View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Scholarship@Western

Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

9-12-2016 12:00 AM

# Using mutual information to reprogram DNA specificity of LAGLIDADG endonucleases

Marcon Laforet
*The University of Western Ontario*

Supervisor
Dr. David Edgell
*The University of Western Ontario*

Graduate Program in Biochemistry
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science
© Marcon Laforet 2016

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Biochemistry Commons

## Recommended Citation

Laforet, Marcon, "Using mutual information to reprogram DNA specificity of LAGLIDADG endonucleases" (2016). *Electronic Thesis and Dissertation Repository*. 4191.
https://ir.lib.uwo.ca/etd/4191

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

## Abstract

A systematic approach to reprogram protein-DNA interactions has yet to be discovered. This study investigates the ability of co-variation analyses to identify potential protein-DNA contacts that regulate specificity. Here, 27 LAGLIDADG Homing Endonucleases (LHEs) and their 22-basepair DNA targets were collated into a Multiple Sequence Alignment (MSA) that was subjected to pairwise co-variation calculations. Using the LHE I-OnuI as a reference, an amino acid-DNA pair, lysine (K) 231 and adenine +3, generated the highest score. To test if the K231/A3 score was biologically relevant we tested protein mutants for altered nuclease specificity at +3 DNA point mutants. Randomizing the 231st amino acid did not alone restore cleavage activity on substrate mutants but randomization in conjunction with aspartic acid (D) 240 restored cleavage activity on A3T and A3G substrates. In conclusion, co-variation analyses identified, in part, amino acids that could be mutated to alter DNA specificity. Future work should focus on mapping more LHE-DNA target sequences to increase MSA diversity.

**Keywords:** co-variation, mutual information, protein-DNA interactions, specificity, multiple sequence alignment, prediction, LAGLIDADG, homing endonuclease

## Acknowledgements

---

I would like to extend a sincere thank you to Dr. David Edgell, without his guidance during my experiments the work presented here would not have been possible. I would also like to thank Dr. Gregory Gloor for his guidance during the computational analysis, rigorous insight and generosity, without which this thesis would not have been possible. Furthermore, I would like to thank my committee member Dr. Patrick O'Donoghue for his helpful insight and support. I would like to extend many thanks to Jason Wolfs for his continual mentorship and advice during my undergraduate and Masters thesis. I would also like to thank Dr. Stephanie Dorman and Mike Ellis for their guidance and advice regarding my Masters thesis. Finally, I would like to thank Thomas McMurrough for his guidance to set up the LHE selection procedure and former lab technician Nancy Friedrich for her helpful advice during my Masters thesis.

## Table of Contents

---

**ABSTRACT** ............................................................................................ **iii**

**ACKNOWLEDGEMENTS** ............................................................................ **iv**

**TABLE OF CONTENTS** ............................................................................ **v**

**LIST OF FIGURES** ................................................................................ **vi**

**LIST OF TABLES** .................................................................................. **vii**

**LIST OF ABBREVIATIONS** ...................................................................... **viii**

**CHAPTER ONE - INTRODUCTION** ........................................................... **1**
- 1.1 Investigating a protein-DNA code ........................................................ 1
- 1.2 LAGLIDADG homing endonucleases – general properties ..................... 3
- 1.3 LAGLIDADG homing endonucleases – reprogramming ......................... 6
- 1.4 Using mutual information to assess protein-DNA interfaces .................. 8
- 1.5 Hypothesis and aims ............................................................................ 9

**CHAPTER TWO - MATERIALS AND METHODS** ......................................... **10**
- 2.1 Multiple sequence alignment ............................................................... 10
- 2.2 Mutual information calculations ........................................................... 10
- 2.3 Alignment sensitivity ............................................................................ 11
- 2.4 Plasmid construction ............................................................................ 11
- 2.5 Two-plasmid selection of I-OnuI and I-OnuI libraries ............................ 12
- 2.6 Bacterial growth curves ....................................................................... 13
- 2.7 *In vitro* nt competition cleavage assay .................................................. 13
- 2.8 Profiling nuclease specificities using MiSeq illumine sequencing ............ 14

**CHAPTER THREE - RESULTS** ................................................................. **15**
- 3.1 LHE MSA MIp calculations and analysis ............................................... 15
- 3.2 Experimentally investigating altered specificity at identified protein-DNA pair 17
- 3.3 Mutant I-OnuI survivors ....................................................................... 20
- 3.4 Deconvoluting I-OnuI mutants ............................................................. 22
- 3.5 Relative bacterial growth rates of I-OnuI and mutants on nt point mutants ........ 22
- 3.6 *in vitro* cleavage activity of I-OnuI and mutants on +3 DNA point mutants ...... 23
- 3.7 Profiling I-OnuI and mutants specificity using Illumina sequencing .............. 25

**CHAPTER FOUR - DISCUSSION** ............................................................... **29**
- 4.1 MI as a technique for characterizing protein-DNA interactions ................ 29
- 4.2 Comparing LHE reprogramming results with previous findings ............... 31
- 4.3 Future directions .................................................................................. 32
  - 4.3.1 Enhancing covariation analysis ....................................................... 32
  - 4.3.2 Additional applications of MI .......................................................... 32

## List of Figures

---

**Figure 1. General homing endonuclease lifecycle. .......................................................... 4**

**Figure 2. LAGLIDADG homing endonuclease features. ............................................. 5**

**Figure 3. Sample MSA extract showing the first LAGLIDADG chain ..................... 11**

**Figure 4. Boxplots of MIp Zpx score distributions for A) aligned MSA iteration and B) K231 and A+3 randomly shuffled DNA MSA (n = 10 000). .................................. 15**

**Figure 5. Crystal structure of I-OnuI in complex with its WT target site (pdb ID: 3QQY) ........................................................................................................................ 16**

**Figure 6. Survival of I-OnuI on +3 DNA point mutants. ........................................... 18**

**Figure 7. Mutant I-OnuI library survival on A3A, A3G and A3T substrates. ......... 19**

**Figure 8. Subcloned I-OnuI mutant survival on +3 DNA point mutants. ................. 21**

**Figure 9. Deconvoluted I-OnuI mutant survival on DNA +3 point mutants............ 21**

**Figure 10. Relative growth rates of I-OnuI and mutants on +3 DNA point mutants. 23**

**Figure 11. I-OnuI and mutants relative cleavage activity on +3 DNA point mutants; *in vitro* nt competition assay. ............................................................................... 24**

**Figure 12. Histograms of rate of change values for 4N sequences in mock and I-OnuI WT, E and YA cleavage assays. ............................................................... 26**

**Figure 13. Heat map of I-OnuI WT, E and YA nuclease activity on +2 to +5 DNA point mutants. ........................................................................................................ 28**

## List of Tables

**Table 1. The 6 highest Zpx scores from the aligned LHE MSA MIp calculations. .. 15**

**Table 2. Summary table of sequence diversity of LHE-DNA pairs at the aa 231 and + 3 nt substrate. .......... 15**

**Table 3. Surviving mutants from I-OnuI 1NNS and 3NNS libraries on + 3 DNA point mutant substrates .......... 20**

**Table 4. Rate of change for sequences robustly depleted during the 20-minute cleavage assay ($p \leq 0.05$) .......... 27**

## List of Abbreviations

| | |
|---|---|
| A | adenosine |
| aa | amino acid |
| bp | base pairs |
| C | cytosine |
| DNA | deoxyribonucleic acid |
| DSB | double-strand break |
| DTT | dinitrothreitol |
| EDTA | ethylenediaminetetraacetic acid |
| FDR | false discovery rate |
| G | guanosine |
| HDR | homology directed repair |
| HE | homing endonuclease |
| I-OnuI | I-OnuI E1 |
| I-OnuI-E | I-OnuI D204E |
| I-OnuI-YA | I-OnuI K231Y D240A |
| IPTG | isopropyl-β-D-1-thiogalactopyranoside |
| LB | luria broth |
| LHE | LAGLIDADG homing endonuclease |
| MGE | mobile genetic elements |
| MIp | phylogenetically corrected mutual information |
| MSA | multiple sequence alignment |
| NHEJ | non-homology end joining |
| nm | nanometer |
| nt | nucleotide |
| OD | optical density |
| ORF | open reading frame |
| p | plasmid |
| PAGE | polyacrylamide gel electrophoresis |
| PCR | polymerase chain reaction |
| SDS | sodium dodecyl sulfate |
| SGE | selfish genetic element |
| SOC | super optimal broth with catabolite repression |
| T | thymine |
| WT or wt | wildtype |
| ZF | zinc finger |

# CHAPTER ONE - INTRODUCTION

## 1.1 Investigating a protein-DNA code

Being able to take protein sequence information and determine which amino acids contribute DNA specificity would greatly improve the ability to modulate protein-DNA interactions, benefiting industrial, medical and academic institutions. Such benefits include: retargeting genome-editing reagents to novel genomic sites, modification of protein binding affinity to control gene transcription, or modification of chromosome organization. The pursuit of a robust protein-DNA code that is minimally dependent on crystal structures and generalizable to any protein-DNA interface is ongoing and perhaps impossible to achieve. Successfully identifying specific protein-DNA contacts will likely require a combination of computational and biochemical approaches, on a case-by-case basis.

In the post-genomic era, the scientific community has shifted its focus from sequencing genomes to studying how they are regulated (Lander, 2011). It has become evident that proteins largely control the expression, organization and lifecycle of DNA within genomes (Mitchell and Tjian, 1989; Ren *et al*., 2000; Muller and Vousden, 2013). As such, aberrant protein-DNA interactions underlie many disease states (Boutell *et al*., 1999; Yu *et al*., 2009; Jimenez, 2010; Lander, 2011), initializing the pursuit of designing custom protein-DNA interfaces to rewire genomic networks or create new pathways. To do so, we must understand the intricacies of DNA recognition administered by proteins. Proteins must facilitate an appropriate level of affinity and specificity for their DNA cognates, obtained in part by directly contacting nucleotide (nt) sequences. Identifying which amino acids specifically contact nts can be challenging as nt recognition can be accomplished by amino acid networks, including metal ions and water molecules. Additionally, proteins may indirectly read out 3-dimensional features of nts like twist, minor groove distance or flexibility that also contribute specificity to the interaction (Rohs *et al*., 2009; Stella *et al*., 2010; Thyme *et al*., 2014).

One of the first DNA-binding proteins studied were the zinc-finger (ZF) proteins. A single ZF motif has a ββα architecture that coordinates a single zinc ion, recognizing a

3 nt triplet. Arrays of ZFs can be artificially assembled to extend the recognition sequence, with one of the first artificial ZFs recognizing a stretch of 18 nts. Notably, this ZF fusion was able to activate or repress expression of genes on a reporter plasmid by fusing the ZF to an activator or repressive domain that interacted with the transcription machinery (Liu *et al*., 1997). Although 18 nts is theoretically sufficient to identify a unique DNA site, it became apparent that ZF specificity is not exact and that different ZF assemblies can tolerate varying degrees of nt mismatches to their predicted binding site (Kim and Pabo, 1997; Beerli *et al*., 1998).

Follow-up studies largely focused on reprogramming ZFs to novel nt triplets. Initially pursuing ZFs that recognized the 16 5' – GNN – 3' variants, studies built and screened a library of ZF mutants for altered binding specificities. These studies identified regions within the α-helix that contributed to binding specificity (Beerli *et al*., 1998). Mutational investigation into this area produced ZF mutants that discriminated against nts at the 3rd (GNN) position (Dreier *et al*., 2000). Follow-up studies used similar mutagenic approaches to expand binding specificity to 5' – ANN – 3' and 5' – CNN – 3' sequences (Dreier *et al*., 2001, Dreier *et al*., 2005). Blancafort *et al*. (2003) simplified the process of assembling ZFs by collating a library of individual ZF motifs that recognized many of the 5' – NNN – 3' triplets. Mutant ZF binding specificities in these studies were tested by their ability to repress or activate endogenous genes in model organisms.

Reprogramming ZFs to recognize all the possible 64 nt triplet combinations was laborious and furthermore, some triplets still cannot be recognized (Dreier *et al*., 2005). These studies began to unravel the complexities of DNA-protein interactions, dispelling notions of a simple one-to-one protein–DNA code that ubiquitously governs specificity of protein-DNA interactions. These studies also demonstrated how amino acids could form networks to create an interaction interface to specifically recognize a DNA sequence (Dreier *et al*., 2005), further complicating these interaction interfaces. Moreover, it was found that some amino acids suspected of binding DNA participated in non-specific nt interactions, non-specifically contributing to the necessary DNA binding energy (Dreier *et al*., 2000). Moving forward in the pursuit of reprogramming protein-DNA interfaces, it is clear that scientists must strive to intimately understand individual amino acids contributions to specific DNA binding.

Advances in computation have lead to the development of simulations that characterize aa-nt interactions within protein-DNA interfaces (Pabo and Nekludova, 2000; Havranek *et al.*, 2004; Rohl *et al.*, 2004; Thyme *et al.*, 2014). Computational models utilize crystallographic data to characterize the structural and thermodynamic features of a protein-DNA interface. These models are then used to predict aa mutations that may recognize nt substitutions, maintaining the necessary structural and thermodynamic characteristics of the interface. Scientists have successfully used this approach to reprogram nt specificity at obvious aa-nt hydrogen contacts, however, they had difficulty in efforts to reprogram extensive protein-DNA interfaces (Ashworth *et al.*, 2006; Thyme *et al.*, 2009; Ulge *et al.*, 2011). Furthermore, when biologically testing the DNA-binding specificity of the predicted protein variants, additional genetic selections are commonly needed to isolate variants with increased activity or specificity (Ashworth *et al.*, 2006; Takeuchi *et al.*, 2011). Arguably, the biggest challenge to computational models of protein-DNA interfaces is their ability to assess proteins ability to indirectly readout intricate details of DNA molecules. Nonetheless, computational advancements have allowed scientists to target mutagenesis studies of proteins, reducing the laborious efforts needed to reprogram protein-DNA interfaces.

### 1.2 LAGLIDADG homing endonucleases – general properties

Homing Endonucleases (HEs) are natural DNA endonucleases that have been intensely studied because of their potential use as genome editing reagents. HEs are site-specific DNA nucleases that introduce a double-stranded break (DSB) into DNA at specific sites lacking the HE ORF. The HE lifecycle accomplishes gene conversion, propagating their own DNA coding region in respect to non-self genetic material. Gene conversion occurs when homology directed repair (HDR) uses a DNA template containing the HE ORF during DSB repair. Following this repair event, the DNA segment includes a HE, disrupting the HE recognition sequence (Fig. 1). This gene conversion process is known as homing. HEs are subject to evolutionary pressure that maintains a balance between DNA binding specificity and permissivity to facilitate the homing process while avoiding cellular toxicity. A common characteristic between HEs is their ability to tolerate significant nt variation within their target sites. This sequence-tolerant binding facilitates cleavage of target sites that have accumulated nt substitutions through genetic drift or other evolutionary processes (Stoddard *et al.*, 2005; Scalley-Kim *et al.*, 2007).

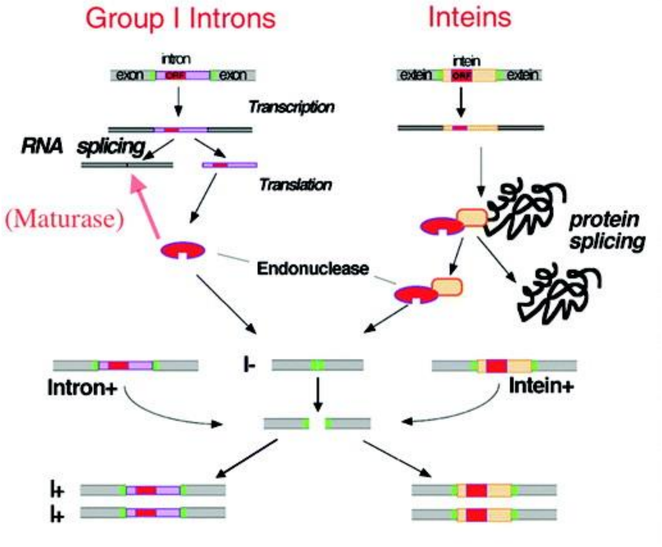

**Figure 1. General HE lifecycle.** HE ORFs are found in group I introns or inteins (shown here) but may also exist in group II introns or free-standing elements. In all cases, HE ORFs code for endonucleases that recognize homing sites that lack the HE ORF. The HE makes a DSB, activating host repair pathways that may be repaired by HDR. HDR events use a template containing the HE ORF. (Printed from Stoddard (2005) with permission from publisher).

Many separate instances have resulted in the evolution of HE families that are uniquely characterized by their method of recognizing DNA targets and by their means of introducing DSBs. The LAGLIDADG homing endonuclease (LHE) family is the best characterized HE and contains the most members. A single LAGLIDADG monomer consists of an αββαββααα structure, with the LAGLIDADG sequence denoting the amino acid consensus sequence that forms an interaction network along an exposed surface of the first α-helix (Fig. 2A-C).

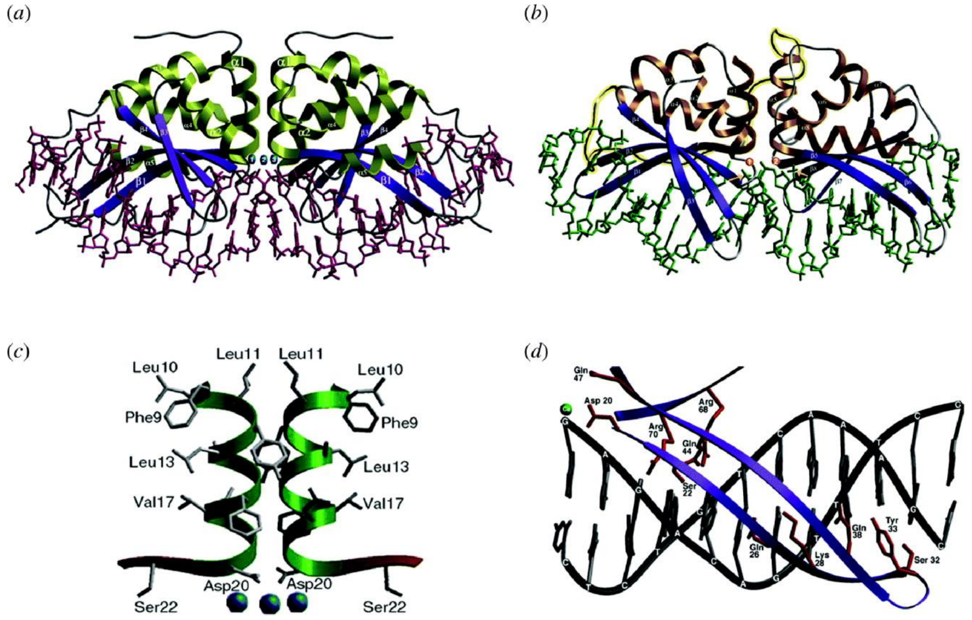

**Figure 2. General LAGLIDADG homing endonuclease features.** A) Homodimeric LAGLIDADG I-CreI. B) Monomeric LAGLIDADG I-AniI. C) I-CreI α–helix LAGLIDADG interface. D) I-CreI anti-parallel β–sheet binding to major groove of DNA target site.

A functional LAGLIDADG protein is formed from the interaction of two LAGLIDADG monomers to form a composite active site at the base of the two α1-helices. LHEs may exist as single genes that homodimerize to recognize a pseudo-palindromic target site or as a single-chain gene-fused dimer, where individual domains can diverge and recognize distinct DNA sequences. DNA recognition by LHEs is accomplished by anti-parallel β-sheets that straddle the major groove of DNA from nts ± 3 to ± 11 of the 22 nt target site, making direct, indirect and water mediated contacts to DNA (Fig. 2D). The interface is under saturated, with respect to aa-DNA hydrogen bonds, participating in 65 – 75 % of possible contacts (Stoddard *et al*., 2005). The central four nts, ± 1-2 positions, are not in direct contact with amino acids and are flanked by scissile phosphates.

## 1.3 LAGLIDADG homing endonucleases – reprogramming

Many efforts have been made to reprogram LHE specificity from their native sites for therapeutic, industrial and academic applications (Seligman *et al*., 2002; Sussman *et al*., 2004; Thyme *et al*., 2014). LHEs are inherently more specific than ZFs as they recognize longer DNA sequences and offer the benefit of intrinsically containing a sequence specific nuclease. To generalize LHE DNA-binding specificity parameters for the LHE family is challenging because residues contributing to target site recognition are not well conserved. This observation suggests that there is, as of yet, no universal protein-DNA code that describes LHE-DNA interactions. Reprogramming LHE specificity is therefore done on a case-by-case basis.

The first approach used to redesign LHE-DNA interfaces relies on crystallographic data to determine which amino acids specifically contact DNA. For I-CreI and other LHEs, the contributions of suspected residues that confer DNA-binding specificity are investigated by mutational analysis, screening for LHE mutants that have altered nuclease properties. Crystallographic and mutational investigations of different LHEs over many years have identified modules of amino acids that contribute to DNA specificity. As summarized by Barry Stoddard on the homingendonuclease.net website, these modules consist of 8-12 amino acids that contribute specificity of up to 3 nts. However, the modules differ between LHEs and to date have been identified only by crystallographic and mutational analyses.

Previous studies that have focused their efforts on reprogramming I-CreI identified variants tolerating many nt substitutions. Seligman *et al*. (2002) screened I-CreI libraries that contained randomized residues suspected of specifically contacting DNA as per the co-crystal structure. One finding from these studies was the realization that crystal structures did not entirely describe the importance or flexibility of protein-DNA contacts, as they identified mutants that ranged from having no effect on binding specificity to those resulting in cellular toxicity. Further studies by Sussman *et al.* (2004) successfully reprogrammed DNA specificity at the ± 6 and ± 10 positions by making mutations at the 26th, 33rd and 66th amino acid positions, but also reported enormous variance in nuclease activity. Taken together, these studies illustrate the challenge of producing mutants that preserve sufficient binding affinity and activity while maintaining site discrimination. These studies also demonstrated drastic changes to activity and affinity of I-CreI on substrates due to single amino acid mutations.

Interestingly, a network of amino acids within I-CreI was identified, mutation of which lead to expansion of nt specificity at the ± 3, 4 and 5 target site positions (Arnould *et al.*, 2006). From crystallographic data, they hypothesized that these target site nts were being recognized by R70, Q44 and R68, respectively. Before the specificity of R70, Q44 and R68 variants were tested, scientists realized that another amino acid would have to be altered to accommodate the negative charge generated during the DSB. As a pre-emptive suppressor screen, R70, Q44 and R68 mutations were made in the D75N background to reduce energetic constraints and allow localized restructuring of the amino acid network (Arnould *et al*., 2006). This study concluded a rough protein-DNA code for which the Q44-4A pair reported as A44-4T or K44-4G. Their analysis did not suggest any clear protein-DNA code for the other positions. In these cases, isolated I-CreI mutants that had altered nt specificity contained randomly assorted amino acids (Arnould *et al.,* 2006).

A second approach to reengineer protein-DNA interfaces utilizes *in silico* approaches dependent on crystal structures and thermodynamic calculations of protein-DNA interfaces. Combinatorial approaches to reprogram DNA specificity of LHEs that integrated computational and mutagenesis methods have demonstrated the most success. This multi-faceted approach has been used to reprogram I-MsoI specificity at ± 6 nt positions (Ashworth *et al*., 2006; Ashworth *et al*., 2010). In 2006, K28L and T83R mutations were made to accommodate a G → C transversion mutation whereas more extensive amino acid mutations were made in 2010 to accommodate substitutions at 3 adjacent nt positions. Perhaps most successively, Thyme *et al*. (2014) were able to reprogram a LHE to a target site containing 12 nt substitutions. Multiple studies have drawn attention to LHE's capacity to indirectly readout DNA sequences (Molina *et al*., 2012; Thyme *et al*., 2014). Modeling improvements that incorporate protein's ability to indirectly readout DNA parameters (Rohs *et al.*, 2009), while accounting for water molecules (Lazaridis and Karplus, 1999; Li and Bradley, 2013) and various backbone conformations (Yanover and Bradley, 2011; Thyme *et al.*, 2012), have independently been implemented, improving computer's abilities to reprogram protein-DNA interfaces.

However, the intricacies of incorporating all these analyses and predictions together have yet to be worked out.

An additional avenue to enhance LHE reprogramming efforts has been to utilize the natural diversity of LHEs and their target sequences found throughout nature. Barzel *et al*. (2011) initialized this approach by developing computational methods to search characterized genomes for novel LHEs and their putative target sites, subsequently validating the putative LHE target sites. Takeuchi *et al*. (2011) then phylogenetically analyzed 211 LHE sequences to illustrate the conservation of the LHE scaffold contrasted by their diverging DNA recognition sequences. This highlighted the potential for the LHE scaffold to be repurposed and direct LHEs to relevant human targets. McMurrough *et al.* (2014) then utilized this phylogenetic diversity of LHEs to identify amino acids that control the catalytic efficiency of the enzyme. Specifically, this study highlighted the potential of using natural LHE diversity to gain insight into LHE function. In this study, we capitalize on the phylogenetic diversity of LHEs and their respective target sites to identify amino acids that confer DNA specificity. We do so by applying a mathematical framework to identify amino acids that are co-varying with nts in their DNA target sequence.

### 1.4 Using mutual information to assess protein-DNA interfaces

Conserved residues within a protein family play a significant role in structural and functional aspects (Clarke, 1995). Some amino acid positions whose mutations are detrimental to protein activity can be rescued by secondary mutations that restore crucial features to the protein. Examples of these include compensatory mutations that restore internal volumes, salt bridges, $H_2O$ contacts or binding and folding energies. This mutational dependency between amino acid positions characterizes an intramolecular coevolutionary relationship, with residues likely close to each other in 3-dimensional space (Atchley *et al*., 2000; Oliveira *et al*., 2002). Analyzing the phylogenetic diversity within a protein family can identify such relationships using a mathematical procedure to characterize sequence entropy in a MSA. Transformed into a mutual information (MI) reading, this statistic characterizes interdependency also noted as co-variation between MSA columns. Tillier and Liu (2003) improved the quality of this statistic by removing sequence variation due to phylogenetic divergence of the protein family. Further corrections were made by Dunn *et al*. (2008) to additionally remove the average entropy within the MSA, resulting in a corrected MIp score. A Z-score procedure is often used to assess MIp scores and identify those with the highest co-variation. Little and Chen (2009) improved this Z-score procedure by subtracting the average MIp score at each position. This calculation was computationally intensive because linear regression on scores obtained from each position had to be conducted but was made more efficient by Dickson *et al*. in 2010. This alteration formed a more robust and efficient statistic, Zpx, which is less sensitive to local misalignments within the MSA.

In this study, we use Zpx scores to identify biologically relevant protein-DNA contacts of LHEs. This approach has been previously applied to a LHE MSA, identifying intramolecular residues coevolving to maintain steric and chemical properties of residues within the active site (McMurrough *et al*., 2014). Furthermore, this co-variation analysis has been previously applied to intermolecular protein-DNA contacts of well-characterized transcription factors (Mahony *et al*., 2007). Mahony *et al*. (2007) used alignments containing > 1000 sequences to validate known protein-DNA interactions that confer specificity to the protein-DNA interaction.

### 1.5 Hypothesis and aims

In light of previous studies, we believe that it is reasonable to assume that co-variation analysis can identify residues interacting in 3-dimensional space. Although most of these studies investigated intramolecular residues, we believe that intermolecular residues can also co-dependently evolve to form complimentary surfaces. Here, we aim to demonstrate how co-variation analysis can identify these co-evolving residues. Moreover, we aim to show that these residues play a role in the binding specificity of a macromolecular interphase. Using LHEs and their DNA cognates as a model system, I hypothesize that co-variation calculations will be able to identify specific amino acids that are co-varying with DNA. Furthermore, because we believe that these residues are specifically recognizing nts in the LHE target site, I hypothesize that mutations of these residues will alter binding specificity.

# CHAPTER TWO - MATERIALS AND METHODS

## 2.1 Multiple sequence alignment

Mapped LHE target sites identified by Thyme et al., (2014) or collected from an online database homingendoniuclease.net, maintained by Barry Stoddard. Twenty-seven monomeric LHEs were identified and collated into a MSA (Fig. 3). Cn3D and structural alignment algorithms were used to produce the attached alignment in FASTA file format. Structural files were used if possible and a local covariation plug-in to Jalview were also used to align these sequences (Dickson and Gloor, 2012). For alignment quality and accuracy, structures in Cn3D were used to largely direct the alignment. Target sites were aligned on the scissile phosphate nts. After we were satisfied with the alignment quality, the MSA was subject pairwise calculations of Shannon's entropy. Under this framework the probability of finding a specific amino acid in a column is determined. As amino acid identity becomes more predicable at a position within the MSA, entropy is lowered and information is gained. Information can be gained if one position in an alignment enables better prediction at a distinct position. This calculation results in a quantity known as MI and is calculated for every possible column pair. Corrections to these calculations were also applied to remove the average phylogenetic entropy producing a MIp score. Further analysis reveals pairs of columns with higher than average MIp scores, suspected of co-evolving in 3-dimensional space.

## 2.2 Mutual information calculations

The MIp Toolset written by Dickson and Gloor (2013) was used. This software imports a MSA as FASTA file and produces a summary spreadsheet that was then further processed and plotted using R. A column in the spreadsheet contains the MIp score for every pair of residues in the alignment and also calculates a Zpx score that shows the number of standard deviations an individual MIp score is from the average MIp score. Boxplots of Zpx scores show the distribution of scores and outliers that display higher than average MIp suggesting coevolution (Fig. 4A).

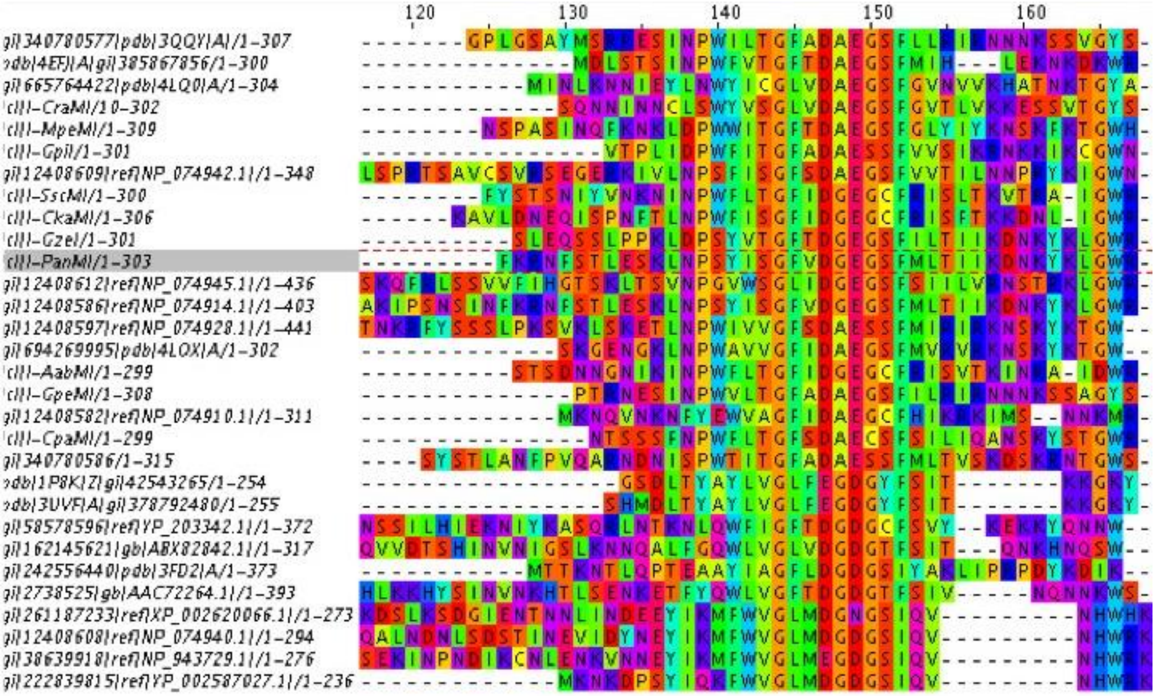
120
130
140
150
160
gi|340780577|pdb|3QQY|A|/1-307
›db|4EFJ|A|gi|385867856/1-300
gi|665764422|pdb|4LQ0|A/1-304
cIII-CraMI/10-302
cIII-MpeMI/1-309
cIII-GpiI/1-301
gi|12408609|ref|NP_074942.1|/1-348
cIII-SscMI/1-300
cIII-CkaMI/1-306
cIII-GzeI/1-301
cIII-PanMI/1-303
gi|12408612|ref|NP_074945.1|/1-436
gi|12408586|ref|NP_074914.1|/1-403
gi|12408597|ref|NP_074928.1|/1-441
gi|694269995|pdb|4LOX|A/1-302
cIII-AabMI/1-299
cIII-GpeMI/1-308
gi|12408582|ref|NP_074910.1|/1-311
cIII-CpaMI/1-299
gi|340780586/1-315
›db|1P8K|Z|gi|42543265/1-254
›db|3UVF|A|gi|378792480/1-255
gi|58578596|ref|YP_203342.1|/1-372
gi|162145621|gb|ABX82842.1|/1-317
gi|242556440|pdb|3FD2|A/1-373
gi|2738525|gb|AAC72264.1|/1-393
gi|261187233|ref|XP_002620066.1|/1-273
gi|12408608|ref|NP_074940.1|/1-294
gi|38639918|ref|NP_943729.1|/1-276
gi|222839815|ref|YP_002587027.1|/1-236

**Figure 3. Sample MSA extract showing the first LAGLIDADG chain.** This extract is coloured using jalviews taylor schema.

## 2.3 Alignment sensitivity

DNA target sites were randomly shuffled using custom R scripts and methods from the seqin R package. MIp calculations were repeated with these randomly shuffled DNA target sequences 10 000 times using custom bash scripts. Bash scripts were used to sort the resulting data files and pull out the Zpx scores of K231 and the +3 DNA target site position. This file was imported into R to analyze its distribution in a boxplot (Fig. 4).

## 2.4 Plasmid construction

DNA point mutants were cloned into pCcbD at two separate sites via restriction enzymes Nhe/Sac and Afl/Bgl respectively. Oligonucleotide inserts were ordered from Integrated DNA Technologies (IDT) with the appropriate overhangs. Inserts were phosphorylated and annealed followed by ligation to appropriately cut and dephosphorylated pCcbD. I-OnuI protein libraries at amino acid positions 231, 238 and 240 were generated by using a NNS codon in a primer used to PCR amplify the coding sequence. PCR libraries were then sewn by PCR to wildtype I-OnuI backbone. These I-

OnuI libraries were restriction enzyme cloned into pMEGA, a pACYCDuet™ derivative purchased from Novagen, using Nco and Not. Individual clones were sequenced for quality assurance and library diversity estimation.

The 231 NNS library (1NNS) has a theoretical complexity of 20 aa's and was cloned with an estimated complexity of 1417. The 231, 238 and 240 NNS library (3NNS) has a theoretical complexity of 8000 aa combinations and was cloned with an estimated complexity of 8636. Ten independent clones from the 231 NNS library and twenty independent clones from the 231, 238 and 240 NNS library were sequenced and nt diversity at the first and second positions were determined. Guanine was most abundant at both the first and second positions while cytosine and adenine were the most underrepresented nts at the first and second positions, respectively leading to biased library synthesis.

## 2.5 Two-plasmid selection of I-OnuI and I-OnuI libraries

A modified bacterial two-plasmid selection was used to screen activity of I-OnuI and I-OnuI libraries on various target sites as previously described (Doyon *et al.*, 2006). A toxic plasmid contains a lactose repressed gyrase toxin and I-OnuI target sequences in wildtype or mutant contexts. Chemically competent NovaXGF' (Novagen) containing the toxic plasmid (pTox) were made as previously described (McMurrough *et al.*, 2014). Different batches of competent *E. coli* corresponded to toxic plasmids with different target sequences. Fifty Nano grams of wildtype I-OnuI or I-OnuI libraries plasmid was transformed into NovaXGF' (Novagen) cells harboring the toxic plasmid. Cells were incubated on ice for 30 minutes, heat shocked at 42°C for 1 minute and returned to ice for 2 minutes. 300 μL of minimal 2x YT medium (16 g/L tryptone, 10 g/L yeast extract and 5 g/L NaCl) was added to cultures that were recovered at 37°C for 10 minutes at 200 rpm. These cultures were transferred to test tubes containing 1 mL of 2x YT induction medium (100 μg/mL carbenicillin, 0.02% L-arabinose) and allowed to recover for 1 hour at 37°C in a rotating wheel for an outgrowth period. Cultures were then diluted accordingly and separated into selective (0.02% L-arabinose, 0.005 mM IPTG) and non-selective media (0.02 % glucose) to obtain a survival ratio (McMurrough *et al.*, 2014). This procedure can be done on plates to determine precise colony numbers or in liquid culture at 200 rpm to survey complex libraries. For liquid cultures, of the transformation was transferred into 2 mL of liquid selective or non-selective media. Both plates and liquid selections are incubated at 37$^{o}$C for 16 hours after inoculation. Here, plates were used to identify wt I-OnuI survival on point mutant substrates whereas libraries were tested both on plates and in liquid culture. Biological triplicates were done to determine survival percentage.

## 2.6 Bacterial growth curves

The two-plasmid selection was used as described above. Outgrown cultures were diluted 5-fold with selective media and 200 μL was loaded into Cellstar® 96-well suspension culture plates that were placed into Thermo scientific Multiskan GO with SkanIt software 3.2. The instrument was set to shake at 200 rpm at 37$^{o}$C, taking OD readings every 15 minutes at 600 nm for 20 hours. Growth curves were completed in technical and biological triplicate. Results were downloaded as a .RTF file and messaged using TextWrangler. They were then imported into R and raw data was a time-series of OD values for each replicate. Technical repeats were averaged and linear models were used to estimate the growth rate using the lm function in R. Results were divided by the growth rate of wt I-OnuI cleaving the wt target sequence using biological replicates to construct error bars (Fig. 10).

## 2.7 *In vitro* nt competition cleavage assay

WT I-OnuI and mutant LHEs were purified as previously described (McMurrogh *et al*., 2014). Proteins with a His tag were purified using a GE nickel column. The His-tag was attached by a sequence containing a Tev cleavage recognition sequence that was removed. Preps were run out on 10 % stacking 15 % separating SDS-PAGE gels to evaluate purity. Bradford assays with a BSA standard (0 – 0.9 mg/mL) were used to estimate protein concentration according to Beer’s law. PCR primers were used to make 2200, 1800, 1600 and 1320 bp fragments equidistance from an I-OnuI cleavage sequence. Each fragment size corresponds to a different nt at the +3 target site position. The appearance of a product was used as an indicator or successful cleavage. A single pot cleavage reaction (5 nM substrate, 50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT and 250 nM protein) was incubated at 37$^{o}$C for allotted time where 10 μL aliquots were removed every 15 minutes for 1 hour. Reactions were stopped with

5X 200 mM EDTA, Bromophenol Blue 30 % Glycerol and 0.2 % SDS stop solution. Reaction time-points were run on a 1 % agarose gel by electrophoresis at 80 V for 2.5 hours. Agarose gels were stained with TAE (2 M Tris-HCl (pH 8.0), 0.06 % glacial acetic acid, 50 mM EDTA (pH 8.0)) containing ethidium bromide for 10 minutes and destained for 15 minutes in TAE (2 M Tris-HCl (pH 8.0), 0.06 % glacial acetic acid, 50 mM EDTA (pH 8.0)). Gels were imaged using an AlphaImager™ 3400 instrument and quantitated using the accompanying spot densitometry toolbox. Biological replicates for each protein were modeled with the lm function in R to give a rate of cleavage. Models were built using band density of the cleavage product over time and then divided by their rate on wt substrate (Fig. 11).

### 2.8 Profiling nuclease specificities using MiSeq illumine sequencing

An I-OnuI target sequence was cloned with random nts at +2, +3, +4 and +5 target site positions creating a 4N library. This library was cloned with an estimated complexity of 42 000 and a theoretical complexity of 256 variants. WT, K231Y, D240A and D240E I-OnuI proteins (250 nM) were incubated with 5 nM of the 4N mp under cleavage conditions as prepared above. Samples of the reaction were taken at 0, 5, 10 and 20 minutes, stopped with stop solution and separated on a 1 % agarose gel. Supercoiled plasmid was isolated from the gel and subjected to barcoding PCR. Five replicates for each protein and a mock sample were completed and sent to the Robarts sequencing facility for Miseq illumina sequencing. The sequencing file was transformed into a count table for each sequence and replicate using a centered-log ratio approach. Plots specific for each protein and sequence were generated plotting the sequence count over time for all replicates. Linear models were generated to estimate the rate of change for each sequences and $R^2$ values were reported, measuring the accordance among replicates. Each sequences rate of change was visualized in a histogram (Fig. 12). Sequences that were drastically depleted, ≤ 5 % likelihood, were reported (Table 4).

# CHAPTER THREE - RESULTS

## 3.1 LHE MSA MIp calculations and analysis

The MSA containing LHE proteins with mapped DNA target sites was subjected to pairwise co-variation calculations to produce MIp between MSA columns (Dickson and Gloor, 2013). MIp scores were reported as Zpx values and visualized using boxplots to determine pairs of residues with the highest co-variation scores (Fig. 4A). Table 1 lists the six highest Zpx scores using I-OnuI as a reference. Notably, the highest Zpx score is an intramolecular pair that has been previously validated (McMurrough *et al*., 2014). The second highest Zpx score stems from an intermolecular aa-nt pair, K231 and A+3. Diversity of amino acids and DNA at these positions in the MSA are summarized in Table 2. Generally, R231-G3, D231-C3, K231-A/T3 and N231-T3 associations are noted.

**Table 1. The 6 highest Zpx scores from the aligned LHE MSA MIp calculations.** Residues with +/– signs represent DNA positions whereas other numbers represent I-OnuI amino acid residues.

| Residue 1 | Residue 2 | Residue 1 ID | Residue 2 ID | Zpx score |
|---|---|---|---|---|
| 25 | 181 | A | G | 6.1 |
| 231 | + 3 | K | A | 5.5 |
| 79 | 278 | N | K | 5.0 |
| - 9 | + 2 | C | C | 5.0 |
| 264 | 266 | G | K | 4.9 |
| 20 | 97 | T | H | 4.5 |

**Table 2. Summary table of sequence diversity of LHE-DNA pairs at the aa 231 and + 3 DNA substrate.**

| Amino Acid | Purine | | Pyrimidine | |
|---|---|---|---|---|
| | A | G | C | T |
| K | 3 | 0 | 0 | 2 |
| D | 0 | 1 | 4 | 1 |
| R | 0 | 6 | 0 | 0 |
| H | 0 | 0 | 0 | 1 |
| V | 1 | 0 | 0 | 0 |
| N | 0 | 0 | 0 | 4 |
| Y | 2 | 1 | 0 | 0 |
| Q | 0 | 0 | 0 | 1 |
| Total | 6 | 8 | 4 | 9 |

To test the procedural sensitivity of this co-variation statistic to the protein-DNA alignment, we randomly shuffled the DNA target sequences and recalculated the Zpx score for the K231 and +3 DNA positions 10 000 independent times (Fig. 4B). The recalculated co-variation scores using shuffled DNA did not obtain a score as extreme when using an aligned MSA. This demonstrates that our original analysis was sensitive to our alignment and not prone to identifying noise within the MSA at this Zpx extreme.

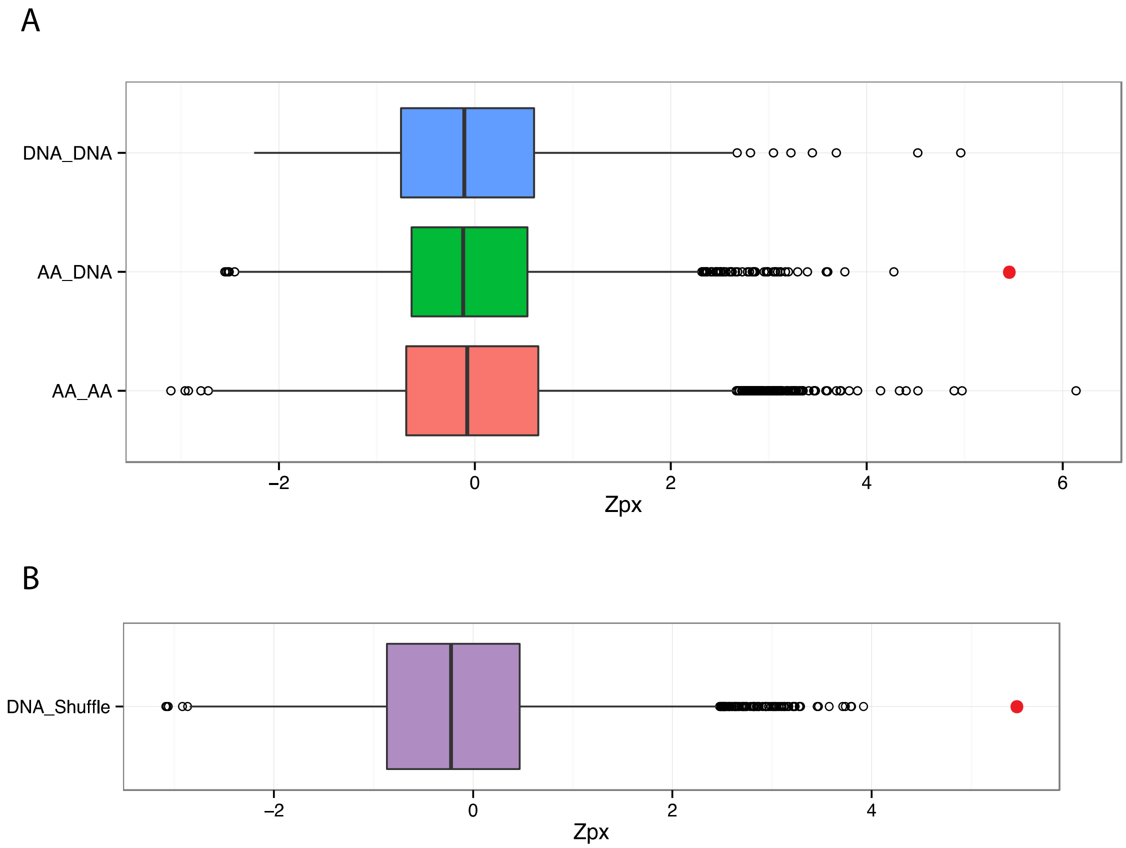

**Figure 4. Boxplots of Zpx scores from the A) aligned MSA iteration and B) K231 with randomly shuffled A+3 MSA (n = 10 000).** Thick vertical lines represent distribution mean. Boxes show the interquartile ranges and whiskers show the remaining quartiles. Open circular points represent significant outliers. The red point in both plots represents the Zpx score for the K231 and A3 positions in the aligned MSA iteration. A) Zpx scores from aligned MSA input separated by those originating between DNA (DNA_DNA), protein (AA_AA) and protein-DNA (AA_DNA). B) Zpx scores for the K231 and A3 pair from 10 000 independent iterations that used uniquely shuffled DNA (DNA_Shuffle) for each calculation.

### 3.2 Experimentally investigating the identified protein-DNA pair K231/A3

We investigated the role of the highest scoring amino acid (K231) in regulating specificity at the +3 DNA position within the context of I-OnuI. Figure 5 is a crystal structure of I-OnuI in complex with its target site.

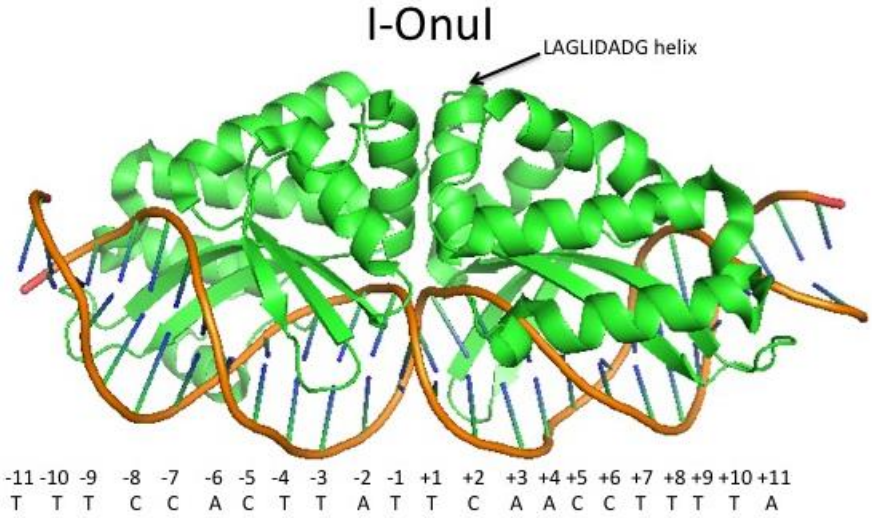

**Figure 5. Crystal structure of I-OnuI in complex with its WT target site (pdb ID: 3QQY).** DNA in complex with the first LAGLIDADG chain are denoted as negative (-) while DNA in complex with the second LAGLIDADG chain are said to be positive (+).

First, we tested I-OnuI-WT activity on +3 DNA point mutants to determine if I-OnuI was sensitive to nt mutations at this position. A two-plasmid selection assay where cleavage activity is coupled to survival (McMurrough *et al.*, 2014) was completed in *Escherichia coli (E. coli)* to measure I-OnuI-WT activity on +3 DNA point mutant substrates (Fig. 6). I-OnuI-WT survived 100 % on the WT (A3) and A3C substrates, but was inactive on the A3T substrate and showed a slow growth phenotype marked by small colony morphology on the A3G substrate (Fig. 6, *). This genetic assay identified A3G and A3T point mutants as substrates that could be used to screen I-OnuI variants for altered DNA specificity.

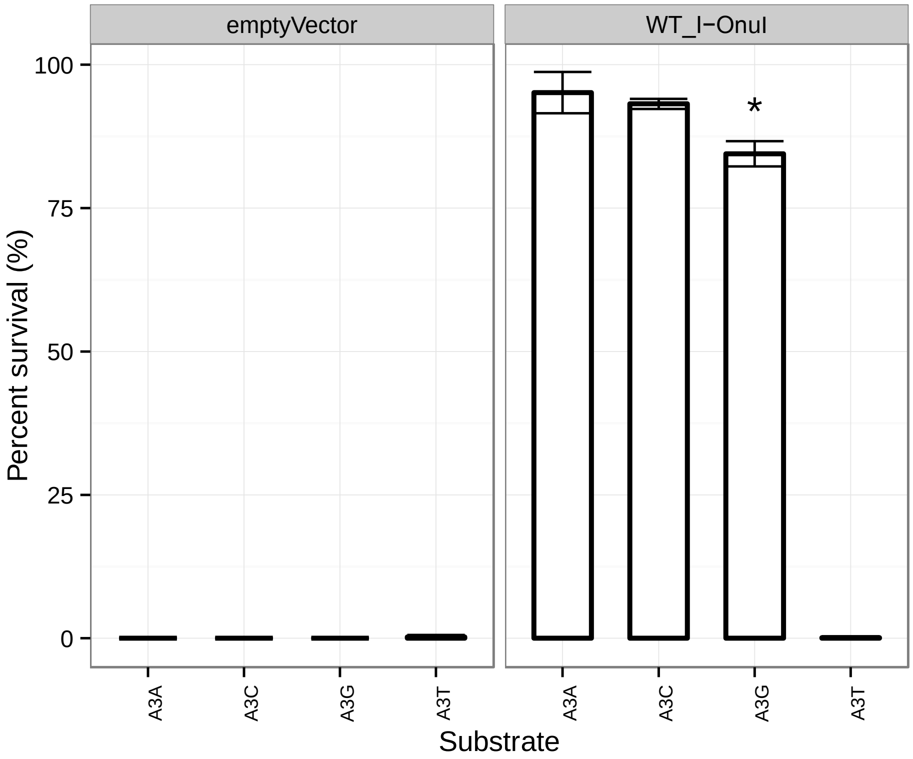

**Figure 6. Survival of I-OnuI on +3 DNA point mutant substrates.** Data represent 3 independent replicates reported as mean +/- standard deviation. The control (left) represents survival of cells transformed without an I-OnuI ORF (emptyVector). The asterix (*) denotes a small colony phenotype.

We screened an I-OnuI library containing all possible amino acids substitutions at the 231st position (1NNS) for survivors on the A3T and A3G substrates (Fig. 7; left panels). After enriching the library for active I-OnuI variants through successive rounds of selection, we only observed survival on the A3 substrate. This result motivated us to randomize nearby residues W238 and D240 to allow local restructuring of the protein-DNA interface, constructing a 3NNS library. Screening the 3NNS library for active variants identified I-OnuI mutants that survived on the A3T substrate and restored normal growth on the A3G substrate (Fig 7; right panels). Survivors were reproducibly isolated on the A3T substrate but only 1 of 3 replicates isolated a survivor on the A3G substrate.

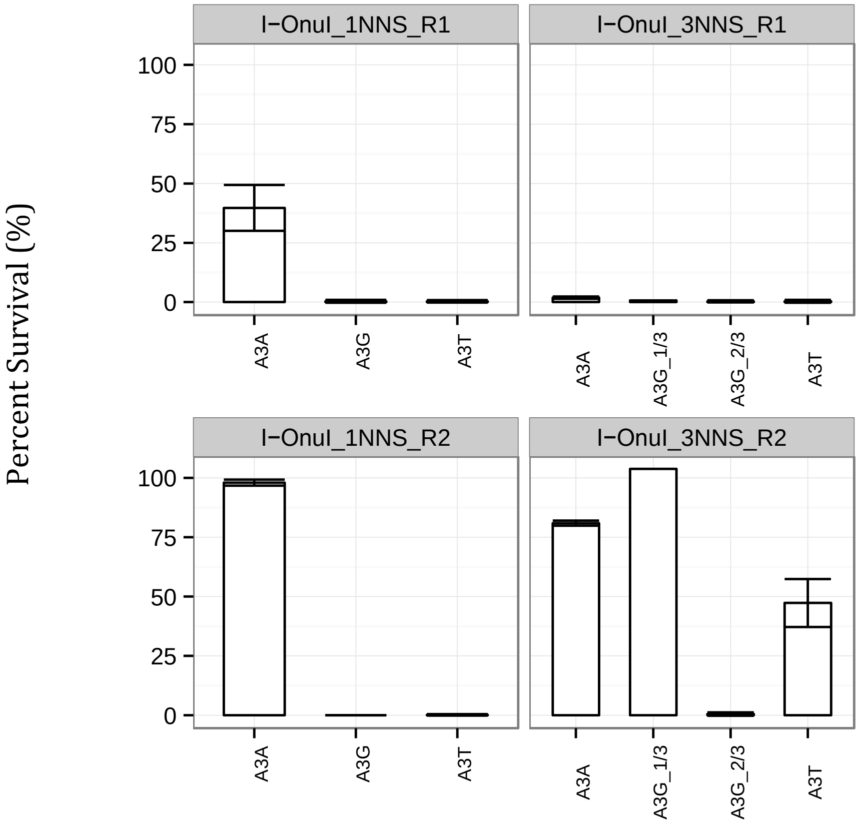

**Figure 7. Mutant I-OnuI library survival on A3A, A3G and A3T substrates.** Two I-OnuI mutant libraries, 1NNS and 3NNS were screened using successive liquid and plate selections (Round1-R1, and Round2-R2) on A3A, A3T and A3G substrates. Libraries were screened in triplicate and survival was plotted +/- standard deviation. The 3NNS library on the A3G substrate was unable to reproducibly replicate the results with only 1/3rd of replicates producing I-OnuI mutant survivors.

Individual survivors from the 1 and 3 NNS libraries were sequenced (Table 3). Mutations were isolated from both the 1NNS and 3NNS libraries on the WT A3 substrate. K231S and K231K along with K231G-W238-D240S, K231G-W238-K240V and K231R-W238-D240 were isolated from the 1NNS and 3NNS libraries respectively (Table 3). Surviving colonies screened on the A3G substrate were all K231Y-W238-D240A (I-OnuI-YA) while survivors on A3T substrate were identified as K231-W238-D240E (I-OnuI-E).

**Table 3. Surviving mutants from I-OnuI 1NNS and 3NNS libraries on + 3 DNA point mutant substrates.** After two rounds of enrichment from randomized libraries, surviving clones were picked and sequenced. Resulting clones are identified below.

| Library | +3 DNA substrate | Mutation | Isolated # of times |
|---|---|---|---|
| 1NNS | A3A | K231K | 2 |
| | | K231S | 2 |
| | A3G | None | N/A |
| | A3T | None | N/A |
| 3NNS | A3A | K231G-W238-D240S | 2 |
| | | K231G-W238-D240V | 2 |
| | | K231R-W238-D240 | 1 |
| | A3G | K231Y-W238-D240A | 5 |
| | A3T | K231-W238-D240E | 15 |

### 3.3 Mutant I-OnuI survivors I-OnuI-YA and I-OnuI-E

I-OnuI-YA and I-OnuI-E mutants were subcloned and independently tested for activity on the +3 point mutant substrates (Fig. 8). I-OnuI-YA survived on the A3A substrate and restored normal colony phenotype on the A3G substrate, but lost activity on the A3C substrate. I-OnuI-E was able to survive on all substrates including A3T, showing expanded activity, however, it still had a small colony phenotype on the A3G substrate.

### 3.4 Deconvoluting identified I-OnuI mutants

To determine the individual importance of the identified I-OnuI mutations, K231Y and D240A substitutions were introduced into WT I-OnuI to produce I-OnuI K231Y (Y) and I-OnuI D240A (A) proteins. Additionally, we were interested in possible synergetic interactions; therefore, K231Y was introduced into the D240E mutant to produce I-OnuI K231Y, D240E (YE).

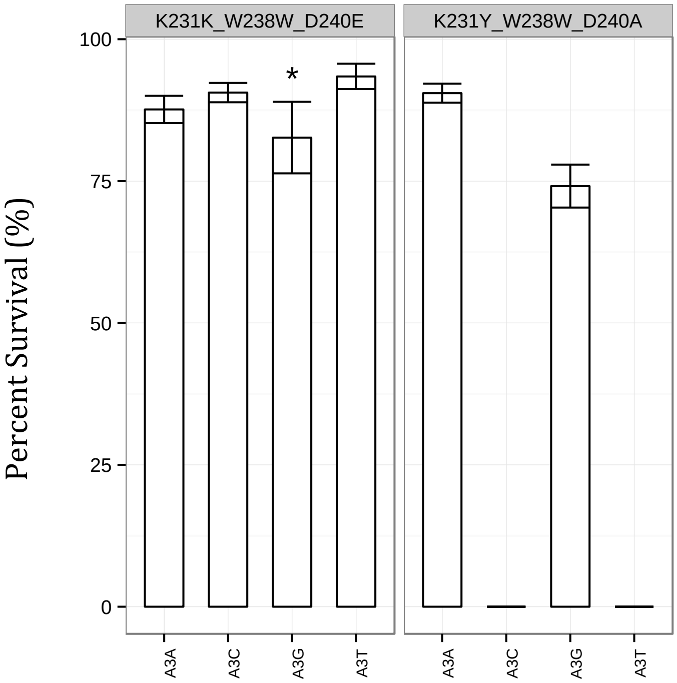

**Figure 8. I-OnuI mutant survival on +3 DNA point mutant substrates.** Isolated I-OnuI mutants K231Y-D240A and D240E were selected on +3 DNA substrates in triplicate with +/- standard deviation. The asterix (*) denotes a small colony phenotype.

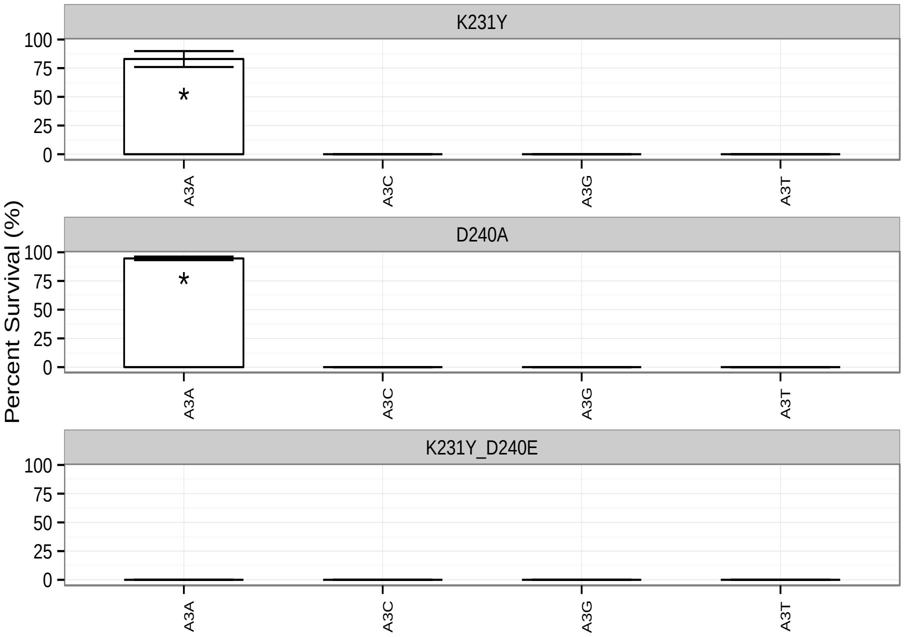

**Figure 9. Deconvoluted I-OnuI mutant survival on DNA +3 point mutant substrates.** Deconvoluted I-OnuI mutants were tested for survival on +3 DNA point mutants in triplicate with +/- standard deviation.

These mutants were assessed for activity on +3 point mutant substrates. The single mutants Y and A were able to survive on the WT (A3) substrate with small colony phenotypes (*) but could not survive on any other point mutant (Fig. 9). The double mutant, YE, was unable to survive on any substrate (Fig. 9), showing that this combination of amino acid substitutions were not synergistic.

### 3.4 Deconvoluting I-OnuI mutants

To determine the individual importance of the identified I-OnuI mutations, K231Y and D240A substitutions were introduced into WT I-OnuI to produce I-OnuI K231Y (Y) and I-OnuI D240A (A) proteins. Additionally, we were interested in possible synergetic interactions; therefore, K231Y was introduced into the D240E mutant to produce I-OnuI K231Y, D240E (YE). These mutants were assessed for activity on the +3 point mutant substrates. The single mutants Y and A were able to survive on the WT substrate with small colony phenotypes (*) but could not survive on any other point mutant (Fig. 7). The double mutant, YE, was unable to survive on any substrate (Fig. 7), showing that this combination of substitutions was not synergistic.

### 3.5 Relative bacterial growth rates of I-OnuI and mutants on nt point mutants

To more accurately quantify the slow growth phenotype of I-OnuI-WT and variants, we performed growth curves on the +3 DNA point mutant substrates. Linear models were generated to estimate the growth rate of the I-OnuI enzymes on +3 DNA substrates, which were then divided by the growth rate of WT I-OnuI on the WT A3 substrate to give a relative growth rate (Fig. 10). As a control, the average growth rate of cells harboring a toxic plasmid without an I-OnuI ORF was determined to evaluate background cell growth, reported as a dashed line. I-OnuI WT had a growth rate on the A3G and A3T substrates that did not exceed background growth rates. The relative growth rate of I-OnuI-D240E was compromised only on the A3G substrate. In accordance with previous data, I-OnuI-D240E showed robust growth on the A3T substrate. I-OnuI-YA is the only enzyme that showed appreciable growth on the A3G substrate, while displaying suboptimal growth rates on the A3C and A3T substrates.

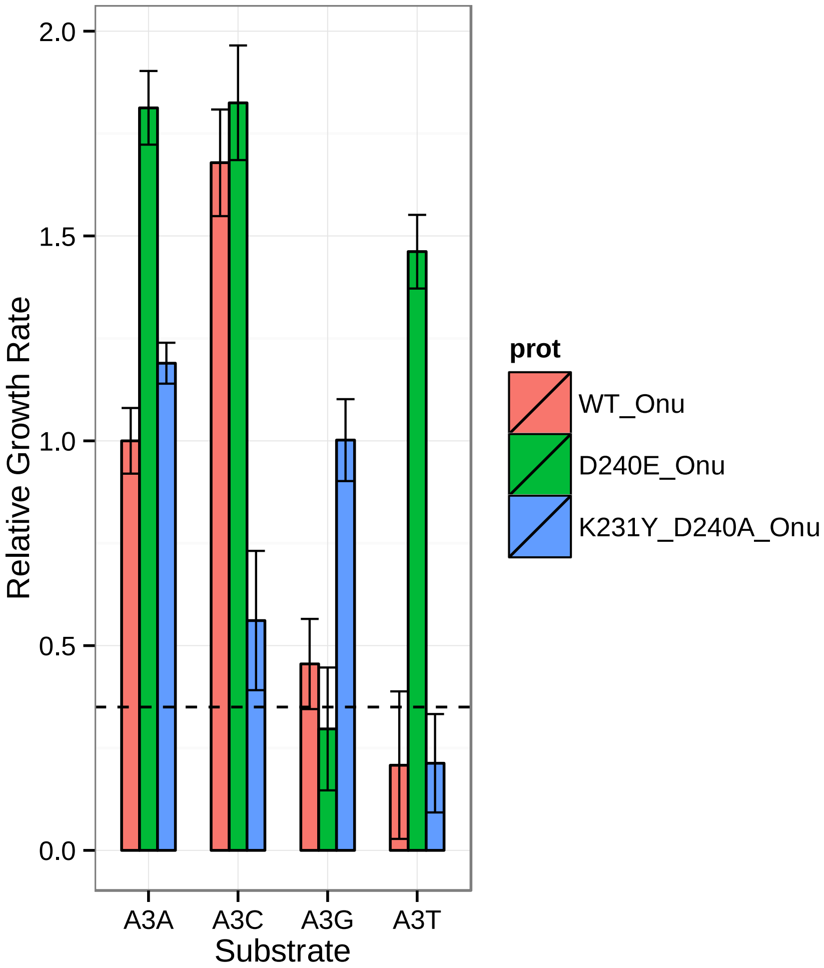

**Figure 10. Relative growth rates of I-OnuI and mutants on +3 DNA point mutant substrates.** Growth rates were estimated from triplicate growth curves with +/- standard error reported. The horizontal dashed line is the average growth rate achieved from cells devoid of the HE ORF plasmid, representing background growth rate of cells.

### 3.6 *in vitro* cleavage activity of I-OnuI and mutants on +3 DNA point mutants

An *in vitro* system was used to confirm the I-OnuI WT and mutant activity on +3 DNA substrates with purified LHEs. A barcode competition assay was performed to simultaneously measure the relative cleavage efficiency of the nucleases on all the possible +3 substrates (Ulge *et al.*, 2011). In this assay, individual +3 nt substrates were PCR amplified such that cleavage products would generate uniquely sized bands. A time-course cleavage assay was performed over 60 minutes and a linear model was used to estimate the rate of product appearance. A relative rate of appearance for each substrate was obtained by dividing by the A3 rate of appearance for each I-OnuI variant (Fig. 11).

I-OnuI WT was found to be most active on the WT substrate, with 50 % and 25 % activity on the A3C and A3G substrates respectively. I-OnuI WT had no measurable activity on the A3T substrate. I-OnuI-E showed comparable activity on all four +3 substrate variants. I-OnuI-YA maintained activity on A3A, lost activity on A3C and preferred the A3G substrate.

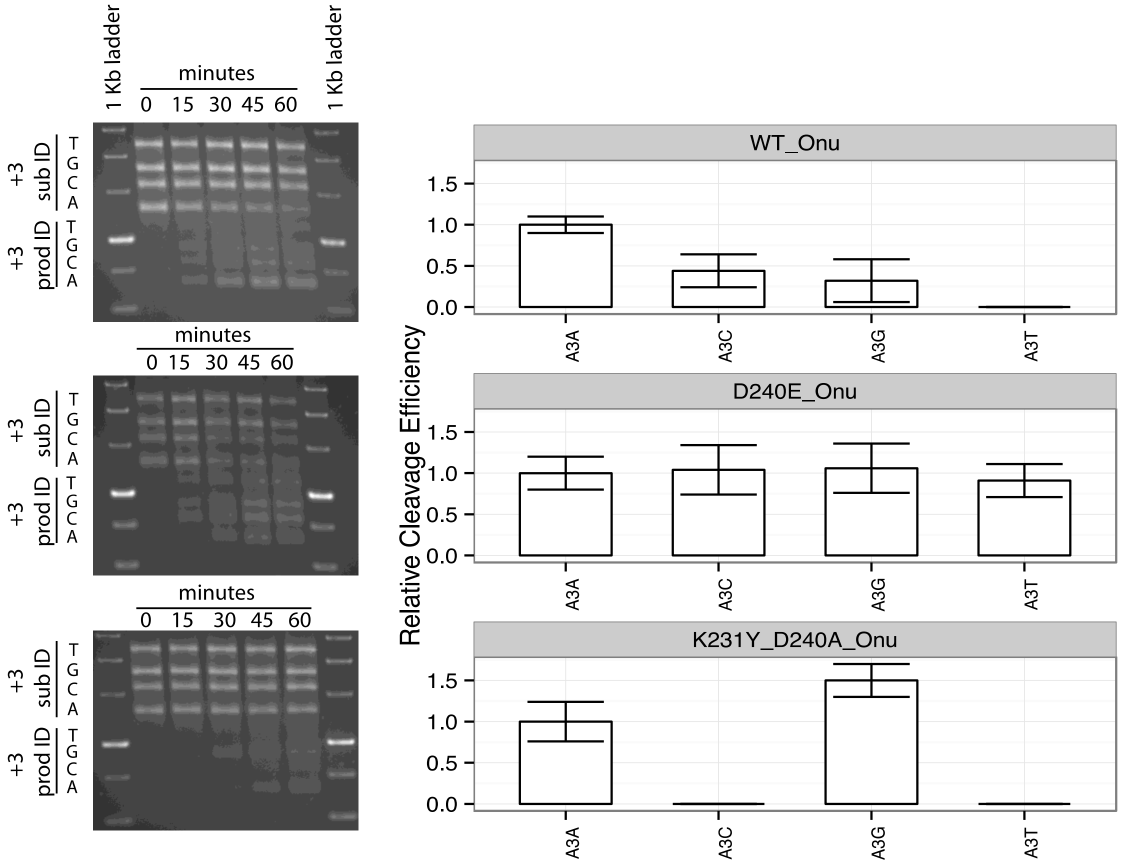

**Figure 11. I-OnuI and mutants relative cleavage activity on +3 DNA point mutant substrates; *in-vitro* nt competition assay.** PCR products of all +3 DNA point mutants were pooled at equimolar ratios followed by cleavage using I-OnuI variants. Left panels are example gels cleavage time-course studies completed in triplicate. SubID shows the uncleaved substrate for each +3 nt. Substrate sizes for T, G, C and A are 2200, 1800, 1600 and 1325 bp respectively. ProdID shows the size of the cleaved product, half the size of the substrate, for each +3 nt. The rate of appearance for ProdID bands over time were estimated and divided by the rate of appearance of WT (A3) substrate, reported with +/- standard error.

### 3.7 Profiling cleavage specificity using Illumina sequencing

WT I-OnuI substrates were randomized from positions +2 to +5 to construct a 4N library. Illumina sequencing of uncleaved DNA substrates collected over time using WT or mutant I-OnuI nucleases was completed to assess how quickly substrates were acted on by the respective nuclease. A mock replicate without the use of any nuclease was completed to estimate the normal variance we could expect from this analysis. The rate of change for each 4N DNA sequence was calculated with respect to its I-OnuI nuclease treatment. The distributions of rate of change values were visualized to compare the variance of mock and I-OnuI nuclease reactions (Fig. 12). The 4N sequences rate of change within the mock replicate shows a normal distribution with less than a 5 % likelihood of depletion rates being $\leq -2$. Sequences that were depleted $\leq -2$ in I-OnuI nuclease samples were subset from the data and their specific depletion values along with a $R^2$ value were reported (Table 4). Nineteen sequences in total fell within this range: 7, 2 and 3 sequences were uniquely called in replicates using I-OnuI WT, E and YA proteins respectively. 1 and 3 sequences were unique to WT & E and WT & YA groups, while 3 sequences were commonly depleted between all proteins. Notably, 4N sequences $\leq -2$ uniquely called in samples treated with the I-OnuI-E protein had I-OnuI-WT depletion scores that were very close to meeting the $\leq -2$ cutoff. Contrastingly, sequences $\leq -2$ uniquely belonging to samples treated with the I-OnuI-YA protein displayed normal depletion rates when treated with other proteins. Therefore, these results suggest that I-OnuI-YA has a more distinct DNA specificity than I-OnuI-E compared to I-OnuI-WT.

To assess the activity of I-OnuI nucleases on DNA point mutants within the 4N substrate, rates of change for these sequences were subset from the data. Rate of change values were divided by the rate of change of each nuclease on the WT 4N (CAAC) sequence and visualized in a heat map (Fig. 13). All nucleases preferred a C or T at position +2 and demonstrated no obvious nt discrimination at the +4 nt position. I-OnuI-E demonstrated most relative activity on A3T and A3G substrates whereas I-OnuI-YA was the only nuclease to show appreciable activity on the C5G substitution.

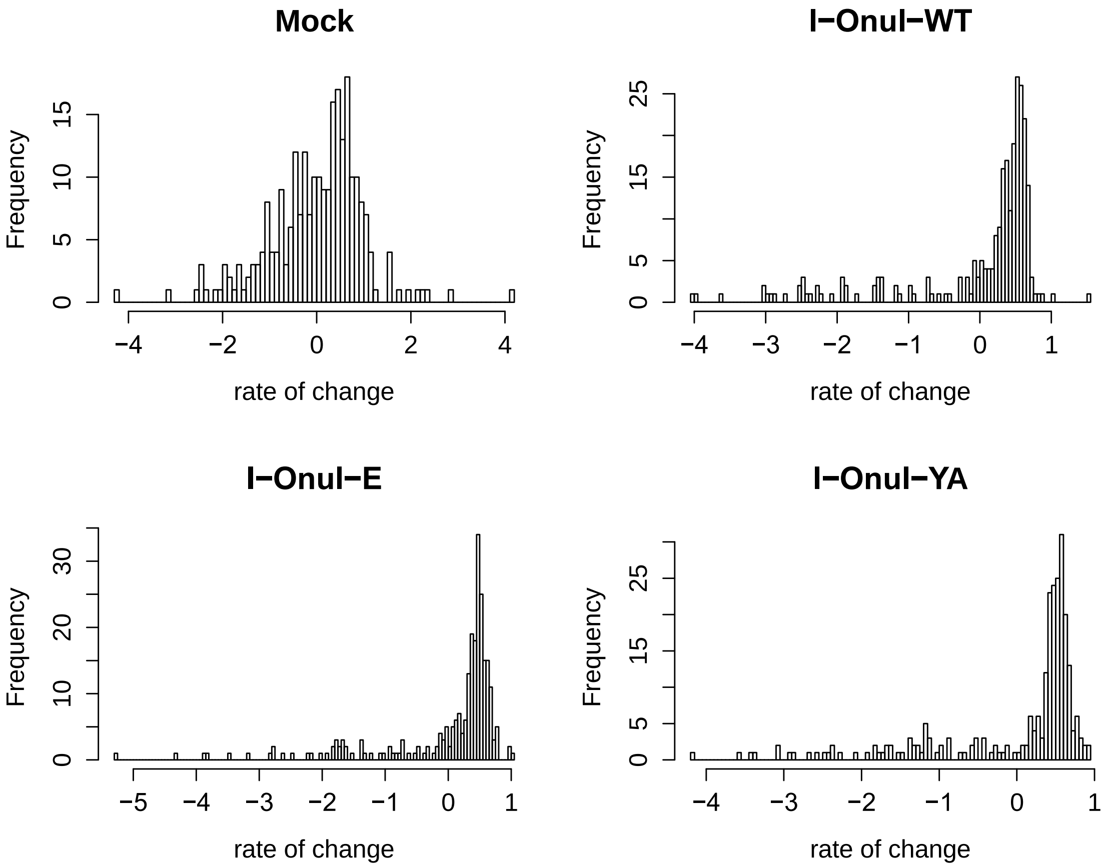

**Figure 12. Histograms of rate of change values for 4N sequences in mock and I-OnuI WT, E and YA cleavage assays.** The rate of change for sequences over the 20 minute time-course were calculated by linear models using 5 replicates for each protein cleavage assay and a single mock replicate.

**Table 4. Rate of change for sequences robustly depleted during the 20-minute cleavage assay ($p \leq 0.05$).** The rate of change for sequences estimated from 5 replicates reported with the associate $R^2$ value including a single mock replicate.

| Sequence | Protein(s) | WT | $R^2$ | E | $R^2$ | YA | $R^2$ | Mock | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| CACA | WT | -2.25 | 0.93 | -1.76 | 0.87 | -1.81 | 0.85 | -1.17 | 0.50 |
| CACC | WT | -2.48 | 0.89 | -1.83 | 0.89 | -1.63 | 0.87 | -1.05 | 0.38 |
| CACT | WT | -2.48 | 0.91 | -1.75 | 0.85 | -1.39 | 0.83 | -1.40 | 0.79 |
| TATC | WT | -2.05 | 0.62 | -0.75 | 0.34 | -1.36 | 0.44 | -0.88 | 0.45 |
| CGTA | WT | -2.37 | 0.86 | -1.25 | 0.67 | -1.70 | 0.80 | -0.61 | 0.77 |
| CATA | WT | -3.64 | 0.77 | -1.35 | 0.51 | -1.68 | 0.76 | -0.22 | 0.02 |
| CAGC | WT | -2.40 | 0.95 | -1.94 | 0.90 | -1.53 | 0.88 | -1.17 | 0.62 |
| CCCT | E | -1.93 | 0.97 | -2.16 | 0.84 | -0.51 | 0.40 | -1.88 | 0.99 |
| CGCT | E | -1.71 | 0.91 | -2.21 | 0.94 | -0.53 | 0.67 | -1.91 | 0.89 |
| CCGT | YA | -0.13 | 0.10 | -0.48 | 0.39 | -3.07 | 0.96 | -1.12 | 0.98 |
| TATA | YA | -1.35 | 0.24 | -0.73 | 0.11 | -4.19 | 0.84 | 0.44 | 0.01 |
| TCGT | YA | 0.07 | 0.04 | -0.05 | 0.01 | -2.37 | 0.86 | -0.74 | 0.40 |
| TCCC | WT, E | -2.46 | 0.96 | -2.82 | 0.94 | -0.40 | 0.32 | -1.23 | 0.84 |
| CGTT | WT, YA | -2.28 | 0.95 | -1.68 | 0.72 | -2.06 | 0.80 | -1.01 | 0.83 |
| CAGT | WT, YA | -2.24 | 0.89 | -0.75 | 0.63 | -2.91 | 0.98 | -0.90 | 0.47 |
| TCTC | WT, YA | -2.51 | 0.84 | -1.75 | 0.63 | -2.66 | 0.88 | -0.91 | 0.46 |
| CATC | WT, YA, E | -3.04 | 0.83 | -2.03 | 0.65 | -2.59 | 0.79 | -1.43 | 0.45 |
| CCTT | WT, YA, E | -2.72 | 0.81 | -2.77 | 0.86 | -2.49 | 0.87 | -1.62 | 0.87 |
| CGTC | WT, YA, E | -2.54 | 0.95 | -2.61 | 0.82 | -2.41 | 0.86 | -1.64 | 0.79 |

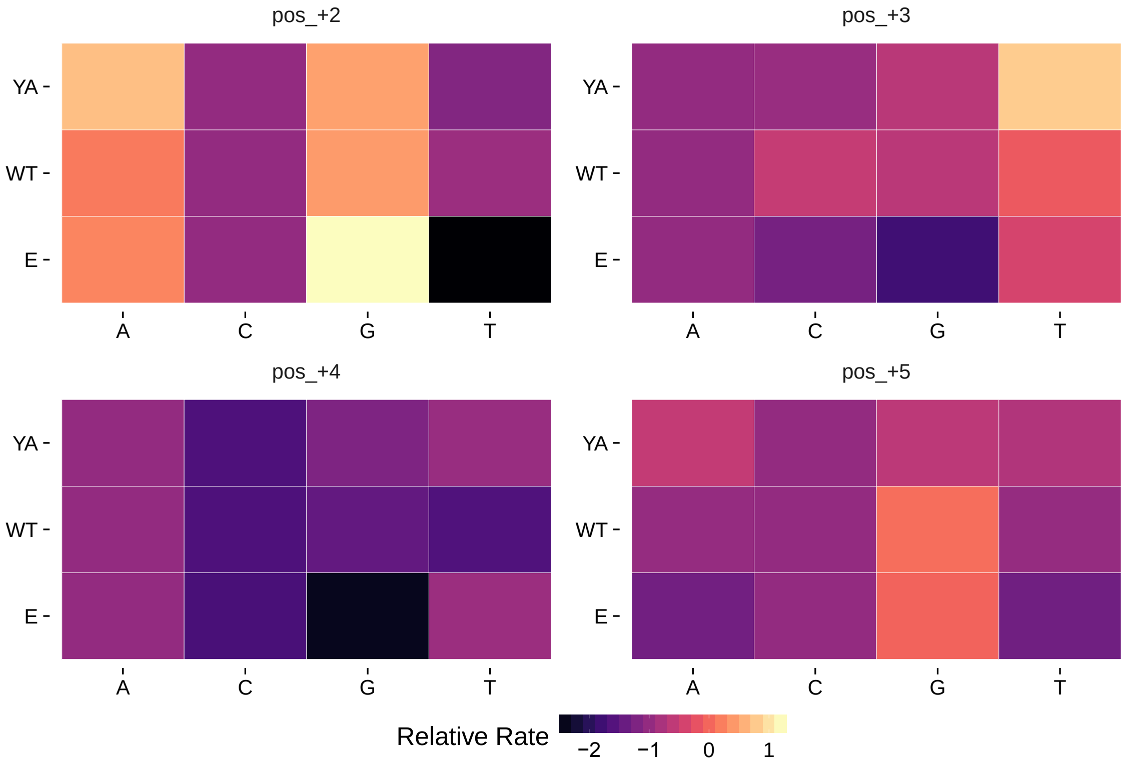

**Figure 13. Heat map of I-OnuI WT, E and YA nuclease activity on +2 to +5 DNA point mutants.** Rates of change for WT (CAAC) and point mutant nt substrates for positions +2 – +5 (pos_+2, pos_+3, pos_+4 and pos_+5) were used to build this heat map. The rate of change for each sequence was divided by their respective nucleases rate of change value calculated on the WT substrate (CAAC) to give a relative rate of change for point mutant substrates.

# CHAPTER FOUR - DISCUSSION

## 4.1 MI as a technique for characterizing protein-DNA interactions

Here we apply a mathematical analysis to a MSA of LHEs with their mapped DNA target sites. This analysis, characterizing pairwise co-variation between MSA columns, identified an aa-nt pair with abnormally high MI. Using I-OnuI as a representative of the MSA, our analysis identified aberrant co-variation between I-OnuI-K231 and the A3 nt of its DNA substrate (Table 1). Randomizing K231 in combination with a local aa D240, resulted in I-OnuI variants that had altered DNA specificity compared to I-OnuI WT (Fig. 8; Fig. 11). I-OnuI-YA's specificity profile was shown to be distinctive from I-OnuI WT while I-OnuI-E appeared to reduce DNA specificity (Table 4). We believe that aa 231 and 240 contribute to an interaction surface governing substrate recognition, in part, at +2 and +5 nt positions (Fig. 13). These results encourage us to accept that positions in a MSA with extreme co-variation can reveal important 3-dimensional interactions that can be targeted for reprogramming efforts. In conclusion, MIp analysis was successfully used to reprogram LHE DNA specificity.

Assessing solved LHE contact maps in light of our MI results reveals discrepancies. Some solved crystal structures agreed that the 231$^{st}$ residue directly contacts the +3 nt, while other crystal structures disagreed with our predictions, suggesting that the +3 nt position was specifically contacted by another aa residue. Specifically calling attention to the contrast between I-OnuI and I-LtrI contact maps (Takecuhi *et al.*, 2011), I-LtrI coincided with our MI analysis, resolving that aa 231 directly contacts the +3 nt substrate. In contrast, the I-OnuI contact map designates aa 231 as directly contacting the +5 nt. Without our co-variation analysis, efforts to reengineer I-OnuI binding at the +3 nt position would have been directed to T203, possibly unable to restructure the specificity at the +3 DNA nt. LHEs have been described as rapidly evolving proteins that have little evolutionary pressure maintaining specific protein-DNA contacts (Lucas *et al.*, 2001). Thus, this co-variation analysis may have identified a variable aa-nt contact that is utilized by some LHEs, like I-LtrI, but has been restructured in other LHEs, like I-OnuI. This would mean that our I-OnuI mutant nt contacts have been restructured to match a distinct evolutionary trajectory taken by homologous LHEs like I-LtrI. However, further investigation into the relationship between I-OnuI 231 and the +5 nt position is required to robustly draw this conclusion.

Takeuchi *et al.* (2011) previously characterized I-OnuI's DNA specificity and its tolerance to nt point mutations. This study revealed I-OnuI to have approximately 25 % relative activity on A3G and A3C nt substitutions compared to its activity on the WT A3 substrate. Furthermore, Takeuchi *et al.* (2011) found I-OnuI to have approximately 10 % relative activity on the A3T substrate. These findings are in accordance with data presented here, showing I-OnuI-WT to have appreciable activity on A3, A3C and A3G substrates (Fig. 6; Fig. 9). Additional observations from Takeuchi *et al.* (2011) showed I-OnuI tolerance to nt substitutions at the +4 nt position. The lack of discrimination against the +4 nt has been attributed to the process of HEs developing nt specificity. HEs target essential genes to ensure conservation of their target sequences and maximize the efficiency of homing. Furthermore, HEs contact strongly conserved nts that contribute essential structural/functional features to the gene. The +4 nt position within the I-OnuI target site is a wobble position (Takeuchi *et al.*, 2011), deterring I-OnuI from strongly recognizing this nt, as it would be poorly conserved (Edgell *et al.*, 2004; Scalley-Kim *et al.*, 2007). Notably, Takeuchi *et al.* (2011) also showed that the C5G nt substitution was detrimental to I-OnuI-WT activity, coinciding with our findings (Fig. 13).

Mutation of I-OnuI at the 231st and 240th positions altered DNA specificity and activity distinct from the WT I-OnuI protein using *in vivo* and *in vitro* assays. Illumina sequencing results showed that the I-OnuI-YA mutant was able to target novel substrates at +3 and +5 positions (Table 4; Fig. 5). Contrastingly, WT I-OnuI and I-OnuI-E seem to be very similar regarding their specificity profiles. This, along with *in-vitro* and *in-vivo* cleavage data (Fig. 8; Fig. 11; Fig. 13) suggests that I-Onu-YA has an altered specificity profile compared to WT I-OnuI, whereas I-OnuI-E is an increased activity mutant of the WT protein. This study used assays that measured cleaved substrate as an indication for binding between LHEs and their DNA substrate. To resolve the changes I-OnuI mutations may have on binding and cleavage, Electric Mobility Shift Assays (EMSAs) should be done. EMSAs should be completed in suboptimal salt conditions to ensure LHEs are binding without cleaving substrate.

Scientists who have previously applied this co-variation analysis suggest that these analyses should be conducted on alignments containing more than 100 sequences (Mahony *et al*., 2007, Dickson *et al*., 2010). MI studies completed by Mahony *et al*. (2007) used alignments containing more than 1000 protein-DNA pairs; the MSA used in this analysis was limited to the number of experimentally determined LHE target sequences. Our co-variation analysis was able to successfully take sequence alignment information, using a modest 27 LHE-DNA pairs, to identify an aa necessary for specifically interacting with the +3 nt. The unexpected success of this study may be attributed to the extensive variability within LHEs and their respective DNA substrates. Variability within the MSA allows co-variation analysis to robustly characterize meaningful dependencies between alignment positions. In summary, results presented here along with those conducted by Mahony *et al*. (2007) were able to use MI analysis of sequence information to pinpoint protein positions that contribute specific intermolecular contacts in 3-dimensional space. MI is suitable as a preliminary analysis to localize mutational efforts aimed at restructuring interaction specificity of molecules.

## 4.2 Comparing LHE reprogramming results with previous findings

Previous studies characterizing homodimeric LHE DNA specificity identified a homologous network of amino acids identified in this study. The homodimer I-CreI was mutated at residues 70 and 75, homologous to monomeric LHE positions 231 and 240, to alter specificity at the ± 3 DNA positions. Specifically, I-CreI Q44A-R70L-D75N and R68A-R70N-D75N mutants were able to accommodate C3T substitutions. Although they did not isolate specific amino acid mutations identified in our study, they found that modulating homologous residues in I-CreI within the homologous beta-sheet were sufficient to alter recognition of nts at ± 3 DNA positions (Arnould *et al*., 2006).

I-OnuI itself has been reengineered to recognize novel substrates. Takeuchi *et al*. (2011) reengineered I-OnuI recognition at 5 nt positions, -11, -10, -4, +2 and +11, to recognize a malignant human gene. N32S, S35R, S40A, T48C, I51N, K80R, E178D, K189N and K229R were the aa substitutions made to accommodate the DNA substrate. All these amino acid mutations are distinct from those identified in this study.

Having worked extensively on reengineering LHEs to recognize novel substrates, Barry Stoddard has identified amino acid modules that recognize stretches of nts. The identified module restructuring I-OnuI recognition at +3 - +5 positions includes K231 and D240 among 6 additional residues. Altering K231 and D240 residues to reprogram I-OnuI specificity at +3 nt, drastically reduces the library complexity to be screened and increases the efficiency of reprogramming efforts.

## 4.3 Future directions

### 4.3.1 Enhancing co-variation analysis

The most significant improvement that could be made to this analysis would be to add LHE-DNA pairs to the MSA. An increased number of LHE-DNA pairs will give the co-variation analysis more power to identify significant co-dependencies. Furthermore, including additional parameters that accurately assess LHEs ability to indirectly readout DNA features would also improve this analysis (Molina *et al.*, 2012; Thyme *et al.*, 2014). Tuning the MI statistic to biological data has been essential to the success of this procedure; further corrections to the MIp statistic to better enable its characterization of biological data is of interest.

### 4.3.2 Additional applications of MI

Working to synthesize genetically modified organisms or accurate disease models can greatly benefit from the ease of CRISPRs, however, it is paramount that genome-editing reagents display stringent specificity to be suitable in clinical use. In a post CRISPR era, LHEs utility for genome editing may not be realized until the limits of CRISPR specificity have been well characterized.

Here we show how computational techniques, specifically co-variation analysis, can identify residues that modulate interaction specificity. Knowledge gained from these computational techniques comes from relatively small amounts of biological data and can greatly reduce uncertainty when initiating study of a biological interaction.. Furthermore, this technique could also be used used to disrupt contacts that hinder utility of a genome editing reagent. Co-variation analysis of the Cas9-CRISPR system could identify residues that govern PAM specificity and possibly alleviate this restriction (Kleinstiver *et al.*, 2015).

Moreover, with the onslaught of the –omics datasets, many view interpretation as the greatest impairment. Interpretation can be cumbersome because of our limited understanding of the intricate genetic networks behind complex cellular processes. Identifying co-variation between transcript expression profiles within the proteome could be useful to identify networks of enzymes underlying expression systems or scaffolding complexes.

# References

---

Arnould, S., Chames, P., Perez, C., Lacroix, E., Duclert, A., Epinat, J.-C., Stricher, F., Petit, A.-S., Patin, A., Guillier, S., *et al*. (2006). Engineering of Large Numbers of Highly Specific Homing Endonucleases that Induce Recombination on Novel DNA Targets. Journal of Molecular Biology *355*, 443–458.

Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Stoddard, B.L., and Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. Nature *441*, 656–659.

Ashworth, J., Taylor, G.K., Havranek, J.J., Quadri, S.A., Stoddard, B.L., and Baker, D. (2010). Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. Nucl. Acids Res. *38*, 5601–5608.

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., and Dress, A.W. (2000). Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. Mol Biol Evol *17*, 164–178.

Barzel, A., Privman, E., Peeri, M., Naor, A., Shachar, E., Burstein, D., Lazary, R., Gophna, U., Pupko, T., and Kupiec, M. (2011). Native homing endonucleases can target conserved genes in humans and in animal models. Nucleic Acids Res *39*, 6646–6659.

Beerli, R.R., Segal, D.J., Dreier, B., and Barbas, C.F. (1998). Toward controlling gene expression at will: Specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. PNAS *95*, 14628–14633.

Blancafort, P., Magnenat, L., and Barbas, C.F. (2003). Scanning the human genome with combinatorial transcription factor libraries. Nat Biotech *21*, 269–274.

Boutell, J.M., Thomas, P., Neal, J.W., Weston, V.J., Duce, J., Harper, P.S., and Lesley Jones, A. (1999). Aberrant Interactions of Transcriptional Repressor Proteins with the Huntington's Disease Gene Product, Huntingtin. Human Molecular Genetics *8*, 1647–1655.

Chevalier, B.S., and Stoddard, B.L. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res *29*, 3757–3774.

Clarke, N.D. (1995). Covariation of residues in the homeodomain sequence family. Protein Science *4*, 2269–2278.

Dickson, R.J., and Gloor, G.B. (2012). Protein Sequence Alignment Analysis by Local Covariation: Coevolution Statistics Detect Benchmark Alignment Errors. PLOS ONE *7*, e37645.

Dickson, R.J., Wahl, L.M., Fernandes, A.D., and Gloor, G.B. (2010). Identifying and Seeing beyond Multiple Sequence Alignment Errors Using Intra-Molecular Protein Covariation. PLOS ONE *5*, e11082.

Dickson, R.J., and Gloor, G.B. (2013). The MIp Toolset: an efficient algorithm for calculating Mutual Information in protein alignment. arXiv preprint arXiv:1304.4573.

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

Dreier, B., Segal, D.J., and Barbas III, C.F. (2000). Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains1. Journal of Molecular Biology *303*, 489–502.

Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D., and Barbas, C.F. (2001). Development of Zinc Finger Domains for Recognition of the 5′-ANN-3′ Family of DNA Sequences and Their Use in the Construction of Artificial Transcription Factors. J. Biol. Chem. *276*, 29466–29478.

Dreier, B., Fuller, R.P., Segal, D.J., Lund, C.V., Blancafort, P., Huber, A., Koksch, B., and Barbas, C.F. (2005). Development of Zinc Finger Domains for Recognition of the 5′-CNN-3′ Family DNA Sequences and Their Use in the Construction of Artificial Transcription Factors. J. Biol. Chem. *280*, 35588–35597.

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

Edgell, D.R., Stanger, M.J., and Belfort, M. (2004). Coincidence of Cleavage Sites of Intron Endonuclease I-TevI and Critical Sequences of the Host Thymidylate Synthase Gene. Journal of Molecular Biology *343*, 1231–1241.

Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. Biochemistry *44*, 7156–7165.

Havranek, J.J., Duarte, C.M., and Baker, D. (2004). A Simple Physical Model for the Prediction and Design of Protein–DNA Interactions. Journal of Molecular Biology *344*, 59–70.

Jiménez, J.S. (2010). Protein-DNA interaction at the origin of neurological diseases: a hypothesis. J. Alzheimers Dis. *22*, 375–391.

Jurica, M.S., and Stoddard, B.L. (1999). Homing endonucleases: structure, function and evolution. CMLS, Cell. Mol. Life Sci. *55*, 1304–1326.

Kim, J.-S., and Pabo, C.O. (1997). Transcriptional Repression by Zinc Finger Peptides EXPLORING THE POTENTIAL FOR APPLICATIONS IN GENE THERAPY. J. Biol. Chem. *272*, 29795–29800.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., et al. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133–152.

Li, S., and Bradley, P. (2013). Probing the role of interfacial waters in protein–DNA recognition using a hybrid implicit/explicit solvation model. Proteins *81*, 1318–1329.

Little, D.Y., and Chen, L. (2009). Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination in Protein Evolution. PLOS ONE *4*, e4762.

Liu, Q., Segal, D.J., Ghiara, J.B., and Barbas, C.F. (1997). Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. PNAS *94*, 5525–5530.

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

Mahony, S., Auron, P.E., and Benos, P.V. (2007). Inferring protein–DNA dependencies using motif alignments and mutual information. Bioinformatics *23*, i297–i304.

McMurrough, T.A., Dickson, R.J., Thibert, S.M.F., Gloor, G.B., and Edgell, D.R. (2014). Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. PNAS *111*, E2376–E2383.

Mitchell, P., & Tjian, R. (1989). Transcriptional Regulation in Mammalian Cells by Sequence-Specific DNA Binding Proteins. *Science*, *245*(4916), 371-378. Retrieved from http://www.jstor.org/stable/1703794

Molina, R., Redondo, P., Stella, S., Marenchino, M., D'Abramo, M., Gervasio, F.L., Epinat, J.C., Valton, J., Grizot, S., Duchateau, P., *et al*. (2012). Non-specific protein–DNA interactions control I-CreI target binding and cleavage. Nucl. Acids Res. *40*, 6936–6945.

Muller, P.A.J., and Vousden, K.H. (2013). p53 mutations in cancer. Nat Cell Biol *15*, 2–8.

Oliveira, L., Paiva, A.C.M., and Vriend, G. (2002). Correlated Mutation Analyses on Very Large Sequence Families. ChemBioChem *3*, 1010–1017.

Pabo, C.O., and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?1. Journal of Molecular Biology *301*, 597–624.

Pabo, C.O., Peisach, E., and Grant, R.A. (2001). Design and Selection of Novel Cys2His2 Zinc Finger Proteins. Annual Review of Biochemistry *70*, 313–340.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al*. (2000). Genome-Wide Location and Function of DNA Binding Proteins. Science *290*, 2306–2309.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein Structure Prediction Using Rosetta. In Methods in Enzymology, (Elsevier), pp. 66–93.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

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

Segal, D.J., and Barbas III, C.F. (2000). Design of novel sequence-specific DNA-binding proteins. Current Opinion in Chemical Biology *4*, 34–39.

Seligman, L.M., Chisholm, K.M., Chevalier, B.S., Chadsey, M.S., Edwards, S.T., Savage, J.H., and Veillet, A.L. (2002). Mutations altering the cleavage specificity of a homing endonuclease. Nucleic Acids Res *30*, 3870–3879.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. *24*, 814–826.

Stoddard, B.L. (2005). Homing endonuclease structure and function. Quarterly Reviews of Biophysics *38*, 49–95.

Stoddard, B.L. (2011). Homing Endonucleases: From Microbial Genetic Invaders to Reagents for Targeted DNA Modification. Structure *19*, 7–15.

Sussman, D., Chadsey, M., Fauce, S., Engel, A., Bruett, A., Monnat Jr, R., Stoddard, B.L., and Seligman, L.M. (2004). Isolation and Characterization of New Homing Endonuclease Specificities at Individual Target Site Positions. Journal of Molecular Biology *342*, 31–41.

Takeuchi, R., Lambert, A.R., Mak, A.N.-S., Jacoby, K., Dickson, R.J., Gloor, G.B., Scharenberg, A.M., Edgell, D.R., and Stoddard, B.L. (2011). Tapping natural reservoirs of homing endonucleases for targeted gene modification. Proc Natl Acad Sci U S A *108*, 13077–13082.

Thyme, S.B., Jarjour, J., Takeuchi, R., Havranek, J.J., Ashworth, J., Scharenberg, A.M., Stoddard, B.L., and Baker, D. (2009). Exploitation of binding energy for catalysis and design. Nature *461*, 1300–1304.

Thyme, S.B., Baker, D., and Bradley, P. (2012). Improved Modeling of Side-Chain–Base Interactions and Plasticity in Protein–DNA Interface Design. Journal of Molecular Biology *419*, 255–274.

Thyme, S.B., Song, Y., Brunette, T.J., Szeto, M.D., Kusak, L., Bradley, P., and Baker, D. (2014a). Massively parallel determination and modeling of endonuclease substrate specificity. Nucl. Acids Res. gku1096.

Thyme, S.B., Boissel, S.J.S., Quadri, S.A., Nolan, T., Baker, D.A., Park, R.U., Kusak, L., Ashworth, J., and Baker, D. (2014b). Reprogramming homing endonuclease specificity through computational design and directed evolution. Nucl. Acids Res. *42*, 2564–2576.

Tillier, E.R.M., and Lui, T.W.H. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics *19*, 750–755.

Yanover, C., and Bradley, P. (2011). Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. Nucl. Acids Res. *39*, 4564–4576.

Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer *9*, 798–809.

# Curriculum Vitae

---

## EDUCATION

**Master of Science** – *Biochemistry with bioinformatics component* **2014 – Summer 2016**

*Western University, London, Ontario*

- Performed literature reviews and stayed current on pertinent literature
- Mined databases for biological data of interest to construct custom datasets
- Applied predetermined and ad-hoc analyses on datasets in R and Unix environments
- Regularly performed common molecular biology procedures such as DNA purification, PCR, gel electrophoresis and sample preparation for sequencing
- Text parsed and visualized Next-Generation Sequencing (NGS) datasets in R
- Maintained meticulous laboratory notes
- Followed standard operating and quality assurance procedures
- Mentored volunteers and undergraduate students

**Bachelor of Science** – *Honors Specialization in Biochemistry and Cell Biology* **2010 – 2014**

*Western University, London, Ontario*

- Gold Medal recipient (highest GPA in major from graduating class)
- Fourth year thesis & biochemistry laboratory – PCR, cloning, DNA preps, analytical digests, mass spectroscopy, spectrophotometry and protein purification
- Cell biology laboratory – advanced microscopy, karyotyping and cell blotting/staining
- Microbiology laboratory – unknown bacterial identification and microbial analysis of body fluids

## WORK EXPERIENCE

**Teaching Assistant (TA)** **Fall/Winter 2015 – 2016**

*Western University, London, Ontario*

- Biochemical regulation (Fall 2015) – 3rd year Biochemistry
    - Led weekly tutorials, marked assessments and answered student questions
    - Content included old, current and emerging sequencing technologies
    - Content also included DNA replication, regulation of gene expression, epigenetics, molecular and synthetic biology
- Biochemistry Laboratory (Winter 2015) – 3rd year Biochemistry
    - Led a group of students during their weekly laboratories
    - Acquainted group with common biochemical and molecular biology procedures
    - Helped students analyze experimental results generated from DNA sequencing, PCR, gel electrophoresis and DNA profiles generated from saliva samples

**Campus Program Coordinator – Leave the Pack Behind** **Fall/Winter 2013 – 2014**

*Western University, London, Ontario*

- Managed a team of 30 volunteers
- Coordinated campus wide smoking cessation campaigns to improve community health
- Liaison between campus health services and LTPB head office personal

**Assistant Aquatic Coordinator / Lifeguard** **2008 – 2012**

*Town of Halton Hills, Georgetown, Ontario*

Managed staff in the operation of a public swimming pool, ensuring adherence to regulations and the safety of visiting patrons.