endonuclease and self-splices under moderate temperature and ionic conditions. *Rna New York N* **16,** 1818–1831 (2010).

96. Moure, C. M., Gimble, F. S. & Quiocho, F. A. The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. *J Mol Biol* **334,** 685–95 (2003).

97. Roberts, R. J. *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* **31,** 1805–12 (2003).

98. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

99. Otwinowski, Z. & Minor, W. in *Macromol. Crystallogr. Part* **276,** 307–326 (Academic Press, 1997).

100. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).

101. Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2184–2195 (2004).

102. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38,** D211–222 (2010).

103. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma. Oxf. Engl.* **25,** 1189–1191 (2009).

104. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinforma. Oxf. Engl.* **24,** 333–340 (2008).

105. Dickson, R. J., Wahl, L. M., Fernandes, A. D. & Gloor, G. B. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *Plos One* **5,** e11082 (2010).

106. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).

107. Winn, M. D., Murshudov, G. N. & Papiz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374,** 300–321 (2003).

108. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).

109. Certo, M. T. *et al.* Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. *Nat. Methods* **9,** 973–975 (2012).

110. Certo, M. T. *et al.* Tracking genome engineering outcome at individual DNA breakpoints. *Nat. Methods* **8,** 671–676 (2011).

111. Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat. Biotechnol.* **26,** 695–701 (2008).

112. Beumer, K. J. *et al.* Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 19821–19826 (2008).

113. Geurts, A. M. *et al.* Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* **325,** 433 (2009).

114. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26,** 808–816 (2008).

115. Bennardo, N., Gunn, A., Cheng, A., Hasty, P. & Stark, J. M. Limiting the persistence of a chromosome break diminishes its mutagenic potential. *Plos Genet.* **5,** e1000683 (2009).

116. Youdim, M. B. H., Edmondson, D. & Tipton, K. F. The therapeutic potential of monoamine oxidase inhibitors. *Nat. Rev. Neurosci.* **7,** 295–309 (2006).

117. Deftereos, S. N. & Andronis, C. A. Discordant effects of rasagiline doses in Parkinson disease. *Nat. Rev. Neurol.* **6,** 1p following 410 (2010).

118. Hauser, R. A. Early pharmacologic treatment in Parkinson's disease. *Am. J. Manag. Care* **16 Suppl Implications,** S100–107 (2010).

119. Olanow, C. W. *et al.* A double-blind, delayed-start trial of rasagiline in Parkinson's disease. *N. Engl. J. Med.* **361,** 1268–1278 (2009).

120. Sampaio, C. & Ferreira, J. J. Parkinson disease: ADAGIO trial hints that rasagiline slows disease progression. *Nat. Rev. Neurol.* **6,** 126–128 (2010).

121. Youdim, M. B. H. Rasagiline in Parkinson's disease. *N. Engl. J. Med.* **362,** 657–658; author reply 658–659 (2010).

122. Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13,** 2633–2638 (1999).

123. Metzger, M. J., McConnell-Smith, A., Stoddard, B. L. & Miller, A. D. Single-strand nicks induce homologous recombination with less toxicity than double-strand breaks using an AAV vector template. *Nucleic Acids Res.* **39,** 926–935 (2011).

124. Szymczak, A. L. *et al.* Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat. Biotechnol.* **22,** 589–594 (2004).

125. Heath, P. J., Stephens, K. M., Monnat, R. J. & Stoddard, B. L. The structure of I-CreI, a group I intron-encoded homing endonuclease. *Nat Struct Biol* **4,** 468–76 (1997).

126. Duan, X., Gimble, F. S. & Quiocho, F. A. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell* **89,** 555–64 (1997).

127. Jurica, M. S. & Stoddard, B. L. Homing endonucleases: structure, function and evolution. *Cell Mol Life Sci* **55,** 1304–26 (1999).

128. Abelson, J. Directed evolution of nucleic acids by independent replication and selection. *Science* **249,** 488–489 (1990).

129. Djordjevic, M. SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomol. Eng.* **24,** 179–189 (2007).

130. Piasecki, S. K., Hall, B. & Ellington, A. D. Nucleic acid pool preparation and characterization. *Methods Mol. Biol. Clifton Nj* **535,** 3–18 (2009).

131. Irvine, D., Tuerk, C. & Gold, L. SELEXION. Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *J. Mol. Biol.* **222,** 739–761 (1991).

132. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* **25,** 778–785 (2007).

133. Schütze, T. *et al.* Probing the SELEX process with next-generation sequencing. *Plos One* **6,** e29604 (2011).

134. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).

135. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20,** 861–873 (2010).

136. Roulet, E. *et al.* High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotech* **20,** 831–835 (2002).

137. Lorenz, C., von Pelchrzim, F. & Schroeder, R. Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nat Protoc* **1,** 2204–12 (2006).

138. Petek, L. M., Russell, D. W. & Miller, D. G. Frequent endonuclease cleavage at off-target locations in vivo. *Mol. Ther. J. Am. Soc. Gene Ther.* **18,** 983–986 (2010).

139. Barzel, A. *et al.* Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Res.* **39,** 6646–6659 (2011).

Vita

Kyle M. Jacoby, son of Glenn and Louise Jacoby and brother of Jenifer Lahiff, was born in Burbank, California in 1986. He graduated from North Hollywood's Biological Sciences (Zoo) magnet program in 2004 with high honors. During his undergraduate studies and the University of California, Santa Barbara, he got his first taste of laboratory research under the tutelage of Dr. David Low. After graduating from UCSB with honors in 2008 he was admitted to the Molecular and Cellular Biology program at the University of Washington and joined Dr. Andrew Scharenberg's lab. He completed his doctor of philosophy in 2013 for his thesis work on homing endonuclease discovery and characterization.