January 2018

Coevolving Residues and the Expansion of Substrate Permissibility in LAGLIDAG Homing Endonucleases

Thomas A. McMurrough
The University of Western Ontario

Supervisor
Edgell, David R.
The University of Western Ontario

Joint Supervisor
Gloor, Gregory B.
The University of Western Ontario

Graduate Program in Biochemistry

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

© Thomas A. McMurrough 2017

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

Part of the Medical Molecular Biology Commons

Recommended Citation
https://ir.lib.uwo.ca/etd/5119

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 tadam@uwo.ca, wlswadmin@uwo.ca.
Abstract

Genome-editing (GE) is a form of genetic engineering that permits the deliberate manipulation of genetic material for the study of biological processes, agricultural and industrial biotechnologies, and developing targeted therapies to cure human disease. While the potential application of GE is wide-ranging, the efficacy of most strategies is dependent upon the ability to accurately introduce a double-stranded break at the genomic location where alterations are desired. LAGLIDADG homing endonucleases (LHEs) are a class of mobile genetic element that recognize and cleave 22-bp sequences of DNA. Given this high degree of specificity, LHEs are powerful GE reagents, but re-engineering their recognition sites has been hindered by a limited understanding of structural constraints within the family, and how cleavage specificity is regulated in the central target site region.

In the present studies, a covariation analysis of the LHE family recognized a set of coevolving residues within the enzyme active site. These positions were found to modulate catalytic efficiency, and are thought to create a barrier to active site evolution and re-engineering by constraining the LHE fitness landscape towards a set of functionally permissive combinations. Interestingly, mutation of these positions led to the identification of a catalytic residue variant that demonstrates cleavage activity against a greater number of central target site substrates than wild-type enzymes. To facilitate these investigations, high-throughput and unbiased methods were developed to functionally screen large mutagenic libraries and simultaneously profile cleavage specificity against 256 different substrates. Lastly, structural studies aimed at increasing our understanding of the LHE coevolving network led to the discovery of direct protein-DNA contacts in the central target site region.

Significantly, these findings increase our understanding of functionally important structural constraints within the LHE family, and have the potential to increase the sequence targeting capacity of LHE scaffolds. More broadly, the methodologies described in this thesis can assist large-scale structure-function studies and facilitate
investigations of substrate specificity for most DNA-binding proteins. Finally, the thorough biochemical validation I provide for computational predictions of coevolution showcases a strategy to infer protein function-structure from genetic information, and emphasizes the need to expand these studies to other protein families.

Keywords
LAGLIDADG homing endonuclease, genome-editing, coevolution, enzyme catalysis, substrate specificity, protein engineering, biotechnology
Co-Authorship Statement

For Chapters 2, 3, and 4, Thomas McMurrough performed the research, with exceptions noted below. Thomas McMurrough, David Edgell and Greg Gloor designed the research, contributed to new reagents and analytical tools, and analyzed the data.

Chapter 2 – Russell Dickson and Greg Gloor generated the LAGLIDADG family multiple sequence alignment and performed the covariation analysis, Greg Gloor and David Edgell completed the phylogenetic analysis, Russell Dickson performed the Rosetta modelling and contributed to experimental design, Stephanie M.F. Thibert contributed to performing the bacterial two-plasmid genetic selections and analyzing data, staff from the London Regional Proteomics Centre made significant contributions to performing the differential scanning calorimetry, and David Edgell and Greg Gloor wrote the manuscript with contributions from Thomas McMurrough and Russell Dickson.

Chapter 4 – Kun Zhang made significant contributions to protein purification and crystallography screens, and Chris Brown and Murray Junop collected and analyzed the diffraction data, and generated the structural models.
Acknowledgments

I would like to thank Dr. David Edgell and Dr. Greg Gloor for the opportunities they have afforded me and the exquisite mentorship they have provided during my graduate studies. Through all my follies as a student your support was unwavering, and I am truly grateful for having you both as my supervisors.

Next, I would like to acknowledge my advisors Dr. Stan Dunn and Dr. Chris Brandl. Your advice and friendship over the years were invaluable and I am thankful for you always providing sound advice and guidance.

Dr. Benjamin Kleinstiver and Dr. Jason Wolfs! Life in the lab would not have been the same without you, and I have missed you both dearly. I have a tremendous amount of respect for you both, and your work ethics are a model to us all. Best. Wet-Labmates. Ever. I would also like to thank my Dry-Labmate Dr. Jean Macklaim for providing significant computation assistance, and your compassion through the roller-coaster ride that was my graduate career.

I am very thankfully for also having met Dr. Henry Dunn, Dr. Stephanie Dorman, Dr. Michael Ellis, Eddy Cruise, and the many other people who had a positive influence on my time at Western. I hope we remain in contact, and I wish nothing but the best for you all.

I am so thankful that I met Carolina Santos after she came to Western from Brazil. If we didn’t peruse our doctorates then we would not have crossed paths, and life would not be the same. I can’t wait to finish grad school and start our life together! Carolina, eu te amo!

Finally, and especially, I would like to thank my amazing family for their love and support during grad school, and at every other moment of my life. I am so fortunate to be supported by such a loving family. Mom and Dad, you don’t need to be concerned about your future because you’ve earned a view of the pond from your nursing home 10-times over!!
Table of Contents

Abstract ........................................................................................................................................... i
Co-Authorship Statement .................................................................................................................... iii
Acknowledgments ............................................................................................................................... iv
Table of Contents ............................................................................................................................... v
List of Tables ......................................................................................................................................... x
List of Figures ......................................................................................................................................... xi
List of Appendices .............................................................................................................................. xiv
List of Abbreviations and Symbols .................................................................................................... xv
Chapter 1 .............................................................................................................................................. 1

1 Introduction ......................................................................................................................................... 1

1.1 Genetic Manipulation and Genome-Editing (GE) ............................................................................. 1
  1.1.1 Historical Perspective ................................................................................................................. 1
  1.1.2 Genome-Editing (GE) Strategies .............................................................................................. 2

1.2 Precision Genome-Editing (GE) Reagents ...................................................................................... 5
  1.2.1 Homing Endonucleases ............................................................................................................. 5
  1.2.2 CRISPR/Cas Systems ............................................................................................................... 6
  1.2.3 Synthetic Genome-Editing Reagents ....................................................................................... 8

1.3 LAGLIDADG Homing Endonucleases (LHEs) ............................................................................. 9
  1.3.1 General Characteristics ........................................................................................................... 9
  1.3.2 The LAGLIDADG Interface .................................................................................................... 11
  1.3.3 Catalytic Mechanism and the Role of Divalent Metal Cofactors ........................................... 13
  1.3.4 DNA Target Recognition and Binding .................................................................................... 15
     1.3.4.1 Direct and Water-Mediated Substrate Readout ................................................................. 15
     1.3.4.2 Indirect or Shape-based Substrate Readout ................................................................. 17
1.3.5 Applying LHEs as Genome-Editing (GE) Reagents ........................................ 19
  1.3.5.1 Re-Engineering of the LHE Protein-DNA Interface ........................... 19
  1.3.5.2 Domain-fusion Chimeras (DFCs).................................................... 21
1.3.6 Advantages and Disadvantages of LHEs for Genome-Editing.................. 22
1.4 Amino Acid Coevolution and Intramolecular Covariation Analyses .......... 24
  1.4.1 Amino Acid Coevolution..................................................................... 24
  1.4.2 Intramolecular Covariation Analyses.................................................. 24
1.5 Summary..................................................................................................... 27
1.6 References.................................................................................................... 30

Chapter 2 ......................................................................................................... 39

2 Control of Catalytic Efficiency by a Coevolving Network of Catalytic and
  Noncatalytic Residues.................................................................................... 39
  2.1 Significance............................................................................................... 39
  2.2 Introduction............................................................................................... 39
  2.3 Materials and Methods .......................................................................... 41
    2.3.1 Oligonucleotides and Plasmids.......................................................... 41
    2.3.2 Bacterial Two-Plasmid Functional Selection....................................... 41
    2.3.3 Mutant Library Synthesis and Screening........................................... 42
    2.3.4 BioScreen Bacterial Growth Assay...................................................... 43
    2.3.5 Protein Expression and Purification.................................................... 43
    2.3.6 Differential Scanning Calorimetry...................................................... 44
    2.3.7 In Vitro Cleavage Assays and Single-Turnover Kinetics...................... 44
    2.3.8 I-Lrtl A28S Suppressor Screen.......................................................... 45
    2.3.9 Rosetta Modeling............................................................................... 45
  2.4 Results and Discussion ........................................................................... 46
2.4.1 Identification of a Coevolving Network of Catalytic and Noncatalytic Residues .............................................................................................................. 46

2.4.2 Saturating Mutagenesis of the Coevolving Network Recapitulates Phylogenetic Patterns ......................................................................................... 52

2.4.3 I-LtrI Variants with Suboptimal Networks Grow Slowly or Not at All ... 56

2.4.4 Coevolving Network Modulates Enzymatic Activity ................................................................................................................................. 59

2.4.5 Suppressor Mutations in the Coevolving Network Can Rescue Suboptimal Variants ......................................................................................... 63

2.4.6 Implications for Engineering LHEs ................................................................................................................................. 65

2.4.7 Conclusions ................................................................................................................................................................................. 68

2.5 References ........................................................................................................................................................................... 69

Chapter 3 .............................................................................................................................................................................. 73

3 The Existence of Additional Coevolving Networks Between LAGLIDADG Domains, and Attempts to Increase Cleavage Activity of the Domain-Fusion Chimera “LtrOnu” ........................................................................................................... 73

3.1 Introduction ........................................................................................................................................................................ 73

3.2 Materials and Methods .................................................................................................................................................................... 74

3.2.1 Oligonucleotides and Plasmids ............................................................................................................................................. 74

3.2.2 Bacterial Two-Plasmid Functional Selection ........................................................................................................................ 75

3.2.3 Mutant Library Synthesis .................................................................................................................................................. 75

3.2.4 Protein Expression, Protein Purification, and In Vitro Cleavage Assays ....................................................................................... 76

3.2.5 LtrOnu Library Screens .................................................................................................................................................. 76

3.3 Results ................................................................................................................................................................................. 76

3.3.1 Additional Covarying Networks are Tolerant to Semi-Conservative Substitutions ......................................................................................... 76

3.3.2 LtrOnu Shows Detectable Cleavage Activity But its Stability is Compromised ............................................................................................................. 82

3.3.3 Saturating Covarying Residue Positions Cannot Increase the Cleavage Activity of LtrOnu ........................................................................................................... 84
3.3.4 Additional Mutagenesis Strategies Cannot Increase the Cleavage Activity of LtrOnu .............................................................. 87

3.4 Discussion ......................................................................................................................... 90

3.5 References .......................................................................................................................... 92

Chapter 4 .................................................................................................................................. 94

4 Catalytic Residue Substitutions Increase LAGLIDADG Homing Endonuclease Cleavage Activity Against Non-Cognate Central Four (C4) Substrates ................................................................. 94

4.1 Introduction ......................................................................................................................... 94

4.2 Materials and Methods ..................................................................................................... 97

  4.2.1 Oligo Nucleotides and Plasmids Protein Libraries ..................................................... 97

  4.2.2 Mutagenic Libraries ..................................................................................................... 97

  4.2.3 Bacterial Two-Plasmid Functional Selection .............................................................. 98

  4.2.4 Protein Expression and Purification ............................................................................. 98

  4.2.5 Endonuclease Assays and Pseudo-Michaelis-Menten Kinetics .................................. 99

  4.2.6 Product Enrichment and Substrate Depletion Assays ................................................. 99

  4.2.7 High-Throughput Sequencing (HTS) and Compositional Data Analysis (CoDa) .... 100

  4.2.8 Structural Analysis ...................................................................................................... 101

4.3 Results ............................................................................................................................... 102

  4.3.1 An in vivo Selection Identifies LHE Variants Active on Substrates with C4 Substitutions ......................................................................................... 102

  4.3.2 Simultaneous Activity Assays on All 256 C4 Substrates Identifies LHE Cleavage Preferences ........................................................................................... 110

  4.3.3 Increases in $k_{\text{cat}}^*$ Allow I-LtrI AEGD to Overcome Defects in Binding and Cleave AACC with Greater Efficiency than Wild-type AEGE ...... 120

  4.3.4 Base Pair Opening Predicts I-Ltrl Catalytic Activity on Cognate and Non-Cognate C4 Substrates ..................................................................................... 124

  4.3.5 Direct Contacts to C4 Nucleobases are Present in LHE Pre-Cleavage Structures ......................................................................................................................... 131
4.4 Discussion .................................................................................................................. 134

4.5 References ............................................................................................................... 139

Chapter 5 ......................................................................................................................... 142

5 Discussion .................................................................................................................... 142

5.1 Covariation Analyses Identify Biologically Important Amino Acid Networks Within Protein Families .............................................................................................................. 142

5.2 Applying Covariation Analysis to Investigate the Genetic Basis of Human Disease ........................................................................................................................................... 144

5.3 Coevolving Network Identity Must Be Taken into Consideration When Re-Engineering LADLIDADG Homing Endonuclease (LHE) Scaffolds .......... 145

5.4 E184D Substitutions Increase LADLIDADG Homing Endonuclease (LHE) Cleavage Activity on Non-Cognate Central Four (C4) Substrates ................ 149

5.5 Nucleotide Identity in the Central Four (C4) Constrains LADLIDADG Homing Endonuclease (LHE) Cleavage Activity, and Base Pair Opening is Favorable to Catalysis .................................................................................................................. 150

5.6 Some LADLIDADG Homing Endonuclease (LHE) Family Members Make Direct Contacts to Central Four (C4) Base Pairs ........................................................................... 152

5.7 Considerations for Re-Engineering LADLIDADG Homing Endonuclease (LHE) Cleavage Specificity and Final Thoughts ........................................ 156

5.8 References ............................................................................................................... 158

Appendices ..................................................................................................................... 161

Curriculum Vitae ............................................................................................................ 183
List of Tables

Table 3.1: Covarying residue combinations that are created in LHE domain fusion chimeras. .................................................................................................................................................. 78

Table 4.1: Nucleotide frequency in the top quartile of central four (C4) substrates for I-LtrI and I-OnuI.................................................................................................................................................. 116

Table 4.2: I-LtrI variant pseudo-Michaelis-Menten kinetics against the cognate (ATAC) and non-cognate (AACC) central four substrates........................................................................................................... 123

Table 4.3: Rate constants for I-LtrI variant cleavage of cognate (ATAC) and non-cognate (AACC) central four substrates. .................................................................................................................................. 123

Table 4.4: Predicted direct contacts to central four base pairs in pre-cleavage LHE structures............................................................................................................................................................................ 133
List of Figures

Figure 1. 1: General genome editing strategies ................................................................. 4

Figure 1. 2: LHEs share a common secondary structure arrangement .......................... 10

Figure 1. 3: The LAGLIDADG helices. .............................................................................. 12

Figure 1. 4: I-LtrI target site contact map ........................................................................ 16

Figure 1. 5: Direct base contacts are modulated by multiple residue positions .......... 20

Figure 1. 6: General domain-fusion chimera (DFC) engineering strategy ................... 23

Figure 2.1. a: LAGLIDADG homing endonuclease (LHE) family tree ..................... 48

Figure 2.1. b: Covarying residues of the LHE family .................................................. 49

Figure 2.1. c: Structure of the LHE coevolving network ............................................. 50

Figure 2.1. d: Naturally occurring coevolving network combinations ....................... 51

Figure 2.2. a: Schematic of liquid competition growth experiments .......................... 54

Figure 2.2. b: Coevolving network residue preference ................................................. 55

Figure 2.3. a: Survival of individual variants under genetic selection ....................... 57

Figure 2.3. b: I-LtrI variants show differences in growth rate ..................................... 58

Figure 2.4. a: Coevolving network variants are thermostable ....................................... 60

Figure 2.4. b: Pseudo-Michaelis-Menten parameters for individual I-LtrI variants ...... 61

Figure 2.4. c: Catalytically efficient variants grow faster under genetic selection .......... 62

Figure 2.5: A suppressor screen identifies residues in the coevolving network .......... 64

Figure 2.6: Coevolving residues create a barrier to active-site evolution in LHEs ....... 67
Figure 3. 1: LHE residues with the second highest covariation score......................... 77
Figure 3. 2: Survival of individual 21:182 variants under genetic selection............. 79
Figure 3. 3: Survival of FLL_E network variants under genetic selection........... 80
Figure 3. 4: The DFC LtrOnu is over 95% insoluble in standard LHE purification buffers. ................................................................. 83
Figure 3. 5: Coevolving network substitutions cannot rescue LtrOnu cleavage activity. 85
Figure 3. 6: Helix stabilization network (HSN) residue saturation cannot increase LtrOnu cleavage activity................................................................. 86
Figure 3. 7: A suppressor screen was unable to increase LtrOnu cleavage activity...... 88
Figure 3. 8: Helix-1 interfacial contacting (HIC) residue saturation cannot increase LtrOnu cleavage activity................................................................. 89

Figure 4. 1: Schematic of I-LtrI, the coevolving network, and central four nucleotides.. 96
Figure 4. 2: Strategy used to screen LHE in vivo cleavage activity......................... 105
Figure 4. 3: Most LHE variants show poor activity on non-cognate substrates......... 106
Figure 4. 4: G-C rich C4 sequences are nonpermissive to I-LtrI activity............. 107
Figure 4. 5: Enrichment values are predictive of survival under functional selection.. 108
Figure 4. 6: Enrichment does not always agree with functional selection data........ 109
Figure 4. 7: Product enrichment and substrate depletion flow-through............... 114
Figure 4. 8: Product enrichment results for I-LtrI............................... 115
Figure 4. 9: Comparison of I-LtrI substrate depletion and product enrichment...... 116
Figure 4. 10: Product enrichment by I-OnuI..................................... 117
Figure 4. 11: Comparison of I-OnuI substrate depletion and product enrichment. ....... 118

Figure 4. 12: The effect of E184D is context-dependent........................................... 119

Figure 4. 13: E184D increases cleavage activity compared to wild-type I-LtrI. ........... 122

Figure 4. 14: Structural models for I-LtrI variants in complex with the cognate ATAC sequence. .................................................................................................................. 126

Figure 4. 15: Least squares alignment of I-LtrI variant N-terminal domains.............. 127

Figure 4. 16: 3DNA outputs for the cognate ATAC substrate in complex with each I-LtrI variant. ............................................................................................................................... 128

Figure 4. 17: Structural models for I-LtrI AEGE and AEGD in complex with the non-cognate AACC sequence. ........................................................................................................ 129

Figure 4. 18: 3DNA analysis of the AACC substrates in complex with I-LtrI AEGE (wild-type) and AEGD.............................................................................................................. 130

Figure 4. 19: I-LtrI residue R311 makes a direct contact to C4 bases....................... 132

Figure 5. 1: The fusion of LHE domains may create unfavorable network combinations. ................................................................................................................................. 148

Figure 5. 2: The I-OnuI contact map. .................................................................................... 155
List of Appendices

Appendix S 1: Copyright Permissions.................................................................161

Appendix S 2: Supplemental Information for Chapter 2........................................163
  S2 Supplemental Figures.................................................................162
  S2 Supplemental Tables.................................................................178
  S2 Supplemental Datasets.................................................................179
  S2 Supplemental References.............................................................179

Appendix S 3 Supplemental information for Chapter 4........................................181
  S3 Supplemental Materials and Methods.............................................180
  S3 Supplemental Datasets.................................................................181
  S3 Supplemental References.............................................................181
# List of Abbreviations and Symbols

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>APC</td>
<td>average product correction</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>Cas</td>
<td>CRISPR-associated</td>
</tr>
<tr>
<td>CRISPR</td>
<td>clustered regularly interspaced short palindromic repeats</td>
</tr>
<tr>
<td>DFC</td>
<td>domain fusion chimera</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSB</td>
<td>double-strand break</td>
</tr>
<tr>
<td>DSC</td>
<td>differential scanning calorimetry</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>GE</td>
<td>genome editing</td>
</tr>
<tr>
<td>HR</td>
<td>homologous recombination</td>
</tr>
<tr>
<td>HE</td>
<td>homing endonuclease</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase</td>
</tr>
<tr>
<td>KD</td>
<td>dissociation constant</td>
</tr>
<tr>
<td>kDa</td>
<td>kilodalton</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>magnesium chloride</td>
</tr>
<tr>
<td>MI</td>
<td>mutual information</td>
</tr>
<tr>
<td>min</td>
<td>minute</td>
</tr>
<tr>
<td>MIP</td>
<td>entropy and phylogeny corrected mutual information</td>
</tr>
<tr>
<td>MSA</td>
<td>multiple sequence alignment</td>
</tr>
<tr>
<td>NHEJ</td>
<td>non-homologous end joining</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>PAM</td>
<td>protospacer adjacent motif</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank identity</td>
</tr>
<tr>
<td>RE</td>
<td>restriction endonuclease</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RVD</td>
<td>repeat-variable diresidue</td>
</tr>
<tr>
<td>s</td>
<td>second</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SOC</td>
<td>super optimal broth with catabolite repression</td>
</tr>
<tr>
<td>TAE</td>
<td>tris base, acetic acid, and EDTA buffer</td>
</tr>
<tr>
<td>TAL</td>
<td>transcription activator-like</td>
</tr>
<tr>
<td>TALEN</td>
<td>TAL effector nuclease</td>
</tr>
<tr>
<td>TBE</td>
<td>tris base, boric acid, and EDTA buffer</td>
</tr>
<tr>
<td>Tm</td>
<td>melting temperature</td>
</tr>
<tr>
<td>tris-HCl</td>
<td>tris (hydroxymethyl) aminomethane hydrochloride</td>
</tr>
<tr>
<td>ZF</td>
<td>zinc-finger</td>
</tr>
<tr>
<td>ZFN</td>
<td>zinc-finger nuclease</td>
</tr>
<tr>
<td>Zpx</td>
<td>Z-score outliers from a MIp analysis</td>
</tr>
</tbody>
</table>
Chapter 1

1 Introduction

1.1 Genetic Manipulation and Genome-Editing (GE)

1.1.1 Historical Perspective

Human civilization as we know it has been made possible by the deliberate genetic manipulation of plants and animals. Before and during the last ice age, nomadic tribes spent much of their time and energy hunting or gathering food, and population levels were limited by the availability of these essential resources (1). Then approximately 11-12 thousand years ago, ancestral peoples began the process of domesticating plants and animals by selectively breeding desirable traits such as increased grain yield, milk production and muscle mass (2). These breeding strategies have continued to the present day and have been combined with the application of large scale mutagens, such as ionizing radiation, to increase trait and species diversity (3). These artificial selection techniques represent a basic form of genetic manipulation but they truly laid the foundation of contemporary society by allowing population levels to expand, and freeing minds to think of things other than food and survival (2).

Today we reap the benefits of domestication and artificial selection but these processes are slow and often result in unexpected outcomes (4). This is because artificial selection generally requires many generations of mutagenesis and inbreeding, and the imprecise nature of these methods causes future generations to have very similar genetic makeups, or genotypes (5). This inherent lack of genetic diversity means that selectively bred organisms are more susceptible to disease and infertility because of the accumulation of recessive alleles, and highly similar defense mechanisms (5).

Everything changed in the mid 20th century when DNA was identified as trait-bearing hereditary material of life (6, 7), and the raw material for both natural and artificial selection (8). Since that time, techniques have been developed that allow for the specific modification of genetic information such that only predictable regions of the genome are affected (9). Advances now permit the deliberate manipulation of genomes within a
single generation, and the resources available to perform these feats have become more powerful with each iteration (10). For these reasons, we now have the ability to breed plants and animals while maintaining genetic diversity (11), and engineer novel therapies to defeat human pathogens and genetic disorders (12).

Today, the influence of genome-editing (GE) biotechnology is expanding at a rapid rate (13), and facilitating the establishment of new economic sectors such as synthetic biology (14). The potential for GE to benefit society could rival the impact of domestication, but regrettably its progression has been limited by poor public perception and a misunderstanding of how the technologies work (15).

1.1.2 Genome-Editing (GE) Strategies

Traditional methods to edit genomes differ depending on the desired outcome and biological system but all are relatively imprecise and inefficient (16). For example, the genome of mouse models was originally modified by injecting DNA into embryonic cells, followed by random integration into the mouse genome (17). Although this effectively delivers the genetic material into cells, inaccuracy of the method means that exogenous sequences may be incorporated into heterochromatic DNA, other epigenetically silenced regions, or even within a coding sequence of the genome (18).

Fortunately, the imprecision and unpredictability of traditional methods are now being overcome with the application of rare-cutting endonucleases that specifically target a genomic sequence of interest (10, 16). These precision enzymes have such long recognition sequences (to be discussed) that they have probabilities of occurring only once within complex genomes. As a result, these tools now permit the accurate modification of almost any genomic sequence, and improvements in specificity have increased the safety of next-generation genomic therapies and biotechnologies (12).

Another advantage of these modern technologies is that the same general methodology can be applied to most eukaryotic systems (Figure 1.1). First, the precision endonuclease is used to introduce a double-stranded break (DSB) at a specific location within the genome. In turn, the DSB activates native DNA repair machinery, and resolution of the
break is biased towards a desired outcome (19). This overview is simplistic, but these approaches have been used successfully in numerous academic and commercial applications.

The major complication for these techniques is that repair pathway utilization is different between biological systems, and is heavily influenced by the temporal state of the cell (20). For example, the predominant repair mechanism in non-dividing mammalian cells is the non-homologous end-joining (NHEJ) pathway (21). Although classical NHEJ is quite efficient, sequence information may be lost during the repair process and result in a loss of function mutation (20). In dividing mammalian cells or other biological systems, such as S. cerevisiae, homologous recombination (HR) uses a complementary sequence as template to resolve the DSB (22). If a sister chromosome is used as template, then the HR event may result in no change or stimulate gene conversion if the cell is heterozygous for the targeted allele. However, if a non-native sequence with homology is provided and used as template then new genetic information can be inserted, or a disease-causing mutation can be corrected (20).

The manipulation of genomes has profoundly influenced our lives, and is arguably one of the most important factors contributing to the rise of modern society (2). Currently GE may be applied to cure human genetic disease, secure food stables to feed an expanding population, produce environmentally-friendly biofuels, and create novel therapeutic compounds (23). The benefits of these technologies are indisputable but the efficacy and safety of GE is reliant upon the molecular reagents used to introduce the targeted genomic DSB. For this reason, we must continue to study these enzymes to improve their precision and ability to recognize any DNA sequence of interest.
Figure 1. 1: Schematic representing the general strategies employed during precision genome editing applications.

*If heterozygous and most efficient in diploids.
1.2 Precision Genome-Editing (GE) Reagents

1.2.1 Homing Endonucleases

Homing endonucleases (HEs) are a class of mobile genetic element that have the ability to selfishly propagate their genetic information into host genomes through a process termed ‘homing’ (24). These enzymes are generally encoded within self-splicing introns (or inteins) and introduce DSBs at highly specific target sites found in intron-less (or naïve) genomes (25). Following introduction of the DSB, the HE open reading frame is inserted proximally to the cut site, and stably propagated to future generations as long as the insertion event is beneficial or phenotypically neutral (i.e. does not negatively impact host survival) (26).

To date six families (LAGLIDADG, GIY-YIG, His-Cys box, HNH, PD-(D/E)-xK and Vsr-like) of HEs have been identified and each group is classified based on the presence of conserved catalytic or structural motifs (24). HEs are found within the genome of organisms in all kingdoms of life, and generally bind 12-40 bp and asymmetric recognition sequences (26). The catalytic mechanism used to introduce a genomic DSB differs between each HE family, but all recognize target sites along an extensive protein-DNA interface (24). Specificity for the target site is predominantly determined by direct bp contacts but indirect or shape-based readout of sequences also plays an important role in some families (27). Together with these readout mechanisms, extensive non-specific backbone contacts increase HE affinity for DNA substrates and result in dissociation constant ($K_D$) values in the nanomolar to picomolar range (25). Unlike other endonucleases such as restriction enzymes, specific DNA binding is accomplished without saturating all possible hydrogen bonding contacts (28). This feature means that HEs can tolerate nucleotide substitutions at most positions within their target sequence while still maintaining some catalytic activity. Although biologically relevant because it improves HE ability to invade new and diverse host genomes, tolerating substitutions at certain positions of the recognition site may also increase the potential of off-target effects (24).
Members of the LAGLIDADG (to be discussed in detail) and GIY-YIG families have been successfully applied in numerous GE applications by re-engineering native enzyme specificity, or combining nuclease domains with other DNA-binding proteins (27). Their exceptional substrate specificity provides the precision required for editing complex genomes, and their small size (relative to other GE reagents) facilitates their packaging into commonly used delivery vectors (18). However, as a consequence of having a highly specific protein-DNA interface it can be difficult to re-engineer their sequence recognition (27). This has restricted their capacity to target any sequence of interest and limits HE applicability as GE reagents.

1.2.2 CRISPR/Cas Systems

Clustered regularly interspaced short palindromic repeat (CRISPR) and CRISPR-associated (Cas) genes are prokaryotic sequences involved with a basic form of adaptive immunity (29). The CRISPR-Cas system is composed of three stages which function together to protect host cells from attack by foreign DNA: (i) recognizing and adapting to foreign sequences; (ii) transcription and processing of CRISPR RNAs; and (iii) an interference response (29).

Separating repeats at the CRISPR locus are short regions of spacer DNA that were incorporated by the system after previous exposure to exogenous sequences (30). These foreign sequences, or protospacers, were incorporated into the CRISPR locus because they encode genetic material that is unique to potential pathogens, and contain a conserved set of nucleotides called the protospacer adjacent motif (PAM) (29). The PAM sequences are critical to the adaptive acquisition of protospacers because they are recognized by Cas proteins through specific protein-DNA contacts during the interference response (31).

Following transcription and processing, CRISPR RNAs (crRNAs) are able to bind foreign DNA using standard Watson-Crick base pairing (29). This DNA-RNA duplex is then recognized by a multi-protein complex that cleaves and eliminates the DNA sequence if there are no mismatches between the spacer/protospacer complex and the PAM sequence is correct (29). Together, the acquisition of new spacers, processing of
CRISPR RNAs, and the recognition and cleavage of foreign DNA form the CRISPR/Cas system and the adaptive immune system of bacteria and archaea (30).

In over-simplistic terms, the acquisition of new spacer sequences in the native type II CRISPR/Cas system requires two RNA molecules and the endonuclease Cas9 (13). The crRNA binds to foreign DNA while a trans-activating RNA (tracrRNA) associates with the crRNA and forms a hairpin loop that activates the Cas9 endonuclease following binding (29). In 2012, Jennifer Doudna and Emmanuelle Charpentier ingeniously simplified this system by fusing the crRNA and tracrRNA to create a single guide-RNA (gRNA) to localize and activate Cas9 (32). In this setup, the gRNA can be interchanged to recognize any DNA sequence adjacent to the correct Cas9 PAM sequence (29). The application of this system has been rapidly adopted within the GE field because of its simplicity (16), and even been repurposed to control gene expression (33). Recently, researchers have also modified other CRISPR/Cas systems using the Doudna/Charpentier method and relaxed or altered the PAM specificity of Cas9 to expand the targetability of the enzyme (34).

Although CRISPR/Cas9 has revolutionized GE and brought a lot of attention to the field, the system is not without limitations. Genetic constructs containing the gRNA and Cas9 are often quite large (greater then 6 kilobases) which limits their packaging into commonly used vectors for in vivo delivery (18). Studies have also highlighted problems with specificity as gRNAs often tolerate mismatches which lead to off-target effects and unexpected mutations throughout the genome (35). Finally, the cleavage events by Cas9 result in blunt-ended DSBs (29) that are excellent substrates for error-free NHEJ (19). These problems not only decrease the efficacy of genome editing in vivo but also mean that extensive screening is required for ex vivo treatment or cell line gene knock-out(20). Researchers have been addressing these issues by decreasing the affinity of Cas9 for DNA (36), identifying smaller Cas proteins (37), and fusing Cas9 to other endonucleases such as members of the GIY-YIG family (38). These studies have been effective at mitigating some problems but also increase the engineering burden and limit the number of sequences that can be targeted within the genome.
1.2.3 Synthetic Genome-Editing Reagents

Before the discovery and modification of CRISPR/Cas systems, zinc-finger nucleases (ZFNs) (39) and transcription activation-like (TAL) effector nucleases (TALENs) (40) were the most commonly used GE reagents. Both constructs have similar modular architectures in which DNA binding domains are fused with distinct nuclease domains to create synthetic enzymes that target a desired genomic sequence.

Zinc-fingers (ZFs) are small protein domains that bind to three nucleotide sequences with high affinity (41). To target a region of interest, multiple domains can be fused together such that six zinc fingers specifically target an 18-bp region of DNA (9). TAL domains have a conserved repetitive alpha-helical sequence that only differ at two amino acid positions which contact (42) DNA. These residues, or repeat-variable diresidues (RVDs), can be modified to various combinations that bind distinct base pair sequences, and allow TAL domains to be concatenated to target any genomic sequence of interest (40).

The first generation of ZFNs and TALENs were both fused to the non-specific nuclease FokI to introduce a genomic DSB (40, 43). This was problematic because the non-specific cleavage activity of FokI was prone to off-target effects (36), and because it functions as a dimer, the engineering burden was increased two-fold (one construct to either side of the target sequence) (43). To overcome these challenges, monomeric HEs from the LAGLIDAG and GIY-YIG families have been successfully fused to ZF and TAL domains to create second generation ZFNs and TALENs (44). These monomeric enzymes improved upon existing technologies, but altering specificity of the HE domain remains burdensome and can limit their targetability. Additionally, the large size of TALENs and the nature of their repetitive sequences renders them difficult to construct and deliver to cells using common viral vectors (45).
1.3 LAGLIDADG Homing Endonucleases (LHEs)

1.3.1 General Characteristics

LAGLIDADG homing endonucleases (LHEs) are the most diverse and well-studied homing endonuclease family (24). They generally target 22-bp regions of intervening sequences such as group I introns, archaeal introns and inteins. Interestingly, the HO endonuclease from S. cerevisiae is an LHE descendant and was the first homing endonuclease to be discovered (46). Unlike most LHEs, the HO gene is free-standing (i.e. not encoded within an intron/intein) and has evolved to perform a function within the host organism by facilitating gene conversion and mating type switching through the introduction a highly specific DSB at the MAT locus (47).

In the broadest of terms, LHEs are classified into two groups based on the presence of one or two copies of the characteristic LAGLIDADG amino acid motif (24). LHEs containing one copy function as homodimers that recognize palindromic or pseudo-palindromic target sites (48), while enzymes that contain two copies are found as two-domain monomers that target more diverse sequences (49). Although the LHE family demonstrates significant sequence diversity, crystallographic structures demonstrate a high degree of structural homology and symmetry with all domains folding with an $\alpha\beta\beta\alpha\beta\alpha$ secondary structure arrangement (Figure 1.2) (50).
Figure 1. 2: LHEs share a common secondary structure arrangement. Structural models of four monomeric LHE family members bound to their cognate recognition sequences.

I-AnI (PDB: 1P8K) from *A. nidulans*, I-HjeMI (PDB: 3UVF) from *H. jecarina*, I-LtrI (PDB: 3R7P) from *L. truncatum* and, I-Onul (PDB: 3QQY) from *O. novo-ulmi*. Notice the symmetry between domains, αβαβαβ secondary structure folds and DNA deformation (bending) along the length of the recognition sequence.
1.3.2 The LAGLIDADG Interface

The two domains of homodimeric and monomeric LHEs interact at an extensive “LAGLIDADG interface,” and contact predominantly through two parallel, right-handed alpha helices (Figure 1.3) (51). Each helix contains one copy of the characteristic motif and interactions at this interface are critical to the structural integrity and catalytic activity of LHEs. This is because packing of the LAGLIDADG helices not only generates the protein hydrophobic core, but also forms two overlapping active sites and positions essential metal-binding residues in close proximity to the scissile phosphates (24,50).

As the LAGLIDADG motif is a consensus sequence, members of the family demonstrate variation in their specific amino acid composition (52). However, these characteristic residues still display the highest level of conservation in an otherwise highly variable LHE amino acid sequence (24). For example, motifs from the closely related LHEs I-Onul and I-LtrI are ~80% conserved while the rest of the protein sequence displays only ~40% conservation (53).

Given the importance of this protein-protein interface to LHE biology, it has been the focus of numerous studies that have identified many critical residue interactions between each domain (52, 54, 55). Silva and Belfort (2004) grafted interfacial residues from the homodimer I-CreI onto the monomeric LHE I-DmoI and assessed the effect on enzyme structure and function. These substitutions created mutants with a variety of phenotypes including enzymes that prefer to nick their DNA targets as opposed to introducing a DSB. The authors attributed mutant phenotypes to a disruption of the LAGLIDADG helix interaction, and identified its similarity to the “GxxxG motif” found in helices of transmembrane proteins (52). These interactions occur predominantly through backbone interactions, as opposed to side-chain interactions, and illustrate that a high level of helix packing occurs at this protein-protein interface (56).
Figure 1. 3: The LAGLIDADG helices.

Structural model of I-LtrI bound to its cognate recognition site (PDB: 3R7P) with LAGLIDADG helices identified in red (upper). Amino acid side chains from the LAGLIDADG motif are identified using a stick representation (panel). In this structure, two acidic residues (Glu) at the base of the helices coordinate divalent metal ions (yellow spheres) in proximity to the DSB. Notice the helix packing with backbone atoms being less than 5Å apart at the centre of the LAGLIDADG helices.
More recently, Baxter et al. (2012) grafted the LAGLIDADG interfacial residues from I-OnuI onto the structure of five other closely related LHEs (I-Ltrl, I-GpiI, I-GzeI, I-PanMI, and I-SscMI). Although the enzymes display complete amino acid conservation at 60% of the positions substituted, all common interface constructs displayed defects in binding and/or cleavage of their native targets, while I-Gpi, I-GzeI and I-SscMI also displayed significant decreases in expression (55).

Together, these studies demonstrate the critical and complex nature of contacts between LHE domains at the LAGLIDADG interface. Specific to their re-engineering for GE is the fact that disrupting these interactions can impact expression, binding, and catalytic activity even though none of the residue sidechains directly contact DNA.

1.3.3 Catalytic Mechanism and the Role of Divalent Metal Cofactors

Following recognition and binding of their 22-bp target site, active LHEs sequentially nick the top and bottom strand of the minor groove (4-bps apart) to produce a DSB with cohesive 3’ overhangs (57). Similar to many other endonucleases, LHEs have been shown to obey single-turnover (pseudo-Michaelis-Menten) kinetics (58) because catalytic rate (kcat*) is limited by dissociation of the enzyme-product complex, and so reactions do not reach steady-state (59). The catalytic mechanism of homodimer I-CreI has been particularly well characterized and is thought to be generalizable to other LHEs including the single-chain monomers (such as I-Ltrl and I-OnuI) (24).

Introduction of the DSB is accomplished using an ATP-independent, two-metal endonucleolytic mechanism with active site residues performing four main functions: (i) to stabilize the transition and intermediate states; (ii) provide basic conditions that promote nucleophilic water formation; (iii) coordinate essential divalent metal (usually Mg$^{2+}$) cofactors, and (iv) provide a general acid to protonate the leaving group (59, 60).

Numerous studies have demonstrated a critical role for Mg$^{2+}$ during catalysis, but there is currently no consensus regarding the number and specific placement of metals within the active site (59, 61). I-CreI employs the canonical two-metal mechanism for phosphodiester hydrolysis, and the overlapping active sites have been found to contain a
total of 3 bound metal ions (61). A shared metal is coordinated by an essential carboxylate residue (E29 and E184 in I-LtrI) from each domain and the scissile phosphates (61). This common metal has been shown to stabilize both the transition state and 3’-hydroxylate leaving group during the sequential hydrolysis reactions (59, 61). Separately, the unshared metals associate with one of the two active site pockets, and charge a water molecule to facilitate nucleophilic attack on the scissile phosphates (59). Intriguingly, the active site of the monomer I-SceI also contains 3 metal ions (62), but structural comparisons show that cofactor positioning and active site architecture is significantly different from I-CreI (59).

The lack of clarity regarding metals and active site architecture has been attributed to LHE family sequence diversity, and the numerous ways in which resides could be arranged to execute the aforementioned essential requirements for catalysis (24, 59). Nonetheless, a two-Mg$^{2+}$ mechanism is common to endonucleases because it contributes to both DNA cleavage and substrate recognition (63). Divalent metals contribute to catalysis by promoting nucleophilic water formation, stabilizing charged intermediates, increasing electrophilicity of the scissile phosphorus, and facilitate protonation the leaving group (63). On the other hand, contributions of metals to substrate specificity are less intuitive, and better explained by considering the two-Mg$^{2+}$ mechanisms of polymerases and type II restriction endonucleases (REs).

It was originally thought that substrate discrimination by polymerases was controlled solely through enzyme-substrate induced-fit (64). In this mechanism, only correct template-dNTP (or NTP) base-pairing induces movement of finger-domains into a closed state that promotes catalysis (64). However, structural and kinetic studies have now shown that conformational movement into the closed state is much faster than the phosphoryl transfer reaction, and so the rate-limiting step must be subtler and occur though a second conformational change within the closed complex (63, 65). Structural evidence has shown that while two metals are bound within mismatched closed complexes, a third carboxylate is incorrectly positioned to aid in their coordination (66). Therefore, the rate-limiting step that confers high fidelity to polymerization could be positioning of the third carboxylate to coordinate metals within the closed complex (67).
In this way, the relatively small difference in free energy associated with a non-Watson-Crick base pair can be amplified to dramatically reduce catalytic efficiency (63).

Type II REs (e.g. EcoRI and EcoRV) mostly target 6-bp palindromic target sites, and single base substitutions within this sequence have been shown to reduce catalytic efficiency by up to 6 orders of magnitude (58, 68). This exquisite level of substrate discrimination is partly controlled by saturating all potential hydrogen bonds at the protein-DNA interface, but this cannot be the sole determinant because REs maintain affinity for many non-cognate sequences (69). Interestingly, metal cofactor binding was determined to be unfavorable when complexes are formed with non-cognate sequences (70). Therefore, precision binding by Type II REs is thought to take advantage of metal sensitivity to deviations in the coordination environment, and the sequence-dependent local structure of backbone atoms within the active site (70). With this mechanism, the binding of non-cognate sequences results in a misaligned active site architecture that cannot provide a favorable environment for metal coordination (63).

1.3.4 DNA Target Recognition and Binding

1.3.4.1 Direct and Water-Mediated Substrate Readout

The 22-bp LHE recognition sequence is generally described as having two halves with the left side (nucleotide positions -11 to -1) being readout by the N-terminal domain, and the right side (nucleotide positions +1 to +11) being readout by the C-terminal domain (71). Crystallographic studies show that LHE specificity is largely controlled by direct or water-mediated amino acid contacts to base pairs -11 to -3 and +3 to +11 (Figure 1. 4) (53, 62, 72). The majority of these interactions are accomplished through the formation of complex hydrogen bonding networks between base pairs and polar or charged side chains projecting from β-sheets into the major groove (73).
Figure 1. 4: I-LtrI target site contact map. Schematic of protein-DNA contacts derived from the x-ray crystallography structure of I-LtrI (PDB: 3R7P).

DNA backbone phosphates are indicated by orange spheres while the scissile phosphates are blue. Central four base pairs of the target site are identified with yellow lettering. Note: Amino acid numbering has been altered from 3R7P to align with I-OnuI (PDB:3QQY) numbering. Modified with permission (Appendix S1.1) from Takeuchi et al. (2011).
Closely related LHEs (and isoschizomers) have been shown to use different protein-DNA interactions to recognize the same bases and nucleotide positions, and no residues have been found absolutely critical for DNA binding in all members of the family (73). Additionally, LHEs make contact with only 65%-75% of hydrogen bond donors and acceptors in the major groove, with little or no contacts with the minor groove (73). These features complicate re-engineering of the LHE protein-DNA interface towards non-native target sites because there is no universal code controlling specificity (in contrast to ZFNs, TALENs and CRISPR-Cas9) (9). However, this is a consequence of a biological trade-off between having enough specificity to limit toxicity to the host while maintaining enough flexibility to invade naïve host genomes with sequence diversity (74).

1.3.4.2 Indirect or Shape-based Substrate Readout

An interesting feature of LHEs is the constrained sequence specificity witnessed at the central four (C4) region of the target site between nucleotide positions ±2 (75). While no direct contacts to the C4 region have been identified in the literature, methylation or substitution of C4 nucleotides severely impacts LHE binding and catalysis (75). Unlike the direct readout mechanism described for the left and right halves of the recognition site, specificity in this region is thought to be controlled through an indirect or shape-based readout of the DNA sequence (76, 77).

Indirect readout has been well characterized in restriction enzymes (REs) and plays an important role in determining the sequence specificity of many other DNA-binding proteins (78, 79). Through this mechanism, proteins take advantage of innate DNA structural properties such as flexibility, elasticity, bending and kinking, major and minor groove widths, and hydration to distinguish between sequences (78, 80, 81). Both nucleotide sequence and composition influence these parameters, and make DNA molecules more or less resistant to deformation away from their native B-DNA structure following protein binding.

This phenomenon is exemplified by contrasting DNA sequences with pyrimidine-purine (Y-R) dinucleotide steps to sequences containing high guanine-cytosine (G-C) content.
The base stacking interactions of Y-R dinucleotides are relatively weak, and their presence within a sequence leads to easy structural deformation (negative ΔG values) and significant protein-induced DNA kinking (82). In contrast, the extensive hydrogen bond networks associated with G-C bps act to stabilize DNA structures, and consequentially sequences with high G-C content are resistant to deformation (positive ΔG values) and protein-induced DNA bending (83). Therefore, when a protein binds to DNA there is a sequence-dependent structural deformation of the double-helix, and alteration of the native properties (83, 84). As a result, DNAs adopt unique shapes along a protein-DNA interface, and DNA-binding proteins have evolved the ability to recognize these differences to “indirectly” readout a specific sequence (78).

In LHEs, target sequences are symmetrically bent across the protein surface by approximately 45° following protein binding (see Figure 1. 2) (24). This DNA bending leads to localized over-winding of the helix, and disrupts normal base pair interactions including propeller twist and base stacking in the C4 region (60, 77). This causes a local narrowing of the minor groove that brings scissile phosphates closer to bound metal ions within the active site for catalysis (59). Given that the deformation of DNA is sequence-dependent (85), LHEs have evolved to position catalytic metals optimally for catalysis of their native (or cognate) target site following DNA binding (76). Therefore, mutation or modification of native C4 nucleotides can result in an altered DNA structure in the active site, suboptimal or abolished positioning of metals, and a loss of catalytic activity (76).

Molina et al. (2012) investigated the indirect readout mechanism of LHEs by crystalizing I-CreI in complex with cognate and non-cognate C4 sequences in both the presence and absence of metal ions. Their study showed that when the cognate sequence is bound by I-CreI, metals are correctly positioned for catalysis, and the scissile phosphates are moved ~5Å closer together (5.6Å vs 10.5Å when unbound). Interestingly, when the C4 region (wild-type: GTAC) was mutated to non-permissive sequences (AGCG and TGCA) under the same crystallographic conditions, the scissile phosphates were found to be 13.3Å and 10.3Å apart, and no metals were visualized (76). These findings were attributed to the energetics of base stacking because mutation of A-T base pairs to G-C increases stacking energy and decreases DNA flexibility (76). Consequentially, the less flexible C4
sequence adopts a non-native structure in the LHE active site, and this prevents binding of essential metals that contribute to correct positioning of the scissile phosphates for catalysis (76, 77). This conclusion is very similar to substrate discrimination by Res (86), and suggests that indirect readout is an important characteristic of two-metal mechanisms of nucleic acid catalysis (63).

1.3.5 Applying LHEs as Genome-Editing (GE) Reagents

1.3.5.1 Re-Engineering of the LHE Protein-DNA Interface

Currently, re-engineering strategies are largely dependent upon well-characterized LHEs (such as I-CreI and I-OnuI) for which structural studies have identified DNA-contacting residues (87). However, the most important consideration when choosing an LHE scaffold is limiting the number of substitutions in the desired sequence (relative to the cognate), and the location of required substitutions within the recognition site (77). Ideally, an LHE scaffold will be chosen that requires the fewest number of changes, and avoids having to target non-native C4 sequences.

Once an LHE is chosen, a contact module, or cluster of amino acids surrounding specificity-conferring positions, is mutated with extensive rounds of genetic selection being applied to directly evolve the native scaffold towards the desired target (88). For example, in order to alter specificity in the -8 to -5 regions of the I-LtrI target site, amino acid positions Met33, Thr35, Ser37, Arg49, Arg51, Ile75, Thr77, Arg85 and Glu87 require mutagenesis, functional selection and optimization (Figure 1. 5) (89). When multiple positions and/or both domains need to be changed, experiments should be attempted in parallel to reduce the number of unfavorable interactions that may develop (90). This strategy has been used to re-engineer LHE specificity towards the human XPC (91) and MAO-B (53) loci (among others), but success has been limited compared to the retargeting strategies employed for other genome-editing platforms (namely CRISPR/Cas9) (92).
The amino acid contact module conferring specificity to positions -8 to -6 of the recognition sequence mapped onto the structure of I-LtrI (PDBID:3R7P). I-LtrI positions M33(red), Thr35(green), Ser37(blue), Arg49(cyan), Arg51(cyan), Ile75(magenta), Thr77(green), Arg85(cyan) and Glu87(orange) are highlighted as well as base pairs -8 to -6 (yellow). (A) view of the entire I-LtrI scaffold with the N-terminal LHE domain and 5’ end of the recognition sequence coding strand on the left-hand side. (B) view of the amino contact module looking down the DNA from the 5’ end of the recognition coding strand. (C) view of the amino acid contact module from underneath the protein-DNA complex (97).
Some enzymes (E.g. I-CreI (93, 94)) appear to be very amenable to re-engineering but in general, significant changes to an LHE’s natural target specificity are limited by extensive rounds of selection and optimization following mutagenesis (88). For example, altering specificity of I-LtrI between the -8 and -6 positions could require the saturation of nine amino acid positions and screening of up to 20-billion variants (95). Techniques have been developed to screen libraries of up to 10-billion variants but this process is laborious and restricts the applicability of LHEs as genome editing reagents (96).

1.3.5.2 Domain-fusion Chimeras (DFCs)

To limit the amount of re-engineering required to target a desired sequence, distinct LHE domains have been fused to create chimeric enzymes that recognize and cleave hybrid target sites (Figure 1. 6) (55, 98-101). These LHE domain-fusion chimeras (DFCs) have the potential to dramatically increase the number targetable sequences by exploiting the reservoir of distinct LAGLIDADG domains that exist in nature (53, 55). This approach also has the potential to reduce the engineering burden of LHEs because it would avoid much of the extensive mutagenesis and selection required to manipulate the protein-DNA interface (55). Instead, domains that bind a sequence most similar to the desired target could be fused to create a chimeric enzyme that could be optimized for activity using genetic selection (55). This method would reduce re-engineering of the protein-DNA interface, and take advantage of both the structural homology and extensive natural target specificity witnessed in the LHE family (27).

Silva et al. (2006) was one of the first groups to show that LHE domains could be fused to create novel enzymes. This was accomplished by fusing individual domains from the single monomer I-DmoI, and successfully created a stable enzyme with novel target specificity. However, the strategy was only successful with one domain of I-DmoI, and the resulting fusion required extensive re-engineering of the LAGLIDADG interface to increase stability and catalytic activity (98).

Li et al. (2009) expanded upon this work by creating single-chain LHE monomers from the homodimers I-CreI and I-MsoI. The monomeric constructs functioned as well as the parental enzymes, and the technique simplified in vivo applications by preventing the
expression and coordination of two independent subunits. However, while the group provided value insight into the peptide linker required to tether the independent subunits, no attempt was made to fuse distinct homodimeric domains (93).

Recently, Baxter et al. (2012) successfully created LHE DFCs using six members of the monomeric I-OnuI subfamily. The group successfully fused domains to create 30 novel enzymes, and provided a clear framework for engineering DFCs using LHEs with diverse amino acid sequences and recognition sites. Unfortunately, many of their constructs were structurally unstable, defective at binding, and/or had catalytic activity well below the wild-type parental enzymes. Re-engineering of the LAGLIDADG interface towards a common set of residues mitigated some of these issues, but this strategy also created new problems and eliminated activity for some wild-type domain fusions against their cognate sequences (55). These studies provided valuable information for future DFC engineering, but illustrated that there are poorly understood structural and functional constraints between LHE domains which require further study.

1.3.6 Advantages and Disadvantages of LHEs for Genome-Editing

Unlike the leading genome editing tool CRISPR-Cas9, LHEs contain DNA-binding and catalytic activity in a single functional unit (24, 29). This property reduces their open reading frame to under 1-kb, and facilitates their packaging into commonly used delivery vectors. The sequential nicking mechanism employed by LHEs to introduce a DSB also results in a 4-bp (57), single-stranded overhang which has been shown to bias repair pathways towards HR (101). Increasing the frequency of HR is a major goal of the GE field and suggests that LHEs could be better suited for these applications (12).

The major disadvantage of LHE technology is the laborious nature of re-engineering DNA specificity towards a sequence of interest (55, 95). In addition, a poor understanding of how LHEs control their indirect readout mechanism (76, 77), and an inability to relax specificity in the C4 region limits the number of sequences that can be targeted. These problems could be mitigated with the construction of DFCs, but many of these constructs are dysfunctional because critical, yet unknown inter-domain interactions exist and need to be considered when engineering DFCs (55).
Figure 1. 6: General domain-fusion chimera (DFC) engineering strategy.

The N-terminal domain of I-LtrI (PDB: 3R7P) is fused to the C-terminal domain of I-OnuI (PDB: 3QQY) to create a novel enzyme LtrOnu. This DFC is expected to be active against a hybrid target sequence with halves from the parental domains’ recognition sequences.
1.4 Amino Acid Coevolution and Intramolecular Covariation Analyses

1.4.1 Amino Acid Coevolution

Multiple sequence alignments (MSAs) are often constructed to identify important amino acid positions that fulfill important structural or functional roles within protein families. When a position within the MSA shows a high level of conservation it is assumed to be critical for protein structure or function while positions that display sequence diversity are thought to be less important (102). While mutagenesis studies confirm the importance of conserved positions, these generalizations do not explain why the mutation of non-conserved positions also affects protein functionality (102).

Given the complex nature of intramolecular protein interactions, it is not surprising that non-conserved positions can also be important for protein structure and function (103). However, in order to account for the amino acid diversity at non-conserved positions in functional proteins, compensatory mutations must occur to offset any unfavourable phenotype (103, 104). Otherwise, these positions would not exhibit amino acid diversity as mutations would be eliminated by purifying selection (105). These secondary mutations can occur simultaneously or might have happened previously without significantly disrupting protein function (104). Over time, these compensatory relationships may constrain amino acid diversity at each position to certain permissible combinations (104, 105). Analogous to predator-prey relationships at a species level, the evolution of certain amino acid positions may become non-independent, and eventually these residues may coevolve to maintain their important structural or function roles within protein families (105).

1.4.2 Intramolecular Covariation Analyses

The easiest way for amino acids to influence one another is if they contact in the folded protein structure (103, 104, 106). For this reason, it was postulated that identification of coevolving positions could facilitate the development of ab initio methods to predict protein folding from primary sequence (107). To identify coevolving positions, statistical methods have been developed to quantify the non-independent assortment of amino acids
within MSAs (102). In particular, statistics that utilize mutual information (MI) have successfully identified covarying residues in numerous protein families (102, 108). It is important to note that *in silico* methods can identify statistical covariation between amino acid positions within a MSA, but biochemical validation is required to demonstrate biological significance and confirm amino acid coevolution (108, 109). Therefore, these terms may seem interchangeable but there is an important distinction between covariation statistics and the biological process of coevolution.

MI theory states that the amount of information you gain about one random variable increases with knowledge obtained from another variable. This feature makes MI analysis ideal for identifying coevolving residues but their widespread application was limited by three factors (102). First, the MI scores for a position within a MSA are influenced by amino acid diversity (entropy), with less diverse positions having lower scores than positions with high entropy (110, 111). Second, false positive MI scores tend to arise when too few sequences are used in the MSA because scores might not exceed background noise (110). Finally, all positions within a MSA have an underlying level of MI based upon common ancestry, or phylogeny (112). This final problem arises because all proteins in the alignment are derived from a common ancestral protein, and can be mitigated by eliminating highly similar sequences from the MSA (110).

The plethora of sequence data available for most protein families has reduced the impact of alignment size on MI scores while the effects of entropy and phylogeny were addressed by Gloor *et al.* (2005). They reasoned that the majority of covarying positions within a protein family result from common ancestry. With this rationale, the effect of phylogeny was assumed to impact all positions equally, and the raw MI scores between two positions were normalized using the product of their joint entropy (105). This method permitted the identification of covarying positions in 23 different protein families, and revealed two classes of coevolving positions (105). The first class consisted of 2 or 3 residue positions that were adjacent in sequence, or had side chains that contacted within the 3-dimensional structure of the protein. As a consequence, the covariation of these positions is thought to be mostly driven by constraints on local protein structure (103, 104, 110). The second class consisted of groups of residues that often surrounded critical
protein-protein or protein-ligand interfaces, and enzyme active sites. These coevolving networks were found to display the highest MI scores because their evolution is constrained by both structural and functional limitations (105).

Dunn et al. (2007) improved upon the application of MI to identify covarying positions by estimating the background phylogenetic signal, or noise, in a MSA using a metric called the average product correction (APC). Rather than normalizing MI scores using the entropy between two positions, the average MI scores for two positions (with all other positions within a protein) are multiplied, and normalized by the average MI score for the entire protein family. Subtracting the APC value from an MI score led to a new covariation statistic, MIp, that was able to increase amino acid contact prediction in all model protein families (102).

A limitation of this method was that positions within a MSA often demonstrate different levels of average MIp covariation (102). This feature complicates the assignment of a universal significance threshold because different cut-offs increase the rate of false positive or false negative MIp scores (102). After demonstrating that MIp scores are normally distributed, the values were converted to Z-scores and reduced the need for subjective MI significance thresholds. This work was continued by Dickson et al. (2010) and their work resulted in a covariation statistic, called Zpx, which identifies Z-score outliers from a phylogeny and entropy corrected MI analysis (113).

The rigorous MI corrections from Gloor et al. (2005) and Dunn et al. (2007), as well as the subsequent development of Zpx, have dramatically increased amino acid contact prediction compared to MI scores alone (102, 105, 108, 110). Given that high Zpx values identify covariation outliers that do not randomly assort within a MSA, it is also very likely that these positions coevolve to fulfil important structural or functional roles (108). However, similar to other methods used to identify coevolving residues, it has been difficult to experimentally validate these covariation predictions (105). One reason for these difficulties is that phenotypes observed by mutating coevolving residues often eliminate protein function or alterations fall below the limit of biochemical detection.
1.5 Summary

Advances in GE technology are rapidly improving the efficiency of genome modification and facilitating the development of new technologies and human genetic therapies (10, 23). While GE reagents such as CRISPR-Cas9 have revolutionized the field of genome editing (16, 92), other tools such as LHEs have properties that make them ideal for niche applications (27, 55). However, LHEs are limited by the laborious process required to re-engineer DNA specificity towards genomic sequences of interest (114). One way to mitigate this problem is to engineer DFCs and exploit the natural reservoir of DNA binding specificity in the LHE family (55).

Although some DFCs have been successfully engineered, the functionality of these enzymes is often suboptimal and unpredictable (55). In many cases these problems likely arise through the disruption of important interactions between LHE domains at the critical LAGLIDADG interface (52, 55). This domain interface has been well characterized structurally, but the inherent sequence variability of the LHE family has made it difficult to computationally identify other significant intramolecular interactions (52, 54, 55). Traditional analyses based upon conservation are insufficient because LHE sequence variability dilutes statistical signals, and makes important interactions indistinguishable from background noise (54). To overcome these issues, methods are required that take advantage of sequence diversity to identify important amino acid interactions between LHE domains.

The overarching hypothesis I had when starting the work in this thesis was that covariation analyses can be used to identify coevolving residues between LHE domains, and that this information can be used to facilitate their re-engineering for GE applications. Prior to starting this research, my colleague Dr. Russell J. Dickson constructed a MSA of the LHE family, and used this alignment as input for a covariation analysis using the Zpx statistical approach (108, 113).

In chapter one, I present the results of a thorough biochemical validation of the LHE covariation analysis. In this work, I definitively show that 4 residues computationally predicted to covary form a coevolving network at the domain interface of LHEs. Given
that the coevolving residues are found within the LHE active site, I was able to demonstrate that the network can modulate catalytic efficiency of LHE enzymes. These alterations of phenotype are accomplished without disrupting thermostability, and detrimental residue combinations can be suppressed by mutations at other positions in the coevolving network. This study represents one of the first experimentally-validated examples of amino acid coevolution in enzyme active sites, and provides a framework for applying covariation analyses to identify intramolecular protein coevolution in other protein families.

In my second data chapter, I expanded my study to additional amino acid positions that were identified as covarying between LHE domains. Initial work was accomplished in the context of wild-type I-LtrI and I-OnuI, and demonstrated that semi-conservative residue combinations at the coevolving positions were permissive to activity. Although these data were suggestive of coevolution, the position of these coevolving residues within the protein core did not facilitate a thorough biochemical validation. Instead, I attempted to apply knowledge gained to increase the cleavage activity and stability of a dysfunctional DFC, LtrOnu.

In chapter three, I investigated whether the active site coevolving network (Chapter 2) influences central four (C4) DNA specificity, and then attempted to expand the diversity of C4 sequences cleaved by the LHEs I-LtrI and I-OnuI.

In order to investigate this question, I developed two unbiased methods to access DNA target site specificity that involve: (i) generation of randomized nucleotide substrates in plasmid libraries; (ii) time-course assays involving the randomized input and purified enzyme variants; (iii) the separation of nicked or fully cleaved reaction products from unreacted supercoiled substrates using electrophoresis; (iv) measurement of target sequence abundance in linear DNA, substrate, and input pools using high-throughput sequencing; and (v) quantitation of changes in abundance from the input library, and identification of permissible substrates using compositional data analysis. These methods permitted the exploration of LHE C4 cleavage specificity, and are applicable to
investigating other DNA-binding proteins, such as CRISPR-Cas9 and transcription factors.

In an attempt to expand C4 target specificity, I screened libraries of 1600 coevolving network variants for activity against different C4 sequences in two different LHE backgrounds. Interestingly, I identified a single variant (E184D) that is active against a broader range of C4 sequences than wild-type enzymes. Although this variant has been described in the literature as an “up-activity” variant, I was able to demonstrate that increases in catalytic efficiency are not extendable to all C4 sequences, and are context-dependent within the active site coevolving network. However, E184D expands C4 substrate targetability by I-LtrI and I-OnuI, and increases their applicability as GE reagents.

Finally, I decided to investigate whether residues at the domain interface influence DNA structure within the LHE active site using x-ray crystallography. In collaboration with Dr. Murray Junop and Chris Brown (PhD Candidate), we demonstrate that cognate and non-cognate DNA substrates adopt different conformations when bound by the various coevolving network variants. Some of these observations helped to explain various enzyme phenotypes, and suggested a role for DNA base pair opening in facilitating the catalysis of C4 target sites. Surprisingly, a final analysis of the protein-DNA interface of our wild-type I-LtrI pre-cleavage structure identified a direct residue contact to the C4 region. While, no direct contacts to these base pairs have been described in the literature, I was able to identify 6 other LHEs that appear to make similar contacts. These findings open the door to predictably altering LHE specificity in the central target site region, and could increase the applicability of these enzymes as genome-editing reagents.
1.6 References


15. Caplan AL, Parent B, Shen M, Plunkett C (2015) No time to waste--the ethical challenges created by CRISPR: CRISPR/Cas, being an efficient, simple, and
cheap technology to edit the genome of any organism, raises many ethical and regulatory issues beyond the use to manipulate human germ line cells. *EMBO Rep* 16(11):1421–1426.


30. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and


86. Vipond IB, Halford SE (1995) Specific DNA recognition by EcoRV restriction


Chapter 2

2 Control of Catalytic Efficiency by a Coevolving Network of Catalytic and Noncatalytic Residues

The work in this chapter is reproduced with no permission requirements (Appendix S1.2) from:


2.1 Significance

Maximizing structural and functional information from multiple sequence alignments is difficult for protein families that exhibit extreme sequence variation. We addressed this issue by identifying covarying positions within the sequence alignment to predict networks of coevolving amino acid residues in LAGLIDADG homing endonucleases, enzymes used for genome-engineering applications. Intriguingly, the predicted coevolving network with the highest score includes the active-site metal-binding residues and adjacent residues. We were able to modulate catalytic efficiency ∼100-fold by substitution of residues in the network. Our data show that the evolutionary trajectory and fitness landscape of LAGLIDADG active sites is constrained by a barrier of coevolving residues and imply that generating an optimal coevolving network is an important consideration when engineering these endonucleases.

2.2 Introduction

The active sites of enzymes are often the most conserved positions in a multiple sequence alignment, as purifying selection for maintenance of function constrains amino acid variation of residues that directly participate in catalysis. Noncatalytic residues, often in close proximity to catalytic residues, contribute to enzymatic function by maintaining the architecture and chemical environment of the active site. Noncatalytic residues often show sequence variation in multiple sequence alignments, yet non-permissive
substitutions at these positions will have an impact on enzymatic function by disrupting the architecture or chemical environment necessary for catalysis. Thus, catalytic and noncatalytic residues must coevolve with each other and surrounding residues to maintain active-site conformation and chemistry and to buffer against potentially deleterious mutations (1, 2). Coevolving residues within a protein family can be predicted by computational methods that use mutual information theory to identify residue covariation in a multiple sequence alignment (2–5). However, because the magnitude of covariation between positions varies with the magnitude of positional variation (4), the identification of catalytic residues as part of coevolving networks is problematic for the simple reason that catalytic residues show little sequence variation. Although others have identified putative coevolution between residues involved in catalysis in a small number of protein families (6–8), to our knowledge there are no examples demonstrating the functional consequences of coevolution between catalytic and surrounding noncatalytic positions.

We examined coevolution between the catalytic and noncatalytic residues of LAGLIDADG homing endonucleases (LHEs), site-specific DNA endonucleases that are typically encoded within self-splicing introns and inteins (9). LHEs show extreme sequence variation and function as homodimers or as single-chain monomers composed of two LAGLIDADG domains that evolved by gene duplication or gene fusion events. The active site of LHEs consists of two parallel α-helices containing the class-defining LAGLIDADG amino acid motif with acidic metal-binding residues (D or E) at the bottom of each α-helix and positioned in close proximity to the DNA substrate (10–12). Outside of the LAGLIDADG α-helices, the extreme sequence variation makes it difficult to build robust alignments and infer functional information (13–15). Moreover, the monomeric and dimeric LHEs likely evolved under different functional constraints, and the phylogenetic signal and functional information for either form is diluted in alignments that include both the monomeric and dimeric LHEs. Because LHEs are currently under investigation for use as genome-editing agents (16–18), a greater understanding of their functional constraints would aid in engineering studies.

Here, we take advantage of high-quality structure-guided multiple sequence alignments of single-chain LHEs to predict coevolving networks using methods based on mutual
information (5, 19, 20). Strikingly, the network with the strongest predictive scores included the metal-binding catalytic residues and adjacent noncatalytic residues that lie on opposite LAGLIDADG α-helices. The coevolving network was experimentally confirmed by functional analyses showing that catalytic activity could be modulated over an ~100-fold range by mutation of either catalytic or noncatalytic residues in the network. Our results show that maintaining integrity of coevolving networks of catalytic and noncatalytic residues is an important consideration when engineering LHEs and may also be applicable to engineering of other enzyme families.

2.3 Materials and Methods

2.3.1 Oligonucleotides and Plasmids

All oligonucleotides used in this study were synthesized by Integrated DNA Technologies, Inc. Plasmid DNA was isolated from E. coli cultures grown in Luria Broth (LB) with an EZ-10 Spin Column Plasmid DNA Kit (Bio Basic Inc.) according to the manufacturer’s protocol. A cesium chloride (CsCl2) gradient was used to obtain supercoiled plasmid DNA from large-scale cultures that was used as substrate for all kinetic analyses. Wild-type I-LtrI and I-OnuI encoding genes (codon-optimized for E. coli) were cloned between the NcoI and NotI sites of plasmid pEndo with a single methionine and glycine sequence added to the N termini. Point mutations were then incorporated into I-LtrI and I-OnuI using a QuikChange II Site-Directed Mutagenesis Kit (Stratagene) to create a variety of amino acid combinations at the coevolving positions, and the mutations were confirmed by sequencing.

2.3.2 Bacterial Two-Plasmid Functional Selection

A bacterial two-plasmid functional selection was used to screen the activity of all LHE variants used in this study, as previously described (39). For liquid media selections, 10 ng of LHE variant (in pEndo) were transformed into 50 µL of competent NovaXGP’ (Novagen) cells harboring the appropriate pTox plasmid. Transformants were allowed to recover in 300 µL of 2× YT medium (16 g/L tryptone, 10 g/L yeast extract, and 5 g/L NaCl) at 37 °C and 200 × g for 30 min. Experiments with I-HjeMI included a 1-h expression period in 2× YT supplemented with 0.02% arabinose and 100 µg/mL
carbenicillin. Following the recovery and expression period, both I-LtrI and I-HjeMI cultures were diluted 200-fold into either nonselective [1× M9 salt, 0.8% wt/vol tryptone, 1% vol/vol glycerol, 1 mM MgSO4, 1 mM CaCl2, 0.2% wt/vol thiamine, 100 µg/mL carbenicillin, and 0.02% (wt/vol) L-glucose] or selective media [nonselective media lacking glucose with the addition of 0.02% (wt/vol) L-arabinose and 0.1 mM isopropyl β-D-1-thiogalactopyranoside]. Cultures were then grown at 37 °C and monitored until cell turbidity reached ∼0.7 at 600 nm.

For solid media selections, 50 ng of LHE variant (in pEndo) was transformed into 50 µL NovaXGF’ (Novagen) cells harboring the appropriate pTox plasmid. Transformants were allowed to recover in 300 µL of 2× YT medium for 10 min, followed by the addition of 2 mL 2× YT medium supplemented with 100 µg/mL carbenicillin and 0.02% L-arabinose. Cultures were then grown at 37 °C for 1 or 4 h of outgrowth, harvested, and resuspended in sterile saline (0.9% wt/vol NaCl), and dilutions were spread onto nonselective and selective agar plates. Plates were incubated at 37 °C for 16–24 h, and the survival percentage was calculated as the ratio of colonies on selective to nonselective plates. Three biological replicates with two technical replicates per selection were performed. An I-LtrI catalytic mutant (E29A) was used as a negative control for all selections.

### 2.3.3 Mutant Library Synthesis and Screening

The I-LtrI and I-HjeMI quartet libraries were constructed by randomizing the noncatalytic positions of I-LtrI (28 and 183) and I-HjeMI (17 and 149) to all 20 amino acids and holding the catalytic positions (29 and 184 of I-LtrI and 18 and 150 of I-HjeMI) to either Asp or Glu (GenScript). The library was screened for activity using two-plasmid bacterial functional selection in liquid media followed by paired-end sequencing on the Illumina MiSeq platform at the London Regional Genomics Centre (London, ON). Reads were parsed for the presence of D or E at positions 29 and 184, and the number of reads in all possible quartets was identified using a custom Perl script. The significance of the proportional abundance of each quartet in the selected versus non-selected condition was determined using the ANOVA-like differential analysis method (25, 26). We generated 128 Dirichlet Monte-Carlo instances of the selected and non-selected datasets, performed t tests on each instance, and estimated the associated false discovery rate for each
Dirichlet instance using the lfdrtool (25). Values reported and plotted are effect sizes calculated for those variant quartets that had estimated false discovery rates of 0.05 or less (25). Data were plotted using R (40) and the ggplot2 package (41).

2.3.4 BioScreen Bacterial Growth Assay

A Microbiology Reader Bioscreen C (MTX Lab Systems, Inc.) was used to measure the growth of individual I-LtrI variants in liquid culture medium. A total of 10 ng of I-LtrI variant DNA was transformed into 50 µL of NovaXGF’ cells (containing pTox-Ltr), and cells were allowed to recover in 400 µL of 2× YT at 37 °C for 30 min. Following the recovery period, 4 µL of culture was aliquoted into 200 µL of both selective and nonselective media within wells of a 10 × 10 Honeycomb 2 plate (Oy Growth Cures Ab, Ltd.). Plates were incubated within the BioScreen apparatus at 37 °C with medium shaking while culture turbidity was measured every 15 min for 24 h. Each I-LtrI variant was tested using four independent transformations (n = 4) and during two separate growth periods.

2.3.5 Protein Expression and Purification

I-LtrI and select variants were cloned between the NcoI and NotI sites of plasmid pProExHta (Invitrogen and Life Technologies), and the 6× histidine-tagged proteins were expressed in E. coli strain ER2566 (New England Biolabs) at 16 °C for 16 h. Cells were harvested at 6,000 × g for 15 min, and pellets were resuspended (40 mL/1 g of cell pellet) in Binding Buffer [50 mM Tris·HCl, pH 8.0, 500 mM NaCl, 1 mM imidazole, and 10% (wt/vol) glycerol] supplemented with SIGMAFAST protease inhibitor (Sigma). Cells were lysed using an EmulsiFlex-C3 high-pressure homogenizer followed by sonication for 30 s. Cell lysates were cleared by centrifugation at 29,000 × g for 30 min at 4 °C, and the supernatant was loaded onto a 1-mL HiTrap column (GE Healthcare Life Sciences), washed, and eluted in 6× 1-mL aliquots using elution buffer [50 mM Tris·HCl (pH 8.0), 500 mM NaCl, 500 mM imidazole, and 10% (wt/vol) glycerol]. Fractions were pooled and dialyzed for 16 h at 4 °C into 50 mM Tris·HCl (pH 8.0), 250 mM NaCl, 30 mM imidazole, and 10% glycerol. The N-terminal 6× histidine tags were removed by adding Tobacco Etch Virus (TEV) protease (6× histidine tagged) to a molar ratio of 1:25 TEV to
LHE. The protein mixture was dialyzed into binding buffer for 4 h at 4 °C and run over an equilibrated 1-mL HiTrap column, and the flow-through was collected and dialyzed for 16 h into storage buffer [50 mM Tris·HCl, pH 8.0, 25 mM NaCl, 1 mM DTT, and 10% (wt/vol) glycerol] and stored at −80 °C.

2.3.6 Differential Scanning Calorimetry

Differential scanning calorimetry (DSC) was performed at the Bimolecular Interaction and Conformation Facility at Western University. Two independent DSC determinations were performed for each LHE construct (0.5 mg/mL). Samples were scanned from 10 °C to 110 °C at 60 °C/h using a MicroCal VP-Differential Scanning Calorimeter (GE Healthcare Life Sciences). Raw data were processed using the nonlinear least-squares regression analysis in Origin 7.0 (MicroCal) by first subtracting a buffer–buffer reference scan and then fitting to a non–two-state transition model.

2.3.7 In Vitro Cleavage Assays and Single-Turnover Kinetics

One copy of the cognate target site for I-LtrI (5′-AATGCTCCTATACGACGTT TAG-3′) was cloned between the AflIII and BglIII sites of pLITMUS 28i (New England Biolabs) and was used as substrate for in vitro cleavage assays. I-LtrI and variant protein constructs were diluted using storage buffer to working (10×) concentrations. Six protein concentrations were assayed over a 25-fold range, and the reaction mixture consisted of 50 mM Tris·HCl (pH 8.0), 100 mM NaCl, 10 mM MgCl2, 1 mM DTT, and 5 nM substrate. Both the protein and the reaction mixture (containing substrate) were incubated separately for 5 min at 37 °C before reactions were started. Time-course experiments involved sampling 37 °C reactions at six independent points (not including t = 0) after the addition of protein. Reactions were stopped using 200 mM EDTA, 30% glycerol, 0.2% SDS, and bromophenol blue and incubated at 50 °C for 5 min. The percentage product formed was calculated as the intensity of the linear product band divided by the sum of the three reactants (supercoiled substrate, nicked plasmid, and linear product). Percentage of the product was plotted against time using GraphPad Prism, and the initial rate was determined after curve-fitting to a one-phase association function. Initial rates from each of the six time-course experiments were then plotted against enzyme concentration and fit
to a Michaelis–Menten model to determine the parameters $k_{cat}^*$ and $K_M^*$. All time-course experiments were repeated three times using protein samples from at least two independent purification procedures.

### 2.3.8 I-LtrI A28S Suppressor Screen

I-LtrI A28S was subject to error-prone PCR (approximately seven to nine nucleotide substitutions per kb) by amplifying the ORF using a GeneMorph II Random Mutagenesis Kit (Agilent Technologies). The I-LtrI A28S mutant library (in pEndo) was then used as the input for an initial round (round 1) of bacterial two-plasmid functional selection. All colonies that survived on selective media (480 colonies) after round 1 were individually inoculated into wells of 96-well plates containing LB supplemented with 100 µg/mL ampicillin, grown overnight, and pooled, and plasmid DNA was isolated. The I-LtrI-encoding insert was recloned to create a second mutagenic library, and 30 colonies that survived the second genetic selection (round 2) were sequenced to identify mutations.

### 2.3.9 Rosetta Modeling

In silico mutants of the I-LtrI structure 3R7P structure (21) were generated using PyRosetta (42) and the `mutate residues` function for all permutations of the residues found in the alignment at positions 28, 29, 183, and 184. Structural refinement and scoring was conducted using Rosetta (43) to evaluate the predicted local energy minima of wild-type and in silico mutants. The Rosetta `relax` protocol was applied to the in silico mutant structures with added constraints to the input model using standard flags for the Rosetta `relax`protocol (`constrain_relax_to_start_coords`). In addition to the mutants, the Rosetta `relax`protocol was applied to the wild-type structure with 20 repeats using the –nstructs flag. Each structure was scored as part of the `relax` protocol. Energies are represented in Rosetta Energy Units. The data were plotted using R (40).
2.4 Results and Discussion

2.4.1 Identification of a Coevolving Network of Catalytic and Noncatalytic Residues

We noted that the occurrence of two different residues, Asp (D) and Glu (E), at the catalytic metal-binding sites of the single-chain LHEs is an unusual feature for a protein family and might influence enzyme activity. In the I-LtrI, the metal-binding residues correspond to positions E29 and E184 of the two LAGLIDADG α-helices (I-LtrI numbering is adopted throughout this article) (21). The evolutionary distribution of these residues was examined with an alignment of 178 single-chain LHE protein sequences generated by using Cn3D (22) to develop a structure-based guide alignment followed by LoCo (19) to identify and correct systematically misaligned segments. The resulting alignment was hand-curated to remove all sequences lacking acidic residues at the catalytic positions (19) (Figure S2.1). A maximum-likelihood phylogenetic tree was derived from the full alignment by PhyML (23), and unexpectedly, the distribution of the catalytic residues sorted into four distinct groups: A, B, C, and D (Figure 2.1. a and Figure S2. 2). We considered the possibility that the groupings reflected an evolutionary and functional interaction between the metal-binding and other residues in the LAGLIDADG α-helices. Examination of the alignments by sequence covariation analysis showed that LAGLIDADG α-helices 1 and 2 were the most strongly covarying segments of the protein family by $Z_{px}$ scores (20). The highest covariation score was between the residue positions immediately preceding the metal-binding residues (A28 and G183 of I-LtrI) (Figure 2.1. b, Figure 2.1. c, Dataset S2. 1). Surprisingly, the positions corresponding to the metal-binding residues (29 and 184) were also identified in this analysis (Dataset S2. 1).

Based on these observations we examined the initial hypothesis that the group-specific distribution of the metal-binding residues was influenced by a covarying network composed of these four positions (28, 29, 183, and 184 in I-LtrI). The first test of this hypothesis examined the frequency of all observed quartet species in the LHE alignment and the distribution of the quartets on the LHE phylogenetic tree (Figure 2.1. a, Figure 2.1. d, and Table S2. 1). There was a strong association between the phylogenetic
position of the metal-binding residue and the adjacent residues and between the identity of the metal-binding residues and the identity of the residues permitted at the adjacent positions. In particular, group A LHEs, including I-LtrI, consisted largely of proteins with E_D and E_E metal-binding residues, with A_G or G_A combinations being favored at the adjacent 28 and 183 positions (this naming scheme is given throughout this article where the first metal binding residue is understood to be orthologous to I-LtrI position 29 and the second to position 184, and similarly for the noncatalytic site residues). In contrast, group B and C LHEs possessed D_E and D_D metal-binding residues and a much more diverse set of permitted residues at positions 28 and 183 (Figure 2.1. a). However, the achievement of statistical rigor is difficult to achieve from such analyses because of the small number of nonindependent sequences sampled.
Figure 2.1. a: LAGLIDADG homing endonuclease (LHE) family tree.

Cladogram of single-chain LHEs generated from an unrooted maximum-likelihood tree made by PhyML. The outer ring is colored according to the identity of the metal-binding residues at positions 29 and 184, and the branches leading to the tips are colored according to the identity of residues at positions 28 and 183, as indicated. Each class is colored by the residue 28_183 combination at the deepest branch, with no assumptions made regarding an ancestral state. The locations of the I-LtrI, I-OnuI and I-HjeMI endonucleases are denoted by triangles containing the first letter of the endonuclease.
Figure 2.1. b: Covarying residues of the LHE family.

Circos plot (44) illustrating covariation scores between residues of LHEs, mapped onto I-LtrI sequence. The outer ring shows I-LtrI residue number and identity for positions within the LHE alignment. The next ring shows amino acid conservation using a heat map (with red indicating “most conserved” and blue indicating “least conserved”) and bar plot. The internal ring shows covariation scores using bar plots, with height of the bar proportional to the covariation score. Red lines connect positions with the highest covariation scores, black lines connect positions with intermediate scores, and gray lines connect positions with the lowest scores.
Figure 2.1. c: Structure of the LHE coevolving network.

Covarying positions A28 (yellow), G183 (teal), and E29 and E184 (red) mapped onto the structure of I-LtrI (blue, Protein Data Bank 3R7P) in the presence of DNA substrate. Magnesium cofactors (violet) are shown as dotted spheres.
Figure 2.1. d: Naturally occurring coevolving network combinations.

Heat map of residue combinations in the LHE sequence alignment for positions 28, 29, 183, and 184 plotted on a log_2 scale. Identities of the metal-binding residue at positions 29 and 184 are on the x axis, and identities of residues at positions 28 and 183 are on the y axis. Combinations of residues at positions 28 and 183 that were observed once in the LHE alignment (G_T, G_R, and G_D) are not plotted. ND, residue combination not detected in alignment.
2.4.2 Saturating Mutagenesis of the Coevolving Network Recapitulates Phylogenetic Patterns

We performed an unbiased functional screen that selects for functional LHE variants to determine if the observed phylogenetic relationships in the predicted coevolving network residues were constrained (Figure 2.2. a). Positions 28 and 183 were each randomized in the context of the I-LtrI backbone to all 20 amino acids, whereas the metal-binding residues at positions 29 and 184 were each held at either D or E to give a complexity of 1,600 possible I-LtrI variants (Figure 2.2. a). This library was screened for enzymatic activity in a liquid culture assay where expression of an active endonuclease (carried on pEndo-Ltr) would cleave an I-LtrI recognition site on a plasmid encoding a bacteriostatic DNA gyrase toxin gene, thereby eliminating the toxin-carrying plasmid (pTox-Ltr) (24). The rate of destruction of pTox-Ltr by the I-LtrI variants correlates with the ability of the enzyme to bind to and cleave the recognition site. Loss of pTox-Ltr is reflected as an increase in relative abundance for cells carrying active I-LtrI variants over weak or inactive variants (Figure 2.2. b).

We performed seven independent transformations of the pEndo-Ltr library into pTox-Ltr–competent cells, harvested total plasmid DNA from selective and nonselective conditions after 16 h outgrowth, and examined the mutated positions by Illumina paired-end sequencing. Because the strategy captured all four positions in paired-end sequencing reads, it was possible to determine the relative enrichment of all 1,600 quartet combinations in the selected condition versus the unselected condition (25, 26) (Figure S2. 3). As shown in Figure 2.2. b, I-LtrI variants with small side-chain residues at positions 28 and 183 were strikingly enriched relative to other combinations, indicating that these positions were not randomly assorting. In particular, combinations of A and G were favored, possibly indicating that steric restrictions at the base of the LAGLIDADG α-helices surrounding the active site may influence residue identity. Residue combinations at positions 28 and 183 that occurred at low frequency in the LHE alignment, or that were not observed, were also enriched (V_G, T_G, S_G), confirming that the frequency of residues in the current LHE alignment reflects an ascertainment bias. Variants with D_E or D_D combinations at positions 29 and 184 within the context
of the I-LtrI backbone were underrepresented relative to I-LtrI variants with E_D or E_E combinations. These observations strongly parallel the preference of the metal-binding residues on the LHE phylogeny (Figure 2.2. a), supporting the hypothesis that a covarying network of residues influences the amino acid composition of these positions. We conclude that the distribution of metal-binding residues within related LHEs is likely constrained by these context-dependent interactions.

We performed a similar experiment with the LHE I-HjeMI to extend these observations to other LHEs. I-HjeMI is a representative of the type B LHEs that are predicted from the phylogenetic analysis to prefer D_E at the catalytic residues (positions 18 and 150 of I-HjeMI) (Figure 2.1. d). We found that variants with D_D or D_E at the catalytic positions were strongly enriched over variants with E_D or E_E (Figure S2. 4). This result is in stark contrast to that observed with I-LtrI where E_D or E_E catalytic residues were highly enriched (Figure 2.2. b). At the I-HjeMI noncatalytic positions (residues 17 and 149), the wild-type G_A pairing was not the preferred pairing, but rather we observed that variants with an A or G at position 17 could tolerate a wide range of residues at position 149, agreeing with the observed diversity of residues at these positions in the LHE alignment (Figure 2.1. d).
Figure 2.2. a: Schematic of liquid competition growth experiments.

Schematic of the liquid competition growth experiment and sequencing strategy used to screen an I-LtrI mutant library.
Figure 2.2. b: Coevolving network residue preference.

Heat map containing the log₂ effect size (25) for quartet pairs with a false discovery rate of less than 0.05. The metal-binding residue identities for 29 and 184 are found along the x axis, and the identities of residues 28 and 183 are found along the y axis. Relative abundance values for all 1,600 possible quartets are found in Figure S2. 3.
2.4.3 I-LtrI Variants with Suboptimal Networks Grow Slowly or Not at All

To investigate the basis for the low abundance of D_D and D_E variants in the I-LtrI liquid competition experiments, we tested individual I-LtrI network variants in a bacterial two-plasmid functional survival assay on solid media (24) and found that many variants with suboptimal combinations at positions 28_183 and 29_184 survived poorly or not at all (Figure 2.3. a, Left; Dataset S2. 2). Similar findings were found when the same mutations were made in I-OnuI (Figure 2.3. a, Right; Dataset S2. 2), a LHE enzyme with ~30% sequence identity to I-LtrI. Variants of both I-OnuI and I-LtrI with G_A or A_G at positions 28_183 in combination with D_D and D_E at the metal-binding positions survived, but had small colony sizes relative to E_D or E_E variants (Figure S2. 5). This phenotype was supported by liquid growth experiments that measured the time individual I-LtrI variants took to reach mid-log phase (A600 = 0.35) (Figure 2.3. b and Figure S2. 5). Notably, D_D or D_E metal-binding residue variants all had longer times to reach mid-log phase than E_E or E_D variants with the same residues at positions 28_183. Collectively, with the I-HjeMI experiment, these data show that interactions between the metal-binding and adjacent residues are generalizable to other LHE sequence backbones and that the network of coevolving residues influences both survival and growth rate.
Figure 2.3. a: Survival of individual variants under functional selection.

Heatmap depicting the log$_2$ survival for individual I-LtrI (Left) and I-OnuI (Right) mutants in a bacterial two-plasmid functional selection. Identity of the metal-binding residues at positions 29 and 184 are on the $x$ axis, and identity of residues at positions 28 and 183 are on the $y$ axis. ND, not determined.
Figure 2.3. b: I-LtrI variants show differences in growth rate.

Boxplots of time (min) for I-LtrI variants to reach midlog growth (\(A_{600} = 0.35\)) in selective media, with individual data points shown as dots. I-LtrI variants are indicated on the x axis, with colors indicating variants that were analyzed kinetically.
2.4.4 Coevolving Network Modulates Enzymatic Activity

It is possible that the low enrichment and low survival in the liquid competition and plate experiments was due to structural defects of the network variants. We tested this first using Rosetta modeling of sequence variants on the I-LtrI backbone and found that the range of values of predicted Rosetta energy units for different initial random seed values using the wild-type sequences encompassed the values observed for the variants (Figure 2.4. a, Center). Thus, there was no obvious difference in predicted Rosetta energy units for I-LtrI and any variants (Figure 2.4. a). These predictions were supported, using differential scanning calorimetry on select purified I-LtrI and I-OnuI variants to show that the melting temperatures and enthalpy of denaturation were not obviously different for wild-type enzymes or network variants (Figure 2.4. a and Table S2. 2). Intriguingly, the A_A:E_E enzyme displayed a much higher melting temperature than the other variants. It is thought that many enzymes exhibit full activity in a narrow band of optimal stabilities (27); thus, we suggest that the A_A:E_E variant may be locked into a conformation that is incompatible with efficient catalysis.

To test the hypothesis that the covarying network modulates the catalytic efficiency of LHE enzymes, the pseudo-Michaelis–Menten parameters (28, 29) $k_{cat}^*$ and $K_M^*$ were determined from single-turnover reaction conditions for a number of I-LtrI variants and compared with the wild-type enzyme (Figure 2.4. b, Figure 2.4. c and Figure S2. 6). For example, a single change at position 183 (G183A) to create the suboptimal A_A:E_E network resulted in an ~65-fold decrease in $k_{cat}^*$, whereas a single change at position 28 (A28G) to create G_G:E_E generated an enzyme with an ~3.5-fold decrease in $K_M^*$ over the wild-type enzyme (Figure 2.4. b). Interestingly, the G_G combination is infrequently found in the LHE alignment (Figure 2.1. a) and exhibits slower growth in liquid selections compared with A_G variants (Figure 2.2. b and Figure 2.3. b), suggesting a penalty for highly efficient enzymes in a cellular context. Plotting catalytic efficiency ($k_{cat}^*/K_M^*$) versus time to mid-log phase in liquid culture for I-LtrI network variants revealed a striking correlation (Figure 2.4. c and Figure S1), suggesting that changes in enzymatic efficiency are sufficient to explain the differential growth phenotypes (Figure 2.2. b, Figure 2.3. a, and Figure 2.3. b) and the class specificity observed (Figure 2.1. a).
It is possible that many of the I-HjeMI variants that are more active than wild type (Figure S2. 4) may also be less fit in their normal context.

Figure 2.4. a: Coevolving network variants are thermostable.

Plot of experimentally determined melting temperatures (temperature in °C, open circles) and Rosetta energy units (black triangles). *(Left)* Experimentally determined temperature and predicted Rosetta energy units for the indicated proteins. *(Center)* The range of Rosetta energy units for 20 predictions performed using the wild-type I-LtrI structure. *(Right)* The Rosetta energy units relaxed with minor constraints for the indicated I-LtrI variants.
Figure 2.4. b: Pseudo-Michaelis-Menten parameters for individual I-LtrI variants.

Log_{10} plot of $k_{cat}^*$ (nM/min) on the x axis versus $K_M^*$ (nM) on the y axis for the indicated I-LtrI variants. Three replicates for each variant are shown as open circles colored according to quartet combination.
Figure 2.4. c: Catalytically efficient variants grow faster under functional selection.

Log$_{10}$ plot of catalytic efficiency versus time to mid-log phase for the indicated mutants. Fit of the data to a quadratic regression model is shown by a black line with the 95% confidence interval as a gray-shaded area.
2.4.5 Suppressor Mutations in the Coevolving Network Can Rescue Suboptimal Variants

The S_G:E_E variant displays both slow growth and low survival in the plate assay after an extended outgrowth period (Figure 2.5) and was thus a logical candidate for a suppressor screen. We conducted a random PCR-based mutagenesis of the entire S_G:E_E variant gene and selection for functional variants to determine if poorly active variants could be rescued by mutations in the coevolving network. The mutated S_G:E_E library was subjected to two rounds of selection using the two-plasmid survival assay, over which the average survival of the library increased from ~10 to ~95% (Figure 2.5). DNA sequencing identified 18 unique clones, 5 of which contained an E184D mutation (Figure 2.5). This change of the metal-binding residue from E to D creates S_G:E_D from S_G:E_E, restored $k_{\text{cat}}^*$ and $K_M^*$ to wild-type levels (Figure 2.5), and resulted in a more rapid growth (Figure 2.2, b and Figure 2.3, b). Mutations at other residue positions were also isolated in the S_G:E_E suppressors, in particular positions Y2, Q11, and A12 that are in the N-terminal tail of I-LtrI. We suggest that these N-terminal mutations are selected for increased expression of the suppressors, as was noted by a recent study that revealed a relationship between codon use in the N-terminal region and protein expression for *Escherichia coli* proteins (30).
Figure 2.5: A suppressor screen identifies residues in the coevolving network.

(Top) Boxplot of survival of the S_G:E_E I-LtrI variant in the two-plasmid genetic selection with 1- or 4-h outgrowth. Red dots indicate data points for different experimental replicates. (Bottom) Survival of a randomized S_G:E_E library in the two-plasmid genetic selection over two successive rounds of selection with 1-h outgrowths. Genotypes of individual clones sequenced after two rounds of selection. All clones possessed the parental A28S mutation, and clones with the E184D mutation are highlighted in yellow.
2.4.6 Implications for Engineering LHEs

Our study identified a number of coevolving networks in single-chain LHEs, the highest scoring of which involved the catalytic and adjacent residues in the LAGLIDAG α-helices. Because catalytic positions in sequence families are often invariant, these positions are refractory to computational methods designed to identify coevolving networks. However, as the acidic metal-binding residues in the LAGLIDAG family exhibited variation, we were able to identify and experimentally validate a coevolving network that included the two acidic metal-binding residues, E29 and E184, and the two adjacent residues, A28 and G183. Indeed, the importance of these residues to LHE structure and function has not gone unnoted. Silva and Belfort identified the A28 and G183 positions as part of a GxxxG network in LAGLIDAG helices (14) and suggested that residue combinations at these positions were influenced by packing constraints at the base of the helices. Similarly, random mutagenesis screens for increased activity of I-CreI derivatives identified the A28 position in dimeric and monomeric I-CreI derivatives (G19 of I-CreI) (17, 31). Likewise, up-activity mutants of I-AniI included randomly selected E-to-D substitutions in the equivalent position to E184 (32). In all cases, the coevolutionary context of the residues involved was not appreciated. Computational modeling using Rosetta and biophysical analyses revealed that I-LtrI and I-OnuI network variants do not exhibit structural defects, which is further supported by our kinetic analyses showing significant effects on $k_{cat}^*$ and $K_M^*$. The observation that mutation of these positions does not cause structural defects is somewhat surprising, given their location in the core of the molecule and their proximity to the active site. It is commonly thought that the maintenance of protein stability plays a key role in limiting the rate of protein sequence evolution (27, 33); however, the generally low expression of these proteins is likely to mitigate this effect greatly (34). Thus, the coevolving network directly impacts on catalytic function, possibly because mutations in the network subtly affect positioning of the acidic metal-binding residues relative to DNA substrate. It is unlikely that subtle differences in positioning of residues in the network would be readily detectable in I-LtrI, I-OnuI, or I-HjeMI variants by structural studies or computational modeling.
The observation that the coevolving network that we studied lies across the LAGLIDADG interface has obvious implications for engineering of chimeric LHEs. A common strategy for engineering LHEs with altered specificity is to fuse two halves (or domains) from different LHEs, each with different DNA-binding specificity (14, 16, 35–38). This strategy has been successful, particularly when the N- and C-terminal domains are derived from closely related LHEs. Our data indicate that fusion of more distantly related domains may require subsequent fine-tuning, as indicated by the contrasting preferences for acidic catalytic residues in I-HjeMI and I-LtrI and the observation that some variants are more active than wild type. Our data show that the identity of the acidic residue in the first LAGLIDADG helix (position 29 of I-LtrI) is the major determinant to transitioning between residue combinations (Figure 2.6), which would involve multiple substitutions to create inactive or hyperactive intermediates unlikely to be tolerated in a cellular environment. Moreover, the greater diversity of residues at noncatalytic positions that create highly enriched variants in the I-HjeMI backbone also supports the hypothesis that chimeric enzymes created by fusion of distantly related domains will require optimization of the noncatalytic positions.
Figure 2.6: Coevolving residues create a barrier to active-site evolution in LHEs.  

The wild-type quartet of residues in the coevolving network of I-LtrI and I-HjeMI are boxed, with the most frequently observed combinations of residues in the other sequence types shown in Figure 2.1. d connected by green or red lines and arrows. The green lines indicate observed and permissive transitions between quartets, and the dashed red lines indicate infrequently observed and nonpermissive transitions, with the number of amino acid changes required for each transition indicated by a circled number.
2.4.7 Conclusions

Our data support the existence of a network of coevolving residues in LHEs, which can be manipulated to modulate catalytic efficiency over an ~100-fold range. Efforts to re-engineer LHEs for genome-editing applications have met with variable success, and we suggest that this is due in part because such efforts inadvertently created a suboptimal complement of residues within the coevolutionary network that we have identified here. Our analyses predicted additional coevolving networks in LHEs, experimental validation of which may be required to enhance success in re-engineering efforts. More generally, our results suggest that robust, structure-guided alignments will facilitate the identification of coevolving catalytic and noncatalytic residues in other protein families, experimental validation of which would add information to enzyme engineering.
2.5 References


Chapter 3

3 The Existence of Additional Coevolving Networks Between LAGLIDADG Domains, and Attempts to Increase Cleavage Activity of the Domain-Fusion Chimera “LtrOnu”

3.1 Introduction

LAGLIDADG homing endonucleases (LHEs) are an ideal scaffold to create genome-editing reagents because they are small (<50 kDa) and have a high level of target specificity (1). LHEs also have the benefit of containing both DNA recognition and cleavage activity within a single functional unit (2, 3). This is in contrast with other genome-editing platforms, such as zinc-finger nucleases (ZFNs) and TAL effector nucleases (TALENs) that use distinct domains for both DNA recognition and cleavage (4-6). However, the aforementioned platforms, as well as CRISPR/Cas9, have well defined mechanisms of DNA recognition and can be retargeted to a desired sequence with relative ease (4, 7, 8). LHEs on-the-other-hand share no conserved mechanism of DNA recognition (3, 9), which limits our ability to alter target specificity and thus their usefulness as a scaffold for creating genome-editing reagents.

To limit the amount of engineering required to target a desired sequence, distinct LHE domains have been fused to create chimeric enzymes that cleave hybrid target sites (Figure 1. 6) (10-14). Chimeric LHEs have the potential to dramatically increase the number of available target sites by exploiting the reservoir of distinct LAGLIDADG domains that exist in nature (1, 13, 15). This approach has the potential to reduce the engineering burden of LHEs because it would avoid the extensive mutagenesis and selection required to manipulate the protein-DNA interface (13).

Baxter et al. (2012) sought to create LHE chimeras using 6 members of the I-OnuI subfamily. However, of the 30 chimeras produced in this study, only 14 formed stable structures and only 9 showed detectable levels of enzymatic activity. The majority of constructs did not fold correctly, had structural defects and/or had catalytic activity well
below that of the wild-type parental enzymes. The most intriguing example of DFC dysfunction included the two fusions (OnuLtr and LtrOnu) that were created using parental domains from I-OnuI and I-LtrI. While the cleavage activity of OnuLtr is comparable to the parental scaffolds, LtrOnu exhibits defects in expression, binding, and cleavage of the predicted hybrid target site. One explanation for these unexpected results is that structural constraints and relationships exist between LHE domains beyond the conserved interactions of the LAGLIDADG motif residues (13).

We previously identified covarying residues between LHE domains, and demonstrated that the highest scoring positions form a coevolving network that can modulate catalytic efficiency (16). Furthermore, our data suggested that the coevolving network creates a barrier to active site evolution by restricting the fitness landscape of LHEs to a limited number of permissible residue combinations. Given that these residues, and the other top scoring networks, interact at the LHE interface, we reasoned that these positions can affect LHE cleavage activity by influencing the stability of domain interactions. Therefore, we hypothesized that mutation and functional selection of covarying positions at the LHE domain interface could rescue DFC instability, and increase cleavage activity against their predicted hybrid target sites.

3.2 Materials and Methods

3.2.1 Oligonucleotides and Plasmids

Oligonucleotides in this study were synthesized by Integrated DNA Technologies, Inc. All plasmids and substrates were isolated from E. coli cultures grown in Luria Broth (LB) with an EZ-10 Spin Column Plasmid DNA Kit (Bio Basic Inc.) according to the manufacturer’s protocol. The DFC LtrOnu ORF was generated as previously described (13) and cloned between the Ncol and NotI sites of plasmid pEndo (17). Two copies of the LtrOnu hybrid target site were cloned between the AflIII/BglII and NheII/SacII sites of pTox for use in genetic selection assays. Point mutations were incorporated into I-LtrI and I-OnuI using a QuikChange II Site-Directed Mutagenesis Kit (Stratagene) to create a variety of amino acid combinations at covarying positions, and the mutations were confirmed by sequencing.
3.2.2 Bacterial Two-Plasmid Functional Selection

Bacterial two-plasmid functional selections were accomplished as previously reported (16, 17) without modification.

3.2.3 Mutant Library Synthesis

The FLL_E covarying network library was synthesized by saturating mutagenesis of the corresponding positions within the I-LtrI scaffold (Genscript).

Unless otherwise stated, all LtrOnu mutagenic libraries were constructed using customized sewing PCR strategies. Briefly, open reading frame fragments were amplified from pEndo templates using oligonucleotides with randomized codon positions (NNN or NNS to limit nonsense codons). Reaction products were gel purified using a EZ-10 Spin Column DNA Gel Extraction Kit (BioBasic Inc.), and fragments were subsequently sewn together by nested PCR. The final sewing primers containing 5’ NcoI and 3’ NotI sites, as well as 10bp landing pads to facilitate endonucleolytic digestion. Mutagenized open reading frames were then NcoI/NotI digested and ligated into previously prepared pEndo vector. All libraries were cloned with 10-fold coverage to ensure that all expected variants were included, and a minimum of 3 independent clones were sequenced (London Regional Genomics Centre) to ensure the correct amino acid positions were mutagenized.

The LtrOnu suppressor screen library was generated by subjecting the DFC open reading frame to error-prone PCR (2-6 substitutions per kilobase) using a GeneMorph II Random Mutagenesis Kit (Agilent Technologies) and primers that contained a 5’ NcoI site and start codon, as well as a 3’ stop codon and 3’ NotI. Mutagenic amplicons were cleaned up using an EZ-10 Spin Column PCR Products Purification Kit (BioBasic Inc.), NcoI/NotI digested and cloned into previously prepared pEndo vector.

3.2.4 Protein Expression, Protein Purification, and In Vitro Cleavage Assays

LtrOnu over-expression, purification, and in vitro cleavage assays were accomplished as previously reported (16) without modification.
3.2.5 LtrOnu Library Screens

With the exception of the LtrOnu random mutagenesis experiments, LtrOnu libraries (in pEndo) were used as input for a single round of solid media bacterial two-plasmid genetic selection as previously described (16, 17). Three surviving colonies were isolated, subject to individual solid media genetic selection, and the open reading frames were sequenced (London Regional Genomic Centre) to identify residue combinations. LtrOnu random mutagenesis experiments were accomplished by performing an initial of liquid media bacterial two-plasmid genetic selection (as previously described (16)). Samples were plated on solid media selection plates to identify survival percentages but the remainder was grown overnight in a rotary wheel at 37°C. pEndo plasmids were isolated from the overnight culture, the LtrOnu open reading frames were PCR amplified using Taq polymerase (New England Biolabs) flanking primers, and subcloned between the NcoI/NorI sites of fresh pEndo. This round 1 library was used as input for a second round of genetic selection using solid media selections, and three surviving colonies were isolated, subject to individual solid media genetic selection, and the open reading frames were sequenced (London Regional Genomic Centre) to identify residue combinations.

3.3 Results

3.3.1 Additional Covarying Networks are Tolerant to Semi-Conservative Substitutions

The second highest covariation scores from the LHE analysis (Dataset S2.1) were found between positions 21 and 182 (I-LtrI numbering) (16). These two residue positions (wild-type I-LtrI T21 and A182, or T:A) had a Zpx score of 5.0 and were measured to be 6.5Å apart (Figure 3.1A). While these positions were not in direct contact, they do interact between LHE domains and are both found at the LAGLIDADG interface (Figure 3.1B). To assess their ability to impact LHE functionality, positions 21 and 182 were mutated to various combinations that are naturally occurring, and/or would be found in commonly engineered DFCs (Table 3.1).
Figure 3.1: LHE residues with the second highest covariation score.

(A) Dot plot illustrating that positions 21 and 182 (I-LtrI numbering) have a covariation (Zpx) score of 5.0 and are positioned 6.5 Å apart in the I-LtrI structure. (B) Positions 21 and 182 (magenta) mapped onto a structural model of I-LtrI in complex with its cognate DNA target (PDB: 3R7P).
Table 3.1: Covarying residue combinations that are created in LHE domain fusion chimeras. Data was collected from Baxter et al. (2012) and show the residue pairings that are created when the N-terminal domain from one parental scaffold is fused to the C-terminal domain of a second parental scaffold.

<table>
<thead>
<tr>
<th>Parental Domain Scaffold</th>
<th>Covarying Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal</td>
<td>C-terminal</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-LtrI</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>I-A</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-D</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-S</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-D</td>
</tr>
<tr>
<td>I-Gpil</td>
<td>I-LtrI</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-D</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-F</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-D</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-Gpil</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-LtrI</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-S</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-Ggel</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-D</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-Ggel</td>
</tr>
<tr>
<td>I-Gzel</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-Onul</td>
<td>I-Ggpil</td>
</tr>
<tr>
<td>I-PanMI</td>
<td>I-D</td>
</tr>
<tr>
<td>I-SscMI</td>
<td>I-S</td>
</tr>
</tbody>
</table>
Figure 3. 2: Survival of individual 21:182 variants under genetic selection.

Bar graphs showing the survival percentage of (A) I-LtrI and I-OnuI (B) variants in solid media two-plasmid genetic selection. Checkered bars represent the wild-type combination of residues (positive control) in each scaffold, while solid grey bars represent naturally occurring variants or combinations that would be produced in domain fusion chimeras. Note that each selection was also performed using a vector only negative control. Survival percentage is shown ± standard deviation.
Figure 3.3: Survival of FLL_E network variants under genetic selection.

(A) Residues of the FLL_E covarying network (red) mapped onto a structural model of I-LtrI in complex with its cognate DNA target site (PDB: 3R7P). (B) Bar graphs showing the survival percentage of I-LtrI FLL_E network variants in solid media two-plasmid genetic selection. Checkered bar show survival of the wild-type I-LtrI positive control. Solid grey bars represent survival of vector only (negative control), the FLL_E mutagenic library, and three variants that were tested individually after surviving a first round of selection. Survival percentage is shown ± standard deviation.
The effect of mutations to the 21:182 pairing was studied in the I-LtrI and I-OnuI backgrounds using two-plasmid genetic selection. Survival data demonstrate that these positions are tolerant to semi-conservative substitutions because all variants demonstrated some level of survival, with the exception of a F:D pairing in the effects I-OnuI background (Figure 3. 2). In a recently published study by Baxter et al. (2012), position 21 was investigated in the I-OnuI scaffold by saturating its amino acid composition, and selecting for activity using a functional genetic selection (13). In agreement with our results, it was found that functional proteins generally contained a large hydrophobic or non-polar residue. Upon further study, they determined that position 21 is critical for enzyme stability because unfavorable combinations dramatically reduced soluble protein levels in yeast surface-display experiments (13). Although residues 21 and 182 appear to be tolerant to semi-conservative amino acid substitutions, it is significant that these positions were identified in our covariation analysis as being potentially important to LHE structure and/or function.

Given our success with the first two covarying networks, we investigated the third highest set of covarying residues from the LHE covariation analysis. Positions 25, 56, 64, and 184 formed a covarying network with the following Zpx scores and contact distances: positions 25 and 56 had a score of 3.0 and were 3.0Å apart; positions 56 and 64 had a score of 4.8 and were 3.8Å apart; and positions 64 and 184 had a score of 4.2 and were 13.3Å (Dataset S2.1). When the positions were mapped to the structure of I-LtrI (Figure 3. 3A), the side chains of positions F25, L56 and L64 were located in the core of the N-terminal domain, and position E184 is the C-terminal domain catalytic residue that coevolves with other active site residues (given the residue identities were F, L, L and E, the network will be referred to as FLL_E).

To investigate the functional impact of the FLL_E network, positions 25, 56, and 64 were subject to saturating mutagenesis in the context of position 184 being held to D or E (library complexity of 16 000). The library was screened using solid media bacterial two-plasmid selection, and then individual survivors were isolated, sequenced and re-tested individually. Library survival in the initial screen was 2.96%, and all three surviving clones demonstrated wild-type levels of survival when tested individually (Figure 3. 3B).
Sequencing results demonstrated that the clones had residue identities of FFL_D, FLP_E, and FLL_E. These combinations include the wild-type network, and the most abundant pairings within the LHE family (data not shown). Given their position within the protein core (25 and 56) and within a disordered loop region (64), the isolated residue combinations were not unexpected. Although residue identity is restricted between these positions (less than 3% library survival), they were not subject to further study because it is assumed that they impact intra-domain (N-terminal) stability rather than interactions between LHE domains.

3.3.2 LtrOnu Shows Detectable Cleavage Activity But its Stability is Compromised

Provided that the DFC LtrOnu has previously demonstrated poor stability and cleavage activity in yeast surface display (13), we purified the scaffold to facilitate *in vitro* analysis. A his-tagged version of LtrOnu purified using a standard HisTrap protocol, and buffers that have been successfully used for the purification of numerous LHE variants. Following elution from the column there was no detectable peak fraction even though there was a clear induction band following over-expression (data not shown). Given the evidence of protein instability, the purification procedure was repeated with 4L of induced culture (rather than 1L), and this resulted in clearly defined peak fractions following elution. Following concentration of the final preparation, SDS-polyacrylamide gel electrophoreses (PAGE) visualization showed a clear band at the expected molecular weight of 34.6kDa (Figure 3. 4A). Unfortunately, over 95% of the sample was lost following cell lysis, and the sample continued to demonstrate concentration-dependent precipitation. Before the sample was completely lost, fixed end-point endonuclease assays were performed using a plasmid substrate containing the predicted hybrid target sequence (5’-AATGCTCCTATTCAACCTTTTA-3’). Interestingly, LtrOnu demonstrated cleavage activity but the fusion only linearized 30% of the provided substrate in 30min (Figure 3. 4B). This demonstrated that the DFC retains cleavage activity against its predicted target site, but the structural integrity appears to be compromised.
Figure 3. 4: The DFC LtrOnu is over 95% insoluble in standard LHE purification buffers.

(A) 12% SDS-PAGE gel electrophoresis image of a HisTrap affinity chromatography purification of LtrOnu. Lanes are as follows: (1) molecular weight ladder; (2) Soluble fraction following cell lysis; (3) Insoluble fraction following cell lysis; (4) binding buffer wash; (5) 30mM imidazole wash; (6) empty; (7-10) peak elution fractions; (11) post-TEV protease digestion HisTrap flow-through; (12) final concentrated preparation. (B) 1% agarose-TAE gel electrophoresis image of an LtrOnu endonuclease assay. 500mM LtrOnu was reacted with 5nM supercoiled substrate for 30min at 37°C. Lanes are as follows: (1) molecular weight ladder showing 3kb band; (2) EcoRI-digested plasmid substrate size marker; (3) no protein negative control; (4) LtrOnu reaction.
3.3.3 Saturating Covarying Residue Positions Cannot Increase the Cleavage Activity of LtrOnu

Given that I-LtrI displays detectable cleavage activity, we mutagenized the interfacial coevolving networks and screened for variants with rescued phenotypes using two-plasmid functional selection. The first network to be tested was the previously identified active site coevolving residue quartet. LtrOnu residues corresponding to I-LtrI positions 28, 29, 183 and 184 were saturated and the library was initially screened for activity, and surviving clones were sequenced and retested individually (Note that the nomenclature for this network was revised for simplicity. Variants will hence forth be named according to their numeric residue numbering such that the native LtrOnu scaffold A28, E29, G183 and E184 is referred to as “AEGE”). Library survival was not found to be increased above native LtrOnu (AEGE) levels, and the three isolated clones (WEPE, AEYD, and PESD) did not survive above backgrounds levels when tested individually (Figure 3.5).

We next decided to test the effect of saturating residue positions 21:182 in the LtrOnu scaffold. Given the low complexity of this library, we also decided to include positions 27 and 174 (referred to as the helix stabilization network [HSN] library) to increase the screen to 160 000 variants (Figure 3.6A). Although these positions were conserved and formed a stable planar-polar interaction in the structure of wild-type I-LtrI, they were chosen because these positions are the reciprocal positions of positions 21:182 at the domain interface. Therefore, these residues occupy the same positions as 21:182 would be expected to fill in homodimeric LHE enzymes. Given HSN complexity, the low levels of library survival were not unexpected, but similarly no increase in survival was witnessed for the three clones that were tested individually following the initial screen (Figure 3.6B).
**Figure 3.5: Coevolving network substitutions cannot rescue LtrOnu cleavage activity.**

Bar graphs showing the survival percentage of LHE variants in solid media two-plasmid genetic selection. Wild-type I-LtrI, wild-type I-OnuI, and vector only act as negative controls for the hybrid LtrOnu target. AEGE is the network combination in the original LtrOnu scaffold, library represents the LtrOnu coevolving network library survival, and the three other LtrOnu variants were tested individually after surviving an initial library screen. Survival percentage is shown ± standard deviation.
Figure 3.6: Helix stabilization network (HSN) residue saturation cannot increase LtrOnu cleavage activity.

(A) Residues positions 21 and 182 (magenta), as well as 27 and 174 (cyan) mapped onto a structural model of I-LtrI in complex with its cognate DNA target site (PDB: 3R7P). Note that only the two LAGLIDADG helices are shown for clarity. (B) Bar graphs showing the survival percentage of LHE variants in solid media two-plasmid genetic selection. Wild-type I-LtrI, wild-type I-OnuI, and vector only act as negative controls for the hybrid LtrOnu target. LtrOnu is the original scaffold, library represents the survival of the HSN library following an initial screen, and Survivors 1-3 were tested individually after surviving the initial library screen. Survival percentage is shown ± standard deviation.
3.3.4 Additional Mutagenesis Strategies Cannot Increase the Cleavage Activity of LtrOnu

Two attempts were then made to rescue the activity of LtrOnu that did not involve covarying residue positions. First, LtrOnu was subject to a suppressor screen by randomly mutating the open reading frame, and selecting for functional variants using two rounds of bacterial two-plasmid genetic selection. The first round of library selection did not yield survival above background levels (Figure 3. 7), but pEndo plasmids were recovered from overnight liquid media selections, and LtrOun mutagenic open reading frames were subcloned into fresh pEndo vector backbones. This round 1 library was then used as input for a second round of solid media two-plasmid genetic selection, and individual clones were sequenced and retested individually. Round 2 library survival was not found to increase compared to round 1, and there was no suppression of the phenotype by either of three isolated clones: (i) no sequence data was obtained after two attempts; (ii) 1N9K/Y105C/I172V/G183A; and (iii) K40R/I69S).

Finally, analysis of the I-LtrI domain interface revealed four N-terminal domain interfacial residues (positions W20, I22, T23, and L64) that contact the C-terminal domain (Figure 3. 8A). In addition, none of these positions were involved with a covarying residue network(16), and they have been previously found to influence LHE cleavage activity (13, 14). The four residues were saturated (160 000 variants), and it was named the Helix-1 interfacial contacting (HIC) library even though position 64 was not in LHE helix1. Once again, library survival was not significantly above background, and no significant survival was witnessed for either of the three clones re-tested following the initial screen: (i) S20/W22/T23/T64; (ii) V20/P22/P23/V64; (iii) Y20/T22/F23/V64 (Figure 3. 8B).
Figure 3. 7: A suppressor screen was unable to increase LtrOnu cleavage activity.

Bar graphs showing the survival percentage of LHE variants in solid media two-plasmid genetic selection. Wild-type I-LtrI, wild-type I-Onul, and vector only act as negative controls for the hybrid LtrOnu target. LtrOnu is the unmutagenized scaffold, library round 1 is mutagenic library survival after an initial screen, library round 2 is survival of enriched variants from round 1, and R2 Survivors 1-3 are variants that were tested individually after surviving two rounds of genetic selection enrichment. Survival percentage is shown ± standard deviation.
Figure 3.8: Helix-1 interfacial contacting (HIC) residue saturation cannot increase LtrOnu cleavage activity.

(A) Bird’s eye view of residues positions W20, I22, T23, and L64 (orange) mapped onto a structural model of I-LtrI in complex with its cognate DNA target site (PDB: 3R7P).

(B) Bar graphs showing the survival percentage of LHE variants in solid media two-plasmid genetic selection. Wild-type I-LtrI, wild-type I-Onul, and vector only act as negative controls for the hybrid LtrOnu target. LtrOnu is the original scaffold, library represents the survival of the HIC library following an initial screen, and Survivors 1-3 were tested individually after surviving the initial library screen. Survival percentage is shown ± standard deviation.
3.4 Discussion

Once again, covariation scores determined with the Zpx method (18, 19) correctly identified sets of interacting residue pairs in LHE family enzymes. Residue identity within the two additional networks (21:182 and FLL_E) is restricted to semi-conservative alterations, and residue combinations that are found naturally within the LHE family are generally permissive to function in the I-LtrI and I-OnuI backgrounds. These findings support LHE scaffolds being tolerant to amino acid substitutions while retaining cleavage activity against their cognate recognition sequences (11). However, these covarying networks did not restrict LHE activity comparable to the previously studied coevolving residue quartet, and it is unclear whether two, or four, random residue combinations would exhibit a similar level of combinatorial permissibility. Considering that the coevolving network quartet contains two catalytically essential residues, it is not surprising that comparatively, the 21:182 and FLL_E networks do not appear to impose a barrier to the LHE fitness landscape. Accordingly, the covariation scores for these two networks were 2 standard deviations lower than the coevolving quartet, and illustrate the importance of biologically validating covariation scores to support the identification of coevolutionary relationships.

Similar to the coevolving network quartet, positions 21:182 was unable to rescue activity of the dysfunctional DFC, LtrOnu. Additionally, suppressor screens using random mutagenesis, and saturating mutagenesis of interfacial residues were also unable to increase the cleavage activity of LtrOnu. These findings suggest that modifications to the domain interface are insufficient to rescue the functionality of problematic DFCs. However, it is possible that more extensive directed evolution experiments (20) could rescue LtrOnu by exploring a broader range of the amino acid sequence space, and drastically altering the protein structure. While those studies could provide valuable information regarding LHE biology, the blunt force nature of that method contrasts with our attempts to pin-point potential defects by identifying biologically important residue interactions from sequence information.

LtrOnu demonstrates cleavage activity (albeit low levels), but structural integrity of the chimeric enzyme is clearly compromised. Therefore, a more realistic path forward would
be to direct the evolution of LtrOnu variants towards increased stability, rather than
directly trying to rescue cleavage activity. Although we were unsuccessful at rescuing
LtrOnu activity, results from this study indicate that conserved, yet unidentified,
interactions may still exist between domains at the LHE interface.
3.5 References


14. Silva GH, Belfort M, Wende W, Pingoud A. From monomeric to homodimeric


Chapter 4

4 Catalytic Residue Substitutions Increase LAGLIDADG Homing Endonuclease Cleavage Activity Against Non-Cognate Central Four (C4) Substrates

4.1 Introduction

LAGLIDADG homing endonucleases (LHEs) are a class of mobile genetic element that introduce double-stranded breaks (DSBs) at highly specific 22-bp recognition sequences (Figure 4.1A) within single-cell, mitochondrial, and chloroplast genomes. Natural cleavage events catalyzed by LHEs function to activate host DNA repair machinery, and facilitate the precise insertion of their open reading frames without impacting host cell survival. Given their small size and incredible target specificity, LHEs are ideal genome editing reagents but their applicability has been limited by difficulties in re-engineering their cleavage activity towards any sequence of interest.

LHE binding specificity is largely a function of direct or water-mediated contacts to bases outside of the central four (C4) region of the target site. Although multiple rounds of mutagenesis may be required, base pair recognition at these positions is readily re-engineered using well established directed evolution experiments. In contrast, LHE cleavage activity is governed by indirect readout of DNA sequences in the C4 base pairs, and regulation of this mechanism is less well understood. Recent studies of LHE indirect readout have highlighted the role of protein-induced DNA bending because the energy required to deform C4 base pairs is sequence-dependent. As a result, DNA backbone atoms adopt a sequence-specific conformation within the LHE active site, and this is recognized by precisely aligning the catalytic centre towards the unique shape of a cognate substrate.

For these reasons, many LHEs are intolerant of changes to the C4 bases, and this has severely limited their sequence-targeting capacity compared to other genome editing technologies, namely CRISPR-Cas9. Interestingly, substitutions to the C4 region have been shown to impact LHE cleavage activity while often maintaining affinity for the
DNA sequence. Given that the previously identified coevolving network in LHEs includes active site residues (Figure 4. 1B) and modulates catalytic efficiency, we hypothesized that changing residues in the coevolving network residues could influence LHE catalytic efficiency on substrates with C4 combinations poorly cleaved by wild-type LHEs.

To pursue this line of investigation, mutagenic coevolving network libraries were screened for activity against various C4 sequences using selective growth experiments (SELEX). The C4 cleavage preference of isolated LHE variants was then profiled and compared to the wild-type enzymes to assess the effect of mutations on the sequence targeting capacity of each scaffold. Unfortunately, we did not have access to the elaborate infrastructure commonly used to determine LHE cleavage preference, and by tethering protein and substrates into close proximity, these methods limit the direct comparison substrate activity levels. Therefore, we decided to develop high-throughput methods to profile LHE cleavage activity simultaneously against all 256 possible C4 sequence combinations. The result was two techniques that identify permissive C4 sequences by their accumulation in cleaved product pools, or by their time-dependent loss from the input library substrate pool. Finally, x-ray crystallography was used to identify the structural basis of the phenotypic differences between coevolving network variants on both the cognate and non-cognate C4 sequences.
Figure 4.1: Schematic of I-LtrI, the coevolving network, and central four nucleotides.

(A) Structural model of I-LtrI bound to its cognate DNA target site and schematic representation of the 22-bp cognate sequence. The left side of the I-LtrI target is composed of nucleotide positions -11 to -1, and the right side of the target is composed of positions +1 to +11. (B) Close-up of the I-LtrI LAGLIDAG helices, coevolving network residues and the central four (C4) nucleotides (gold). A28 (green) and G183 (magenta) are represented by spheres, E29 and E184 (blue) are represented by sticks, and secondary structure elements (grey) are denoted by ribbons.
4.2 Materials and Methods

4.2.1 Oligo Nucleotides and Plasmids Protein Libraries

All oligonucleotides used in this study were synthesized by Integrated DNA Technologies (IDT), Inc. unless otherwise stated. All plasmid DNA was isolated from *E. coli* cultures grown in Luria Broth (LB) and isolated using an EZ-10 Spin Column Plasmid DNA Kit (Bio Basic Inc.) according to the manufacturer’s protocol, unless otherwise stated. Substrate for all kinetic analyses were isolated using Plasmid DNA Extraction Maxiprep Kit (Bio Basic Inc.) according to the manufacturer’s protocol. I-LtrI and I-OnuI encoding genes (codon-optimized for *E. coli*) were cloned between the *NcoI* and *NotI* sites of plasmid pEndo (1) for genetic selection assays, or pProExHta (Addgene) for protein purification. I-LtrI and I-OnuI target sites were synthesized by IDT Inc., or ThermoFisher Scientific, and cloned between the *NheI* and *SacII* sites of the pTox plasmid or genetic selection assays, or between the *EcoRI* and *BamHI* sites of pLitmus28i (Addgene). Individual I-LtrI and I-OnuI coevolving network variants were constructed as previously described (2).

4.2.2 Mutagenic Libraries

Unless otherwise stated, all I-LtrI and I-OnuI mutagenic libraries were constructed using customized sewing PCR strategies. Briefly, open reading frame fragments were amplified from pEndo templates using oligonucleotides with randomized codon positions (NNN or NNS to limit nonsense codons). Reaction products were gel purified using a EZ-10 Spin Column DNA Gel Extraction Kit (BioBasic Inc.), and fragments were subsequently sewn together by nested PCR. The final sewing primers containing 5’ *NcoI* and 3’ *NotI* sites, as well as 10bp landing pads to facilitate endonucleolytic digestion. Mutagenized open reading frames were then *NcoI/NotI* digested and ligated into previously prepared pEndo vector. All libraries were cloned with 10-fold coverage to ensure that all expected variants were included, and a minimum of 3 independent clones were sequenced (London Regional Genomics Centre) to ensure the correct amino acid positions were mutagenized. Note that the I-LtrI coevolving network library was constructed as previously described (2).
4.2.3 Bacterial Two-Plasmid Functional Selection

Bacterial two-plasmid functional selection (1) was used to screen the activity of all LHE variants and libraries used in this study. All solid media (plate) selections were accomplished as previously reported (2).

For selective growth experiments (SELEX) in liquid culture, 20 ng of LHE variant (in pEndo) were transformed into 50 µL of chemically competent NovaXGF’ (Novagen) cells harboring the appropriate pTox plasmid (i.e. the correct I-LtrI or I-OnuI target site and central four sequence). Transformants were allowed to recover in 2mL of 2× YT medium (16 g/L tryptone, 10 g/L yeast extract, and 5 g/L NaCl) at 37 ºC in a rotary wheel for 30 min. Cultures were then supplemented with 2mL of 2× YT, arabinose (0.02% final) and carbenicillin (100 µg/mL final) to induce endonuclease expression. Following recovery and a 1hr expression period, both I-LtrI and I-OnuI cultures were diluted 200-fold into either nonselective [1× M9 salt, 0.8% w/v tryptone, 1% vol/vol glycerol, 1 mM MgSO4, 1 mM CaCl2, 0.2% w/v thiamine, 100 µg/mL carbenicillin, and 0.02% (w/v) L-glucose] or selective media [nonselective media lacking glucose and with the addition of 0.02% (w/v) L-arabinose and 0.1 mM isopropyl β-D-1-thiogalactopyranoside]. Cultures were then grown in a rotary wheel at 37ºC for 16hrs before cells were pelleted and pEndo plasmids were isolated.

4.2.4 Protein Expression and Purification

All protein over-expression was performed from pProExHta (Addgene) templates and in ER2566 (New England Biolabs) as previously described (2). I-LtrI purifications intended for kinetic analyses were accomplished as previously reported (2), while purifications for crystallography were modified as follows:

Following the 16hr incubation at 4ºC, I-LtrI variants were purified from Tobacco Etch Virus (TEV) protease by cation exchange chromatography using a 5-ml HiTrap SP HP column (GE Healthcare) on an AKTA fast-performance liquid chromatograph (FPLC). Peak fractions were then pooled and incubated with 50mM Ethylenediaminetetraacetic acid (EDTA) at 4ºC for a minimum of 4hrs to remove residual divalent metals. Preparations were finally exchanged into storage buffer (250mM NaCl, Tris-HCl, pH 8.0,
10%[v/v] glycerol, and 30mM CaCl$_2$) before concentrating to ~5mg/mL (30 mL Amicon filters, Merck), and storage at −80°C.

Crystallization, data collection, and structural determination methods are described in Appendix S3.

4.2.5 Endonuclease Assays and Pseudo-Michaelis-Menten Kinetics

The endonuclease assays and kinetics experiments in this study were performed as previously reported (2), and are based on the single-turn over kinetic models described by Halford and colleagues (3).

4.2.6 Product Enrichment and Substrate Depletion Assays

I-LtrI and I-OnuI recognition sequences with randomized C4 regions (I-LtrI 5’-AATGCTCCTNNNGACGTITTA-3’; I-OnuI 5’-TTTCCACTTNNNNACCTTTTA-3’) were first cloned between the EcoRI and BamHI sites of pLitmus28i (Addgene). These plasmid libraries were then used as input for time-course endonuclease assays using purified I-LtrI or I-OnuI variants. Reaction products were subsequently separated from unreacted substrates using 0.9%-TBE gel electrophoresis at 3.5V/cm for 75 minutes. Linearized product, or supercoiled substrate bands were then cut out of the gel matrix, and isolated using EZ-10 Spin Column DNA Gel Extraction Kits (Bio Basic Inc.). Following elution from the columns, linearized plasmid DNA was re-circularized using T4 DNA Ligase (New England Biolabs) to restore an intact recognition site. The C4 region of unreacted supercoiled substrates, re-circularized products, and input library were then amplified by GoTaq® Hotstart PCR (Promega) using customized barcoding primers that planked the recognition site. Amplicon length was verified by 1%(w/v) agarose-TAE gel electrophoresis, and equimolar volumes of each sample were then pooled and subject to high-throughput sequencing.
4.2.7 High-Throughput Sequencing (HTS) and Compositional Data Analysis (CoDa)

Abundance of coevolving network variants from the liquid genetic selection assays were measured using paired-end sequencing on the Illumina NextSeq platform at the London Regional Genomics Centre (London, ON), or the Illumina HiSeq platform at The Centre for Applied Genomics (SickKids, Toronto, ON). Genetic selection based reads were parsed for the presence of D or E at positions 29 and 184, and the number of reads for all possible coevolving network variants was identified using a custom Perl script (G. Gloor).

Abundance of C4 substrates in the input libraries, and substrate or linearized product pools were measured using paired-end sequencing on the Illumina NextSeq platform at the London Regional Genomics Centre (London, ON). Reads were parsed for the correct I-LtrI or I-OnuI target sequence between nucleotide positions -11 and -4, and read counts for each central four (C4) substrate was identified using a custom Perl script (G. Gloor).

The proportional abundance of coevolving network variants (in the selected versus non-selected condition), enriched central four products (in linearized pool versus input library), or depleted central four substrates (substrate pool versus input library) within the sequenced libraries was determined using the Bayesian CoDa tool ANOVA-Like Differential Expression 2.0 (ALDEx2) (4, 5). Briefly, the underlying per-feature technical variation within each sample was estimated by taking 128 Monte-Carlo sample instances from a Dirichlet distribution for all datasets. Instances within each distribution were subsequently converted into proportions using a centre-log ratio ($clr$) transformation. This transformation normalizes instances to a common scale (typically the geometric mean) and removes the dependency between features which allows for standard statistical analyses. Note that $clr$ transformations were accomplished using the inter-quartile log-ratio ($iqlr$) denominator method to correct, or mitigate, the influence of asymmetry (read counts of 0) within each dataset (unpublished). Finally, Welch’s $t$ tests were performed on each instance and the raw $p$-values were Benjamini-Hochberg adjusted to correct for multiple hypothesis testing and determine the false discovery rate (FDR) for each feature ($q$-value) (5-7). In addition to the statistical analysis, ALDEx2 generates other important
descriptive information that can be used to compare the two groups of data (experimental versus control): the median log2 difference in clr values between each group (diff.btw) which represents the magnitude of differences between features within each group; the median of the largest log2 difference in clr values within each group (diff.win) which represents the intra-condition variation; the median log2 effect (diff.btw/max[diff.win]) which represents both the magnitude of clr differences between each group and the confidence one has in the inferred differences; and the overlap or proportion of effect size distributions that overlap. Unless otherwise stated, data were called significant if they had an absolute diff.btw greater than 1, an absolute effect greater than 2, a FDR (or q-value) less than 0.1, and an overlap of less than 1%. Following the CoDa procedure, data were plotted using basic R and the ggplot2 package (8).

4.2.8 Structural Analysis

Structural models of the individual Protein Data Bank (PDB) files were visualized using Coot (9) and MacPymol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) to produce the figures in this study. The structural parameters of DNA substrates were analyzed using 3DNA (10) software on the web-3DNA (11) web server (http://w3dna.rutgers.edu/), and outputs plotted using R. Analysis of all protein-DNA interfaces was accomplished using the 'Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html).
4.3 Results

4.3.1 An *in vivo* Selection Identifies LHE Variants Active on Substrates with C4 Substitutions

To identify LHE variants that display increased cleavage activity on non-cognate C4 sequences, libraries of 1600 coevolving network variants were created by randomizing residue positions 28 and 183 (I-LtrI numbering will be used for simplicity) to all twenty amino acids, while the metal-binding residues at positions 29 and 184 were held to either Asp (D) or Glu (E). LHE variants are identified by amino acid combinations, with the residues ordered 28, 183, 29, and 184. In this classification scheme, the wild-type I-LtrI protein is AEGE. Libraries were constructed in both the I-LtrI and I-OnuI backgrounds and first screened for activity against C4 sequences using a strategy involving selective growth experiments (SELEx), high-throughput sequencing (HTS) and compositional data analysis (CoDa) as previously described (2, 4) and outlined in (Figure 4. 2). In our SELEx strategy, I-LtrI or I-OnuI target sites are cloned into toxic plasmids (pTox) with nucleotide substitutions in the C4 region (DNA sequences are identified by an underline, i.e. ATAC). Competent cells containing a pTox variant are then transformed with the corresponding LHE library (I-LtrI or I-OnuI), and recovered cultures are split and grown in either selective (experimental) or nonselective (control) media. Active LHE-substrate combinations permit bacterial growth in selective media because cleavage of the target site induces exonucleolytic degradation of the toxic plasmid and prevents expression of a DNA gyrase inhibitor (1). In contrast, non-permissive LHE-substrate combinations do not cleave the toxic plasmid, and are bacteriostatic due to expression of the DNA gyrase inhibitor (1). Over time, exponential differences in growth rate lead to differences in LHE variant enrichment within selected populations, whereas the cultures grown under nonselective conditions control for disparities in LHE variant copy number within the input library. Following outgrowth, HTS is used to readout LHE variant abundance within the experimental and control populations, and the data is quantified using the CoDa tool ALDEx2.0 (4, 5). This method of analysis allows us to determine changes in the relative abundance (enrichment) of LHE variants in selective versus nonselective conditions as well as to assess the significance of enrichment by determining variability between experimental replicates. Variants were called significant if they had a positive
enrichment value, an effect size greater than 2, an overlap < 1%, and false-discovery rate less than 0.1. Note that SELEX data is presented using strip charts for convenience but these plots do not illustrate variance or statistical confidence. With that said, variants with the largest enrichment values generally have the largest effect sizes, and lowest false discovery rates (Figure 4.3 a).

We individually screened the I-LtrI and I-OnuI variant libraries for activity against 26 and 19 C4 substrates that typically differed by 1 or 2 bases from the I-LtrI (ATAC) or I-OnuI (ATTC) cognate substrates (Figure 4. 3 b and c). In agreement with previous experiments (2), we found that numerous LHE variants demonstrated activity against the cognate I-LtrI and I-OnuI substrates. Interestingly, one variant in both protein backgrounds (E184D or AEGD) was significantly enriched relative to all other variants against 6/26 non-cognate I-LtrI C4 substrates (AAAC, AACC, AAGC, ACAC, ATAT, ATCC) and 5/19 non-cognate I-OnuI C4 substrates (AATC, AGTC, ATAT, ATTT, TTTC). No significant enrichment of LHE variants was found against the remaining C4 substrates, but it is possible that longer outgrowth times than used in these experiments (16 hrs) may yield active variants. We next expanded the I-LtrI library screens to an additional 24 C4 substrates that differed by 2 to 4 bases from the ATAC cognate substrate, many of which were GC-rich and shown to be poor substrates for LHEs (Figure 4. 4). We found significant enrichment of the I-LtrI AEGE wildtype and AEGD variant against 1 of the 26 additional substrates (TTTC).

To confirm the results from the SELEX screens, the AEGE wildtype and AEGD variant were individually tested against various C4 substrates against which they were enriched. This was accomplished using a well-described two-plasmid genetic selection (1) where survival is estimated by the ratio of colonies on selective versus non-selective plates (Figure 4. 5). We found that enrichment values of LHE variants in the SELEX experiments were generally predictive of LHE variant survival against individual C4 substrates, although there was no direct correlation between enrichment values and survival. For instance, in the I-LtrI background, both variants conferred 100% survival against ATAC (cognate) and AAAC C4 substrates, while the AEGD variant conferred significantly greater survival against AACC, AAGC, ACAC, ATAT and ATCC. A
similar trend was found for I-OnuI where both variants conferred 100% survival against the ATTC (cognate) and TTTC substrates, while the AEGD variant demonstrated significantly greater survival against AATC, AGTC, ATAT, and ATTT. Additional LHE variants (ADRE, IEVE, EEAE, TDGD, VDGD and DEGD) were found to be enriched above background levels in other I-LtrI (AGGT, CGGC, TCCG, TTTC, ACCC, ACTC, AGTC, CTAC, ATTA) and I-OnuI (AGGG, ATTA) non-cognate substrate SELEX screens. However, these coevolving network variants are not found naturally in any wild-type LHE and none demonstrated significant survival when tested individually against their identified target (Figure 4. 6).

The main conclusions from the in vivo selection experiments are:

1. Many I-LtrI and I-OnuI variants with substitutions in the coevolving network are active against the cognate substrates, consistent with these scaffolds being tolerant of substitutions while remaining active against their native substrates.

2. In support of previous findings (2), there is a strong association between phylogenetic presence of a coevolving network variant and catalytic efficiency (i.e. variants not found naturally within the LHE family demonstrate poor activity).

3. The E184D substitution (typically AEGD) was enriched against multiple C4 substrates in both the I-LtrI and I-OnuI backgrounds, including substrates on which the wild-type enzymes show no or substantially reduced activity. This finding is consistent with E184D having increased catalytic activity, and is in agreement with previous studies where directed evolution was applied to identify more active LHE variants (12).

4. Substrates with 2 or fewer substitutions in the C4 region are preferred, whereas substrates with 3 or more substitutions are generally not tolerated.
Figure 4.2: Strategy used to screen LHE in vivo cleavage activity.

Schematic flow-through of the experimental strategy used to screen the LHE variant library for activity against various C4 target sites. The approach includes library construction, selective growth experiments (SELEx), high-throughput sequencing (HTS), and compositional data analysis (CoDa).
Figure 4.3: Most LHE variants show poor activity on non-cognate substrates.

(A) Visualizing SELEX using a strip chart, effect plot (Difference Between vs Difference Within), and volcano plot (Difference Between vs False Discovery Rate). Strip-plots illustrating the Enrichment (log2) of 1600 I-LtrI (B) and I-Onul (C) variants against various central four (C4) substrates. Sequences typically differ by 1 or 2 bases from the corresponding cognate site. Wild-type (cyan) and E184D variants (red) are denoted by diamonds while significantly enriched variants (effect >2, overlap <1%, and FDR <0.1) are shown with magenta circles.
Figure 4.4: G-C rich C4 sequences are nonpermissive to I-LtrI activity.

Strip-plots illustrating the Enrichment (log2) of 1600 I-LtrI coevolving network variants against various central four (C4) substrates that typically differ 2 to 4 bases from the corresponding cognate sequence. Wild-type (cyan) and E184D variants (red) are denoted by diamonds and significantly enriched variants (effect >2, overlap <1%, and FDR <0.1) are coloured magenta. Note that no I-LtrI variant was found to be significantly enriched in this experiment.
Figure 4.5: Enrichment values are predictive of survival under functional selection.

Bar-plots showing the in vivo selection survival of wild-type (cyan) and E184D (red) I-LtrI (upper) and I-Onul (lower) against various central four (C4) substrates. Survival percentage (%) is presented ± the standard deviation.
Figure 4.6: Enrichment does not always agree with functional selection data.

Bar-plots showing the *in vivo* selection survival of various I-LtrI coevolving network variants against central four (C4) substrates upon which they enriched above background levels. Survival percentage (%) is presented ± the standard deviation.
4.3.2 Simultaneous Activity Assays on All 256 C4 Substrates Identifies LHE Cleavage Preferences

The \textit{in vivo} SELEX screens established that the AEGD variant is active against a more diverse range of C4 substrates than the wildtype enzyme in two LHE backgrounds (I-LtrI and I-OnuI). However, because it was not practical to test all C4 substrates \textit{in vivo}, we devised a high-throughput \textit{in vitro} competition assay to profile LHE activity on all 256-possible C4 sequences simultaneously (Figure 4. 7). I-LtrI and I-OnuI target sites with randomized C4 sequences were cloned into plasmid substrate libraries and each protein variant was purified using affinity chromatography. These reagents were then used as input for time-course endonuclease assays, followed by gel electrophoresis to separate uncleaved supercoiled C4 substrates from nicked and linearized reaction products. The supercoiled substrate pool and linearized product bands were subsequently purified, subject to high-throughput sequencing and analyzed using the CoDa tool ALDEx2 (n=5). In addition to the experimental libraries, input libraries were also sequenced to determine the relative C4 substrate abundance before exposure to enzyme. This CoDa strategy enabled the identification of permissible C4 substrates by two methods:

1. Product enrichment assays compared the relative abundance of fully linearized C4 sequences to their relative abundance in the input library (i.e. relative abundance of sequence X in linearized product band > relative abundance of substrate X in the input library).

2. Substrate depletion assays identified C4 sequences that were removed, or depleted, from the supercoiled substrate pool over time (i.e. relative abundance of supercoiled substrate X at t= 5min < relative abundance of substrate X in the input library). Substrate abundance may be compared at a fixed time-point or, multiple time-points can be sampled to determine a rate of substrate depletion.

After incubating I-LtrI variants with the C4 substrate library in reaction buffer for 15mins, 5% (13/256) and 9% (22/256) of C4 sequences were enriched (log2 Enrichment
>1) by AEGE and AEGD respectively (Figure 4. 8A). In addition, all I-LtrI enrichment values greater than 1 were statistically significant with effect sizes (log2) of 3 or higher. AEGD was found to enrich all C4 sequences permissible to AEGE as well as 10 other sequences that were poor substrates for AEGE (AAGT, CTAC, CTCC, GAAC, GAGC, GCAC, GCTC, GTAT, GTCC, GTGC). While the top 10 enriched substrates for AEGE (ATAC, TTAC, AACC, ATCC, ACAC, AAAC, AAGC, GTAC, AGAC, and ATGC) and AEGD (GTAC, ATAC, ATCC, AAGC, AAAC, AACC, TTAC, GTCC, TTCC and ACAC) only differed by two sequences, the sequence rank order was distinct for each enzyme (Figure 4. 8B). Consistent with the in vivo data, 100% (13/13) and 86% (19/22) of AEGE and AEGD substrates contained 2 or fewer substitutions while no sequence with 4 substitutions was permissible to activity. When considering nucleotide preference at each C4 position in the top quartile of enriched substrates, over 60% (39/64 and 42/64) of sequences contained a C at the +2 positions for both AEGE and AEGD (Table 4. 1). In contrast, no other position demonstrated a nucleotide frequency of more than 50% and, no significantly enriched substrates contained G-C dinucleotide steps at the -1/+1 positions (i.e. NCCN, NCGN, NGCN and NGGN were poor I-LtrI substrates).

Levels of I-LtrI product enrichment were then compared to the corresponding rate of substrate depletion for each C4 sequence (Figure 4. 9). In agreement with our initial findings, C4 substrates that were significantly enriched by each enzyme also showed the highest rates of substrate depletion. However, unlike the enrichment assay, the depletion method was not limited by the accumulation of product and so reactions were sampled at time points where differences between the two enzymes was greatest. This comparison demonstrated that AEGD was more active on all significantly enriched substrates and, depleted the cognate C4 substrate (ATAC) at a relative rate 10x higher than AEGE.

Differences in substrate permissibility between AEGE and AEGD were even greater in the I-OnuI background (Figure 4. 10A). After 30mins of enrichment, 8% (21/256) and 19% (48/256) of C4 substrates were enriched (log2 Enrichment >1) by AEGD and AEGD respectively. However, while all enrichment values were statistically significant for AEGD (log2 effect sizes >4), the enrichment of substrates by AEGE was not significantly different from the variability in abundance between experimental replicates.
In addition to all substrates estimated to be permissible to AEGE activity, I-OnuI AEGD significantly enriched an additional 27 C4 sequences including ATAT, ATCT, CCTC, CTTC, GCAC, GGAC, GTCC and GTTT (Figure 4. 10B). The top 10 substrates for I-OnuI AEGE (TTTC, ATAC, ATTC, AATC, ACTC, AGTC, AAGC, AAAC, AGAC) and AEGD (GCTC, ATGC, GTTC, AGTC, ACTC, ACAC, ATAC, AGAC, AATC, GGTC) differed by 4 sequences and the substrate rank order was distinct for each enzyme (Figure 4. 10B). Consistent with the I-LtrI and in vivo data, 100% (21/21) and 88% (42/48) of I-OnuI AEGE and AEGD substrates contained 2 or fewer substitutions while no sequence with 4 substitutions was permissible to activity. Similar to I-LtrI, 64% and 67% of permissible I-OnuI substrates contain a C at the +2 position for AEGE and AEGD, while only 8% (4/48) of significantly enriched sequences contain a G-C dinucleotide step at the -1/+1 positions (Table 4. 1).

Once again, C4 substrates with the largest enrichment values for I-OnuI AEGE and AEGD also show the highest rates of substrate depletion (Figure 4. 11), and AEGD depletes all permissible substrates at a higher rate than AEGE. In contrast to the I-LtrI data, there was only a 2-fold difference between the maximum calculated rates of depletion for I-OnuI AEGE and AGED and, neither enzyme depleted the cognate (ATTC) faster than other C4 substrates.

Given that AEGD confers activity against a greater number of substrates than the wild-type enzyme in two LHE backgrounds, we investigated the C4 preference of two additional coevolving network variants (Figure 4. 12). I-LtrI GEGE (A28G) was chosen because it was previously found more catalytically efficient against the cognate (ATAC) substrate than wild-type AEGE (2). The second enzyme, I-LtrI SEGD (A28S/E184D), was selected because it contains the E184D substitution while demonstrating catalytic activity comparable to wild-type AEGE. Surprisingly, only 4% (11/256) and 2% (5/256) of C4 substrates were enriched (log2 Enrichment >1) by GEGE and SEGD after 15min. All enrichment values for GEGE and SEGD were deemed statistically significant (log2 effect sizes > 2.5) and the cognate ATAC sequence was the preferred C4 substrate for both variants.
The main conclusions from the \textit{in vitro} C4 cleavage preference experiments are:

1. Both product enrichment and substrate depletion strategies can be applied to identify permissible LHE substrates from randomized sequence libraries.

2. Consistent with previous findings (2, 13), the E184D substitution increases activity on all C4 substrates cleaved by the wild-type enzyme in two LHE scaffolds.

3. The E184D mutant is also active against C4 substrates that are not permissible to wild-type enzyme activity. Given that the effect does not extend to all C4 substrates, our data suggests E184D is an up-activity mutant with a broader substrate range than the wild-type I-LtrI and I-OnuI enzymes.

4. I-LtrI and I-OnuI variants prefer C4 substrates with 2 or fewer substitutions from the cognate sequence.

5. Consistent with previous findings for wild-type I-LtrI and I-OnuI (12), E184D variants show a strong preference for C (and a to a lesser extent T) at the +2-nucleotide position and, substrates with G:C dinucleotide steps at the -1/+1-nucleotide positions are not permissible to activity.

6. Increases in C4 substrate activity conferred by the E184D substitution are context-dependent with respect to amino acid identity at the other coevolving positions (i.e. E184D is dependent upon A28, E29 and G183).
Figure 4. 7: Product enrichment and substrate depletion flow-through.

Schematic of the experimental strategy used to simultaneously LHE activity on all 256 possible central four (C4) sequences.
Figure 4.8: Product enrichment results for I-LtrI.

(A) Heatmap illustrating the enrichment values (log2) for all 256 central four (C4) substrates by wild-type I-LtrI (AEGE) and I-LtrI E184D (AEGD). AEGE and AEGD (lower x-axis) data are presented side-by-side to facilitate a direct comparison of substrate enrichment by both enzymes. Sequences were split by the left (-2 and -1) and right (+1 and +2) C4 nucleotides on the left-y and upper-x axes respectively. Note that this presentation is not meant to infer that C4 nucleotides on left and right side are independent. (B) Bar-plots showing the top quartile of C4 sequences that were enriched (log2) by wild-type I-LtrI (cyan) and I-LtrI E184D (red).
Table 4. 1: Nucleotide frequency in the top quartile of central four (C4) substrates for I-LtrI and I-OnuI.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Variant</th>
<th>Top Strand C4 Nucleotide</th>
<th>C4 Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(-2)</td>
<td>(-1)</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>AEGE</td>
<td>44% A</td>
<td>40% T</td>
</tr>
<tr>
<td>I-LtrI</td>
<td>AEGD</td>
<td>45% A</td>
<td>35% T</td>
</tr>
<tr>
<td>I-OnuI</td>
<td>AEGE</td>
<td>38% A</td>
<td>33% T</td>
</tr>
<tr>
<td>I-OnuI</td>
<td>AEGD</td>
<td>48% A</td>
<td>39% T</td>
</tr>
</tbody>
</table>

Figure 4. 9: Comparison of I-LtrI substrate depletion and product enrichment.

Scatterplots comparing wild-type I-LtrI (left) and I-LtrI E184D (right) product enrichment to the corresponding rate of substrate depletion (nM/min) for all 256 central four (C4) substrates. Circles represent the placement of numeric values while the C4 substrate sequence identity is presented above each point.
Figure 4.10: Product enrichment by I-OnuI.

(A) Heatmap illustrating the enrichment values (log2) for all 256 central four (C4) substrates by wild-type I-OnuI (AEGE) and I-OnuI E178D (AEGD). AEGE and AEGD (lower x-axis) data are presented side-by-side to facilitate a direct comparison of substrate enrichment by both enzymes. Sequences were split by the left (-2 and -1) and right (+1 and +2) C4 nucleotides on the left-y and upper-x axes respectively. Note that this presentation is not meant to infer that C4 nucleotides on left and right side are independent. (B) Bar-plots showing the top quartile of C4 sequences that were enriched (log2) by wild-type I-LtrI (cyan) and I-OnuI E178D (red).
Figure 4.11: Comparison of I-OnuI substrate depletion and product enrichment.

Scatterplots comparing wild-type I-Onul (left) and I-Onul E178D (right) product enrichment to the corresponding rate of substrate depletion (nM/min) for all 256 central four (C4) substrates. Circles represent the placement of numeric values while the C4 substrate sequence identity is presented above each point.
Figure 4.12: The effect of E184D is context-dependent.

Heatmaps illustrating the enrichment values (log2) for all 256 central four (C4) substrates by wild-type I-LtrI AEGD, AEGE, GEGER and SEGGER. All data were presented using the same scale, and 15min and 30min enrichment data (lower x-axis) are presented side-by-side to facilitate a direct comparison of substrate enrichment over time. Sequences were split by the left (-2 and -1) and right (+1 and +2) C4 nucleotides on the left y and upper x axes respectively. Note that this presentation is not meant to infer that C4 nucleotides on left and right side are independent.
4.3.3 Increases in $k_{\text{cat}}^*$ Allow I-LtrI AEGD to Overcome Defects in Binding and Cleave AACC with Greater Efficiency than Wild-type AEGE

To help explain differences in activity between I-LtrI AEGE (wild-type) and AEGD (E184D), we determined the kinetic parameters underlying cleavage of the ATAC (cognate) and AACC substrates. Both proteins were purified by affinity chromatography and plasmids containing each I-LtrI C4 sequence were used as substrate for time-course endonuclease assays (Figure 4.13). Initial rates were calculated using the linear range of product accumulation, and single-turnover (pseudo-Michaelis-Menten) kinetics were modeled using six different protein concentrations (Table 4.2).

While the $V_{\text{max}}$ and $k_{\text{cat}}^*$ for AEGE were similar on both substrates, a ~10-fold increase in $K_M^*$ resulted in a ~10-fold reduction in catalytic efficiency ($k_{\text{cat}}^* / K_M^*$) for AEGE against the AACC substrate. AEGD was found to have $V_{\text{max}}$ and $k_{\text{cat}}^*$ values ~4-times greater than AEGE against the cognate (ATAC) substrate but, increases in $k_{\text{cat}}^*$ were offset by a ~3-fold increase in $K_M^*$ and resulted in an efficiency comparable to AEGE. Versus the AACC substrate, the Vmax for AEGD was reduced by 27% compared to ATAC and the $K_M^*$ increased ~3-fold. Although there were subtle differences between the $k_{\text{cat}}^*$ and $V_{\text{max}}$ values for AEGD on both substrates, the $k_{\text{cat}}^*$ for AEGD was ~3-fold higher than AEGE on AACC. Together these kinetic parameters illustrate that AEGD is ~3-times more efficient at cleaving the AACC substrate than wild-type AEGE, and these differences can be attributed to defects in substrate binding by AEGE and a general increase in $k_{\text{cat}}^*$ for AEGD.

Given that the general increase in activity for AEGD has been mentioned in the literature (12), we examined this further by comparing rate constants for the two sequential phosphodiester hydrolysis reactions (Table 4.3). While the $k_1$ constants for wild-type AEGE are similar for both substrates, there was a ~5-fold reduction in $k_2$ for AEGE on AACC compared to ATAC. The $k_1$ for AEGD against ATAC was ~6-fold higher than AEGE while no difference was observed between the $k_2$ values for both enzymes on ATAC. The $k_1$ for AEGD was reduced ~3-fold on AACC compared to ATAC but this value was at least 3-fold greater than the $k_1$ for AEGE on AACC. Finally, the $k_2$ for
AEGD was ~2-fold higher on AACC than ATAC and ~10-fold greater than the $k_2$ for AEGE against AACC.

The main conclusions from the kinetic profiling experiments are:

1. The E184D substitution increases $V_{\text{max}}$ and $k_{\text{cat}}^*$ on the cognate substrate but the mutation also impacts binding (3-fold increase in $K_M^*$) which results in AEGD having a catalytic efficiency comparable to wildtype AEGE on ATAC.

2. Poor activity for wildtype I-LtrI on AACC can be attributed to a ~10-fold increase in $K_M^*$ compared to the cognate (ATAC) substrate.

3. AEGD is twice as efficient at cleaving AACC than AEGE because defects in binding ($K_M^*$) are off-set by increased $k_{\text{cat}}^*$ and $V_{\text{max}}$ values.

4. AEGD increases the first catalytic rate constant ($k_1$) against the cognate ATAC substrate compared to the wild-type AEGE.

5. AEGD increases both catalytic rate constants ($k_1$ and $k_2$) against the non-cognate AACC substrate compared to the wild-type AEGE.
Figure 4. 13: E184D increases cleavage activity compared to wild-type I-LtrI.

0.9% agarose-TBE gel electrophoresis images showing time-course endonuclease assays of I-LtrI AEGE (wild-type) and AEGD (E184D) against either the cognate (ATAC) or non-cognate (AACC) central four (C4) sequence. Over time supercoiled plasmid bands are converted to nicked and fully linear bands and these species were quantitated using densitometry and AlphaImager© software. Note that these images represent one replicate of 500nM protein to 5nM substrate assays.
**Table 4. 2: I-LtrI variant pseudo-Michaelis-Menten kinetics.** I-LtrI AEGE (wild-type) and AEGD (E184D) single-turnover kinetics with cognate (ATAC) and non-cognate (AACC) central four substrates (n=3, values ±SEM).

<table>
<thead>
<tr>
<th>I-LtrI Scaffold</th>
<th>C4 Substrate</th>
<th>$V_{\text{max}}$ (nM / sec)</th>
<th>$k_{\text{cat}}$ (nmol / sec)</th>
<th>$K_m$ (nM)</th>
<th>Efficiency ($k_{\text{cat}} / K_m$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEGE (wt)</td>
<td>ATAC</td>
<td>0.29</td>
<td>5.88 x 10^{-2} ± 2.07 x 10^{-3}</td>
<td>48.2 ± 9.58</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>AACC</td>
<td>0.29</td>
<td>5.76 x 10^{-2} ± 7.66 x 10^{-3}</td>
<td>458 ± 9.58</td>
<td>0.13</td>
</tr>
<tr>
<td>AEGD (E184D)</td>
<td>ATAC</td>
<td>1.06</td>
<td>2.11 x 10^{-1} ± 9.02 x 10^{-3}</td>
<td>163 ± 25.7</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td>AACC</td>
<td>0.77</td>
<td>1.55 x 10^{-1} ± 2.36 x 10^{-2}</td>
<td>561 ± 178</td>
<td>0.28</td>
</tr>
</tbody>
</table>

**Table 4. 3: Rate constants for I-LtrI cleavage.** I-LtrI AEGE (wild-type) and AEGD (E184) rate constants for the sequential nicking reactions of cognate (ATAC) and non-cognate (AACC) central four substrates. Assays were accomplished with 500nM protein to 5nM substrate, and data are presented ± standard deviation (n=3).

<table>
<thead>
<tr>
<th>I-LtrI Variant</th>
<th>C4 Substrate</th>
<th>$k_1$ (nM / sec)</th>
<th>$k_2$ (nM / sec)</th>
<th>$k_1 / k_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEGE (wt)</td>
<td>ATAC</td>
<td>4.63 x 10^{-3} ± 2.07 x 10^{-3}</td>
<td>4.11 x 10^{-2} ± 2.07 x 10^{-3}</td>
<td>11.2 x 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>AACC</td>
<td>3.29 x 10^{-3} ± 7.66 x 10^{-3}</td>
<td>7.47 x 10^{-3} ± 7.66 x 10^{-3}</td>
<td>3.43 x 10^{-4}</td>
</tr>
<tr>
<td>AEGD (E184D)</td>
<td>ATAC</td>
<td>2.79 x 10^{-2} ± 9.02 x 10^{-3}</td>
<td>4.56 x 10^{-2} ± 9.02 x 10^{-3}</td>
<td>5.90 x 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>AACC</td>
<td>1.12 x 10^{-2} ± 2.36 x 10^{-2}</td>
<td>8.05 x 10^{-2} ± 2.36 x 10^{-2}</td>
<td>2.20 x 10^{-5}</td>
</tr>
</tbody>
</table>
4.3.4 Base Pair Opening Predicts I-LtrI Catalytic Activity on Cognate and Non-Cognate C4 Substrates

To help explain differences in catalytic efficiency between I-LtrI coevolving network variants, x-ray crystallography was used to investigate active site architecture and substrate structure in various protein-DNA complexes. To facilitate these aims, I-LtrI variants and DNA substrates were co-crystallized in the presence of calcium to prevent catalysis and attain pre-cleavage structures. I-LtrI AEGE (wild-type), GEGE (A28G), AEAE (G183A) and AEGD (E184D) were solved in complex with the cognate (ATAC) substrate to resolutions of 2.0Å, 2.9Å, 2.92Å and 2.5Å respectively (Dataset S3.1). In addition, I-LtrI AEGE and AEGD were also solved in complex with the AACC substrate to resolutions of 2.28Å and 2.5Å (Dataset S3.1).

With respect to active site architecture in the four ATAC duplex structures, no significant differences were found in positioning of the catalytic residues (E29 and E184) and, all protein-DNA duplexes contained 3 bound metal ions except for GEGE which was missing a central cation (Figure 4. 14). In order to tolerate the coevolving network mutations, it is likely that subtle changes in LAGLIDADG-helix packing, and/or domain interactions, must occur to maintain the precise orientation of active site residues required for catalysis (14-16). To explore this idea, the N-terminal domains from all four complexes were aligned using least squares fitting to emphasize differences in the relative positioning of the C-terminal domains (Figure 4. 15). Although no differences were found between AEGE and AEGD, secondary structure elements in the C-terminal domain of GEGE and AEAE were shifted by up to 4.42Å compared to the wild-type AEGE structure.

To quantitate differences in ATAC substrate structure, all four protein-DNA complexes were analyzed using 3DNA software (Figure 4. 16). The minor groove widths at the centre of AEGE (5Å) and AEGD (5Å) were narrowed to a greater extent than the widths measured for GEGE (5.7Å) and AEAE (6.9Å). A 4° base pair shear and 5.62Å y-displacement were found in the centre of the GEGE-ATAC duplex while these parameters were not significantly different in the other three complexes. The only other parameter that varied between all four complexes was base pair opening in the C4 region.
between the two scissile phosphates. AEGE (6.11°), AEGD (6.31°) and GEGE (2.4°) open the cognate bases around the first scissile phosphate (-2 position), while AEGD (2.67°) and GEGE (7.48°) also open the substrate around the second scissile phosphate (+2 position). In contrast, C4 base pairs adopted a closed confirmation (-6.68°) when bound by the poorly active AEAE variant.

When the structures of AEGE and AEGD were solved in complex with the non-cognate substrate AACC, no differences were found with respect to secondary structure elements or positioning of catalytic residues and 3 metal ions were bound in each duplex (Figure 4. 17). 3DNA analysis of the AACC substrate in both duplexes showed no changes in minor groove width and base pair shear, but disparities were found in C4 base pair y-displacement and base pair opening (Figure 4. 18). Bases around the first scissile phosphate (-2-position) were open in both AEGE (3.31°) and AEGD (5.21°), while only AEGD (2.05°) conferred an open structure to bases around the second scissile phosphate (+2-position) of AACC.

The main findings from analyzing I-LtrI-C4 substrate duplex structures are:

1. The GEGE-ATAC structure was the only duplex that contained 2 bound divalent metal ions within the active site.

2. Mutation of the non-catalytic coevolving residues A28 and G183 (AEAE and GEGE variants) influences the interaction of domains at the LAGLIDADG interface.

3. The minor groove of I-LtrI ATAC substrates were narrowed by wild-type AEGE and AEGD to a greater extent than the GEGE and AEAE variants.

4. The opening of C4 base pairs is predictive of I-LtrI catalytic activity against both the ATAC and AACC substrates.
Figure 4. 14: Structural models for I-LtrI variants in complex with the cognate ATAC sequence.

(A) Close-up of the coevolving network residue side chains (28, 29, 183, and 184) from each I-LtrI variant within the active site of a duplex structure. (B) Bird’s eye view looking at the minor groove of ATAC and divalent cation positioning. Note that these models are generated from a global least squares alignment, and the following colour scheme was used to identify each variant: AEGE is green, AEGD is blue, GEGE is magenta, and AEAE is red.
Figure 4.15: Least squares alignment of I-LtrI variant N-terminal domains.

(A) Frontal (top) and bird’s eye (bottom) view of I-LtrI variant secondary structures with the N-terminal domain on the left. (B) Side view of the structural alignment looking at deviations in the C-terminal domain secondary structures. Note that the ATAC substrate was removed for clarity, and the following colour scheme was used to identify each variant: AEGE is green, AEGD is blue, GEGE is magenta, and AEAE is red.
Figure 4. 16: 3DNA outputs for the cognate ATAC substrate in complex with each I-LtrI variant.

Line graphs illustrating the following DNA parameter measurements: (top) Minor groove width (Å), (upper-middle) Base pair opening angle (°), (lower-middle) Base pair shear (°), and y-displacement (°). Note the following symbol scheme was used to identify each variant: black circles for AEGE, open triangles for AEGD, black diamonds for GEGE, open squares for AEAE, and the central four nucleotide region is shaded grey.
Figure 4.17: Structural models for I-LtrI AEGE and AEGD in complex with the non-cognate AACC sequence.

(A) Close-up of the coevolving network residue side chains (28, 29, 183, and 184) from each variant within the active site of a duplex structure. (B) Bird’s eye view looking at the minor groove of AACC and divalent cation positioning. Note that these models were generated using a global least squares alignment, and the following colour scheme was used to identify each variant: AEGE is green and AEGD is blue.
Figure 4.18: 3DNA analysis of the AACC substrates in complex with I-LtrI AEGE (wild-type) and AEGD.

Line graphs illustrating the following DNA parameter measurements for the non-cognate AACC substrate in complex with I-LtrI AEGE and AEGD: (top) Base pair opening angle (°), and (lower) y-displacement (°). Note the following symbol scheme was used to identify each variant: black circles for AEGE, open triangles for AEGD, and the central four nucleotide region is shaded grey.
4.3.5 Direct Contacts to C4 Nucleobases are Present in LHE Pre-Cleavage Structures

The protein-DNA interface of our wild-type I-LtrI pre-cleavage complex was then compared to a previously reported (12) wild-type I-LtrI post-cleavage complex (PDBID: 3R7P) using the EMBL-EBI tool PDBePISA. The two structures have 40 water-mediated, direct, or backbone contacts in common while the 37RP structure contains an additional 12 contacts outside of the C4 region (data not shown). Unexpectedly, two direct base contacts were identified between I-LtrI and C4 nucleotides in our pre-cleavage complex (Figure 4.19). These energetically favourable minor groove contacts ($\Delta G=-1.64$) are made between R311 and the bottom strand T and G at the +1/+2-nucleotide positions, and have a bond length of 3.63 Å (Figure 4.19B). These contacts are not found in the post-cleavage I-LtrI structure (3R7P) because the R311 side chain density presented a different rotamer conformation (Figure 4.19A). In support of this observation, both I-LtrI structures were solved in the same space group and our pre-cleavage complex was solved to a higher resolution (2.0 Å vs 2.7 Å). However, further validation of these direct contacts is required because the two structures were crystallized using different metal ions, temperatures, and buffer conditions.

While backbone interactions with the C4 region have been reported in the literature, no direct contacts to C4 nucleotides have been discussed in other LHE structures (17-21). Given the unexpected nature of our finding, I decided to analyze every available pre-cleavage LHE structure using PDBePISA (Table 4.4). To our surprise, direct contacts to C4 nucleotides were predicted in six additional LHEs (I-OnuI, I-PanMI, I-AabMI, I-GpeMI, I-GzeII and I-LtrWI) while no mention of these interactions were discussed in the corresponding publications. Interestingly, the nature of these interactions varies in terms of contacting amino acid identity, mechanism of base readout, and whether the contacts occur in the major or minor groove of the C4 region.
Figure 4.19: I-LtrI residue R311 makes a direct contact to C4 bases.

(A) Global alignment of the I-LtrI post-cleavage (PDB: 37RP) and pre-cleavage structures in complex with the cognate ATAC substrate. The side chain rotamer conformation of the R311 (sticks) was different in the pre-cleavage (purple) and post-cleavage (cyan) structures. (B) The direct contact between R311 and nucleotides in the I-LtrI target site. Carbon (green), oxygen (red), nitrogen (blue), phosphorus (orange) atoms are identified to illustrate the hydrogen bonding potential associated with this protein-DNA interaction.
Table 4.4: Predicted direct contacts to central four base pairs in pre-cleavage LHE structures.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDBID</th>
<th>Amino Acid</th>
<th>Base Contact</th>
<th>Bond length</th>
<th>ΔG</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-LtrI</td>
<td>37RP</td>
<td>R311</td>
<td>1T</td>
<td>3.64</td>
<td>-1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R311</td>
<td>2G</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>I-Onul</td>
<td>3QQY</td>
<td>K229</td>
<td>1T</td>
<td>3.14</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K229</td>
<td>2G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-PanMI</td>
<td>5ESP</td>
<td>R229</td>
<td>2G</td>
<td>2.7</td>
<td>-1.93</td>
</tr>
<tr>
<td>I-AabMI</td>
<td>4YIT</td>
<td>R220</td>
<td>T2</td>
<td>3.4</td>
<td>-1.76</td>
</tr>
<tr>
<td>I-GpeMI</td>
<td>4YHX</td>
<td>H210</td>
<td>T1</td>
<td>3.61</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K234</td>
<td>T1</td>
<td>3.05</td>
<td>-0.42</td>
</tr>
<tr>
<td>I-GzeII</td>
<td>4Z1X</td>
<td>D44</td>
<td>T2</td>
<td>3.7</td>
<td>-0.03</td>
</tr>
<tr>
<td>I-LtrWI</td>
<td>4LQ0</td>
<td>R231</td>
<td>T2</td>
<td>3.8</td>
<td>-1.42</td>
</tr>
</tbody>
</table>
4.4 Discussion

In this study, we systematically investigated LHE central four (C4) cleavage preference, and attempted to expand LHE targetability by identifying variants with increased activity on non-cognate C4 sequences. Specificity in the C4 region is thought to be governed by an indirect readout mechanism and LHEs are generally intolerant to C4 nucleotide substitutions because they can reduce binding affinity, or eliminate cleavage activity by disrupting the coordination of catalytically essential divalent cations (17, 18, 21). Here, mutation and functional selection of four coevolving active site residues has identified a single variant, E184D (AEGD), that demonstrates higher levels of activity on non-cognate C4 sequences than the wild-type enzyme.

To fully understand how the E184D mutation affects LHE targetability, the C4 nucleotide preference of wild-type I-LtrI and I-OnuI was compared to their E184D variants by thoroughly profiling cleavage activity against all permutations of the C4 sequence. Towards this aim, we developed two high-throughput and unbiased in vitro methods to simultaneously profile cleavage activity against the 256 possible C4 substrates. While both strategies used the same randomized C4 target plasmid library input, our product enrichment assay identified permissible substrates by recognizing target sequences that were enriched in linearized product pools after exposure to enzyme. In contrast, substrate depletion assays investigated the substrate pool and measured loss of permissible sequences from the randomized input library. Both techniques successfully identified all permissible LHE substrates, and are advantageous over existing methods because each substrate is tested in unbiased competition (22-24). This allows for direct comparison of each substrate, and can identify codependent relationships within a DNA target site because nucleotide positions are not probed independently. Whereas our assays were designed to investigate nucleotide preference in LHE substrates, it is important to note that similar methodologies could be employed to study most DNA-binding proteins and endonucleases, such as Cas9.

Our C4 nucleotide preference analysis revealed that E184D (AEGD) variants expand I-LtrI and I-OnuI targetability by increasing activity on sequences that are poor substrates for wild-type enzymes. Although the E184D mutation has been identified in other
directed evolution experiments (12), it was thought to cause a general increase in catalytic rate on all LHE substrates. Significantly, we have shown that increases in cleavage activity conferred by E184D do not extend to all 256 possible C4 nucleotide combinations, and the resulting effect of this mutation is different between LHE scaffolds. Consistent with previous findings (17, 18), sequences with G-C dinucleotide steps at the centre of the LHE target site are non-cleavable, the best C4 substrates contain two or fewer nucleotide substitutions from the cognate, and I-Ltrl and I-Onul show a strong preference for C at the +2 position (12). In I-Ltrl specifically, we found that E184D is generally more active than wild-type even though both enzymes demonstrate in vitro activity against a similar subset of C4 sequences. Nevertheless, the increases in catalytic rate allow I-Ltrl E184D (AEGD) to over-come defects in binding, and maintain biologically relevant activity on C4 sequences that are poor substrates for the wild-type enzyme. Surprisingly, the impact of E184D was even more significant in the I-Onul background. Once again, the variant was more active on all sequences permissible to the wild-type but I-Onul E184D also demonstrated activity on a much larger subset of the 256 C4 substrates.

The sequence preference we determined for wild-type I-Onul was consistent with previous studies in which yeast-display methods were applied to probe the C4 nucleotide specificity of I-Onul and I-PanMI (18). Lambert et al. (2016), found that I-Onul is active on a smaller subset of C4 substrates than I-PanMI, and these findings were attributed to LHE scaffolds showing unique levels of sequence promiscuity. Our I-Ltrl and I-Onul data confirms that there are innate differences in C4 nucleotide specificity between each LHE, but we also find that C4 substrate permissibility can be expanded with the addition of E184D. Importantly, we find that E184D increases the number of C4 sequence targetable by I-Onul comparable to what has been described for the promiscuous I-PanMI (18). Interestingly, wild-type I-PanMI naturally contains an Asp (D) at position 184 (25) and we believe this residue identity is likely responsible for its increased C4 activity. Unlike mutations that degrade LHE functionality, our findings suggest that D184 enzymes (such as I-PanMI) would be more successful at catalyzing genomic DSBs and invading host genomes with variable C4 sequences than E184 LHEs.
While the presence of E184D expands LHE sequence targetability in I-LtrI and I-OnuI, it is important to note that the positive effects of this mutation are context-dependent. While half of the LHE variants tested in this study contained E184D, only AEGD (A28, E29, G183 and E184D) demonstrated increased activity on non-cognate sequences. This finding supports the coevolution of LHE positions 28, 29, 183 and 184, and illustrates why residue identity in this coevolving network needs to be considered during re-engineering applications. For example, when six different I-OnuI family LHEs were used to engineer domain-fusion chimeras, network combinations were changed to AEAE and GEGE in over 50% of constructs (Table 3.1). While GEGE is active against the cognate ATAC sequence, AEAE is poorly active on all substrates and our data shows that introducing E184D into these backgrounds would not increase cleavage activity.

Changes to the LHE coevolving network at the LAGLIDADG interface do not necessarily influence substrate preference, but they can impact cleavage activity by altering substrate structure in C4 nucleotide region. Following binding, LHEs induce a substantial bend in their DNA substrates (26, 27), and the resulting torsional stress has been shown to unstack base pairs in C4 region of the target sequence. Given that the energetic cost associated with these changes is sequence-dependent (28), recognizing the unique local substrate structure within the active site is an important determinant of preferred sequence indirect readout (17, 18, 21, 29, 30). This mechanism is exemplified in previous studies of I-SmaMI where substitutions to the C4 region altered substrate structure, and excluded catalytically essential divalent cations from the active site (18).

Unexpectedly, crystallographic results show limited differences in cation binding between I-LtrI coevolving network variants on both the cognate ATAC and non-cognate -AACC sequences. This was unexpected because previous studies have demonstrated that the impermissibility of some non-cognate C4 sequences is caused by altered metal binding (17, 18). It is important to note that we cannot rule out changes to I-LtrI divalent cation binding when in complex with C4 substrates that are non-permissible to cleavage. The exception in our findings is the GEGE variant where only two bound metals were found within the active site. Given the degree of base pair shear induced by GEGE, it is possible that this dramatic deformation of the substrate prevents metal binding. While
this does not affect activity on the cognate sequence, this finding could help explain why GEGE demonstrates poor \textit{in vivo} activity on non-cognate sequences where the availability of metals is limited compared to \textit{in vitro} analysis.

Interestingly, side chain mutations or C4 base substitutions that increased base pair opening around the scissile phosphates improved catalytic rate. These results suggest that base pair opening in the C4 nucleotide region is an excellent predictor of substrate permissibility because open complexes lead to the unstacking of bases. This observation for preferred LHE substrates is similar to an investigation of indirect readout by the endonuclease ColE7 (30). The DNA backbone was distorted to a greater extent in preferred ColE7-substrate complexes because non-preferred complexes contain more hydrogen-bonded base pairs that reduce base pair opening. Similar to narrowing of the minor groove by LHEs, base pair opening acts to bring scissile phosphates into closer proximity (31), and presumably increases the susceptibility of electrophilic scissile phosphates to nucleophilic attack.

Surprisingly, our structural studies have also identified direct protein contacts to base pairs at the +1 and +2 positions of the C4 region. Admittedly, these contacts are inconsistent with previous findings for LHE enzymes (18, 21, 32) but they provide a simple explanation for the strong C preference at the +2 position of I-LtrI and I-OnuI target sites. These observations are strengthened by the identification of direct C4 contacts in five other pre-cleavage LHE complexes, and recognition that the same major groove interaction predicted for I-OnuI K229 has been described previously (33) in the p50 subunit of the human transcription factor NF-kappa B (PDBID: 1SVC). However, direct C4 nucleotide contacts are not observed in all LHEs, and no interactions were identified with the -2 and -1 positions of the C4 region. LHE family members are not functionally homogeneous because some members have evolved specialized roles within their host (ex. I-SceI (34)), while others demonstrate maturase activity in addition to their hydrolysis of DNA (ex. I-AniI (35)). Therefore, indirect substrate readout is still an important mechanism controlling LHE specificity in the C4 region, but the possibility of direct readout at the +1 and +2 positions of the LHE target site cannot be ruled out and requires further study.
The studies presented in this article describe two new methods to investigate the nucleotide specificity of DNA-binding proteins, and demonstrate how specific alterations to the LHE active site can increase activity on non-cognate C4 sequences. In addition, base pair opening in the C4 region was found to be an important factor that determines substrate permissibility even when nucleotide substitutions reduce binding affinity. Together, these findings help to expand the targetability of LHEs and increase their applicability as genome-editing reagents in the field of biotechnology.
4.5 References


Chapter 5

5 Discussion

5.1 Covariation Analyses Identify Biologically Important Amino Acid Networks Within Protein Families

Mutual information (MI) based covariation analyses were originally applied to protein families to facilitate ab initio predictions of protein folding (1). This concept was based on the idea that residue identity between two contacting positions is restricted (or covaries) to maintain local protein structure and stability (2, 3). Statistical corrections and refinements have increased the accuracy of covariation methods, and have dramatically improved residue contact prediction from multiple sequence alignment (MSAs) (4-7). Given that this method identifies two or more positions that do not randomly assort within a MSA, it was hypothesized that high levels of covariation could be driven by the coevolution of residues to fulfil important structural, or functional roles within protein families (5, 7). Today, numerous coevolving networks have been identified, but it has remained difficult to experimentally validate the functional impact of covarying positions, especially within enzyme active sites (7).

Prior to the start of this thesis, a covariation analysis was applied to a structure-guided MSA of the LAGLIDADG homing endonuclease (LHE) family using the Zpx method (5, 6). This was performed to identify amino acid positions that represent statistical outliers in the phylogeny and entropy corrected MI analysis.

Similar to previous findings in model protein families (7), the most significant covariation scores within the LHE family consisted of a 4-residue quartet within the enzyme active site. The positions included two catalytic metal-binding residues (E29 and E184) adjacent to non-catalytic positions (A28 and G183) that contact across the LAGLIDADG domain interface (Figure 2.1. c). I-LtrI numbering will be used for simplicity, and network variants will be referred to by their residue identity such that wild-type I-LtrI is AEGE. Two major clades were identified when residue identity at the covarying positions was mapped onto the LHE family tree (Figure 2.1. a). One clade
contained LHEs with an Asp at position 29 in combination with G28 and A184 (i.e. GDA[D/E]), while the other clade was comprised of Glu29 in conjunction with A28 and G183 (i.e. AEG[D/E]). In contrast to the other positions, identity at the second catalytic residue (184) showed more variability, and had no clear clade preference.

Mutation and functional selection of the quartet recapitulated the LHE phylogenetic distribution (Figure 2.1. d and Figure 2.2. b), and demonstrated that amino acid combinations within the coevolving network are generally permissible to activity if they were found naturally within the LHE family. That said, some combinations that were absent from our MSA were found to be permissible to activity. Although residue permissibility could be scaffold specific, this is more likely caused by ascertainment bias, or our inability to sufficiently sample the sequence space of functional enzymes.

Most striking from the mutagenesis study was that D29E substitutions were not tolerated in the I-HjeMI background, and E29D substitutions were not tolerated in I-LtrI (Figure 2.2. b). These findings suggest that the fitness landscape of LHEs is specifically constrained by the residue identity at position 29, and supports the existence of two evolutionary distinct LHE clades.

Interestingly, substitutions within the network were found to modulate LHE catalytic efficiency against the cognate substrate over a 100-fold range (Figure 2.4. b), and the cleavage activity of poorly active network variants, such as SEGE, could be rescued by secondary mutations within the network (e.g. SEGd) (Figure 2.5). Importantly, the observed differences in catalytic efficiency were caused by alterations to binding affinity and/or catalytic rate rather than disrupting thermostability of the protein scaffold (Figure 2.4. a).

Together, these data strongly supported my hypothesis that the covariation statistic Zpx could identify coevolving residues within the LHE family. The thorough biochemical validation I present in Chapter 2 clearly demonstrated that LHE positions A28, E29, G183, and E184 are non-independent, and interact to fulfil a functional role during enzyme catalysis. Significantly, this study demonstrates that Zpx can identify biologically
relevant amino acid networks from primary sequence information, and provides a framework for identifying coevolving residues in other protein families.

Another interesting finding from this study was that many of the highest covariation scores were from residue combinations that clustered at the LAGLIDADG interface (Figure 2.1. b). Although this is not the first analysis to identify contacting interfacial residues (7), these data confirm that the statistic $Zpx$ can accurately predict the position of a protein-protein interface from sequence data. While the LADLIDADG interface is an intra-molecular structure, I believe that these analyses could be expanded to identify inter-molecular protein-protein contacts between protein families. This could be accomplished by creating independent MSAs (one for each protein of interest), and then concatenating the two MSA into one alignment prior to the covariation analysis. These de novo predictions of protein-protein interfaces would undoubtedly require significant time and computational power. However, this type of analysis will become more feasible as the accuracy of automated MSA generation increases, and could generate valuable information from the plethora of sequence information that is available for most protein families.

5.2 Applying Covariation Analysis to Investigate the Genetic Basis of Human Disease

The work from Chapter 2 of this thesis provides one of the first experimentally validated examples of amino acid coevolution within an enzyme active site. These findings illustrate that covariation analyses can identify important interactions between amino acids from sequence information, and this can be accomplished without prior knowledge of protein structure. My results provide valuable information regarding LHE biology, and have implications for re-engineering LHE scaffolds toward sequences of interest in genome-editing applications. However, these studies also provide a framework that can be used to expand covariation analyses to other protein families, and have the potential to identify the genetic basis of human abnormalities and disease that cannot be recognized by traditional methods.
We have shown that while coevolving positions display sequence diversity, certain residue combinations can severely impact protein structure and function (Figure 2.3. a). Therefore, disease-associated human mutations could occur at coevolving positions within the human genome, and substitution to non-permissive residue combinations could impair functionality, reduce fitness, and potentially lead to disease. I envision comparing the results from a covariation analysis to databases of human disease mutations, such as the Cancer Genome Atlas (8), and identifying overlapping residue positions. Our findings demonstrate that the fitness landscape of proteins can be constrained by networks of coevolving amino acids, and disruptions to these significant interactions could be detrimental to cells, and ultimately human health.

Unfortunately, the application of covariation analyses to study the human proteome is currently limited by a lack of sequence diversity. In the covariation analysis presented in Chapter 2, LHE sequences with 95% sequence identity were removed from the input MSA because highly similar sequences suppress the scores MI-based analyses (9, 10). For this reason, there is insufficient genomic diversity within our species to facilitate the identification of coevolving residues in most protein families. These limitations could be mitigated by endeavours such as the United Kingdom’s 100 000 Genomes Project (11), but similar to Genome-wide Association Studies (GWAS), it will be difficult to differentiate true disease-causing mutations from natural polymorphisms (12). Fortunately, advances in sequencing technology, and the reduction in cost for whole-genome sequencing, is expediting the Genome 10k (13) project that aims to sequence genomes from ~66 000 vertebrate species. This project will not only increase our understanding of evolutionary history, but also has the potential to provide the sequence diversity required to identify biologically significant coevolving networks that impact human health.

5.3 Coevolving Network Identity Must Be Taken into Consideration When Re-Engineering LADLIDADG Homing Endonuclease (LHE) Scaffolds

Domain-fusion chimeras (DFCs) represent one of the most promising avenues for engineering because they can exploit the natural reservoir of DNA recognition specificity
and increase the targeting capacity of LHEs (14). Baxter et al. (2012) provided a detailed framework for engineering DFCs using I-OnuI subfamily members, but despite their efforts, many constructs were dysfunctional in expression, binding, and/or cleavage activity. Some of these defects were mitigated by engineering a common set of interfacial residues at the LAGLIDADG interface, but these efforts had the unforeseen consequence of reducing cleavage activity of wild-type LHEs (14).

To help explain these findings, I examined a previously engineered DFC and assessed residue identity at the 4 LHE coevolving positions from Chapter 2 (Table 3.1). To my surprise, coevolving networks were disrupted in 50% of I-OnuI family DFCs by creating residue combinations that were under represented, or not identified, within the LHE family (Figure 2.1. d). Given that amino acid substitutions at the coevolving positions can modulate catalytic efficiency over a 100-fold range (Figure 2.4. b), it is my belief that inadvertent disruption of these critical interactions is in part responsible for reducing the functionality of many DFCs.

The overwhelming majority of I-OnuI subfamily enzymes contain an A_G (e.g. I-LtrI and I-OnuI) or G_A (e.g. I-PanMI and I-SscMI) residue combination at positions 28 and 183 (Figure 5. 1, Figure 2.1. a). When DFCs are engineered using domains from only A_G (e.g. OnuLtr) or G_A (e.g. PanSsc) LHEs, network combinations are not affected and disruption of the coevolving positions is not causative of enzyme dysfunction. However, when domains from A_G LHEs are fused with domains from G_A LHEs (e.g. OnuSsc or PanLtr) then coevolving network combinations can be altered to A_A or G_G (Figure 5. 1, lower).

Interestingly, while A_G and G_A combinations are the most abundant pairings at positions 28 and 183 within the LHE family (Figure 2.1.1), no scaffold in our MSA contained an A_A pairing (Figure 2.1. d). Genetic selection and kinetic data showed that A_A (specifically AEAE) combinations in I-LtrI and I-OnuI dramatically reduce catalytic efficiency against the cognate target sites (Figure 2.3. a, Figure 2.4. b). Structural analysis of I-LtrI AEAE demonstrated a reduced ability to narrow the substrate minor groove
compared to wild-type I-LtrI, and the central four (C4) cognate base pairs adopted a closed conformation that is unfavorable to catalysis (Figure 4.14, Figure 4.16).

Interestingly, substitution to a G_G combination in I-LtrI (specifically GEGE) increased catalytic efficiency against the cognate target site by lowering the binding constant, $K_M$ (Figure 2.4. b). While this caused a moderate increase in survival under genetic selection (Figure 2.4. c), G_G pairings at positions 28 and 183 are poorly represented in the LHE family (Figure 2.1.4). Structural analysis of I-LtrI GEGE showed that the substitution impacted LHE helix packing and domain positioning (Figure 4.15), and caused significant deformation of the cognate substrate (Figure 4.16). This had the interesting effect of eliminating one of the divalent metals that was found within the active site of wild-type I-LtrI duplexes (Figure 4.14). These changes were beneficial against the cognate C4 target site, but I-LtrI and I-OnuI GEGE variants demonstrated poor activity against non-cognate C4 sequences in all genetic selection screens (Figure 4.3, Figure 4.4).

Together these findings demonstrate that the A_A and G_G combinations are likely to negatively impact DFC activity by affecting divalent metal coordination, altering substrate structure within the active site, restricting C4 nucleotide preference, and/or reducing catalytic efficiency. Importantly, these profound effects occurred with relatively simple changes to the LHE domain interface at the base of the LAGLIDADG helices. This implies that significant modification of interfacial residues will undoubtedly affect LHE functionality, and attempts should be made to limit alterations of this critical interface.

An ideal scenario would be the creation of DFCs using domains from any member of the LHE family, but successful DFC engineering has been limited to the I-OnuI subfamily (14). While these represent some of the most well studied LHE scaffolds, I believe the focus on this subfamily is caused by the incompatibility of domains between the two major LHE clades (Figure 2.1. a, Figure 2.2. b). The functional limitations imposed by residue 29 (D29 or E29) are likely to differentially constrain the fitness landscape of their respective scaffold. This would be expected to maintain the precise architecture that is
required by the carboxylate to fulfil its essential role in catalysis (15). Over time, these pressures could have contributed to the sequence divergence necessary for clade separation (Figure 2.1. a), and reduced domain compatibility because their distinct evolutionary history has imposed disparate structural limitations. Therefore, more work is required to understand the biochemical basis of the D/E divergence at position 29, and how this constraint has affected the structural and functional evolution of LHEs.

![Diagram of Ala-Gly and Gly-Ala domains](image)

**Figure 5. 1: The fusion of LHE domains may create unfavorable network combinations.**

The four residue combinations that are possible at positions 28 and 183 when domains from A_G and G_A LAGLIDADG homing endonucleases (LHEs) are used for the engineering of domain-fusion chimeras (DFCs). *Upper left:* A_G LHE such as I-LtrI, or DFCs including OnuLtr. *Upper right:* G_A LHE such as I-PanMI, or DFCs such as PanSsc. *Bottom left:* A_A DFCs such as OnuSsc. *Bottom right:* G_G DFCs such as PanLtr.
5.4 E184D Substitutions Increase LADLIDADG Homing Endonuclease (LHE) Cleavage Activity on Non-Cognate Central Four (C4) Substrates

As described in Chapter 4 of this thesis, E184D mutations confer I-LtrI and I-OnuI with biologically relevant cleavage activity on a greater number of non-cognate C4 substrates than wild-type scaffolds (Figure 4.3, Figure 4.4). Importantly, the effect of this mutation is context-dependent with respect to the other coevolving network residues, and must be combined with A28, E29, and G183 in I-LtrI and I-OnuI (i.e. only works in AEGD variants) (Figure 4.12). Further study is required to determine if E184D can also expand the cleavage activity of D29 LHE scaffolds, but the phylogenetic distribution of network combinations suggest that the substitution could work in the context of G28 and A183 (i.e. GDAD) (Figure 2.1. a, Figure 2.1. d).

The significant consequence of E184D for LHE re-engineering is that the substitution increases catalytic efficiency against non-cognate C4 sequences (Figure 4.13), and expands the targeting capacity of certain scaffolds for genome-editing applications (specifically I-OnuI). What is less clear is the biological role for this variant, and an explanation for the broad distribution throughout every clade of the LHE family (Figure 2.1. a). Data presented in this thesis demonstrate that the effect of E184D is not uniform because both scaffolds that were tested had unique phenotypes. In I-LtrI, the substitution increases cleavage activity against all sequences cleaved by the wild-type (Figure 4.8), while its presence in I-OnuI not only increases cleavage activity, but also facilitates cleavage of C4 sequences that are poor substrates for the wild-type enzyme (Figure 4.10).

In the right context, it is possible that the presence of D184 could increase the success rate of natural LHE homing events and the horizontal transfer of LHE genes. This would be especially beneficial to LHE fitness if the new host genome displays sequence diversity (i.e. a non-cognate sequence) at the homing site. However, the increased activity of E184D enzymes could also prove toxic to host cells (16), and this could explain why the mutation is not maintained in all LHE backgrounds (Figure 2.1. a). Any toxicity to host cells would provide selection pressure in favour of down-regulating LHE genes, and/or the accumulation of loss function mutations within an LHE open reading frame.
These forces would drive the degradation of enzyme functionality, and could contribute to the age-dependent loss of LHE catalytic residues that has been observed in organisms such as Scleractinia and Corallimorpharia (18).

From a mechanistic point of view, the E184D substitution is interesting because Asp residues are the preferred metal coordinating carboxylate in many two-Mg$^{2+}$ mechanisms of catalysis (19). In polymerase enzymes, the preference for Asp has been attributed to the influence of conformational isomerism in positioning active site residues (20). Glu side chains have the potential to adopt a greater number of stable rotamers than Asp, and therefore have the potential limit the efficiency of catalysis by increasing the plasticity of active site architecture. This thought is supported by a broad analysis of 178 non-homologous enzyme active sites that aimed to determine the frequency and function of catalytic residues (21). While there was similar involvement of Asp and Glu in acid/base interactions, Asp was the preferred carboxylate to activate waters, cofactors, or other residue side chains (21).

Although their catalogue of catalytic residue combinations is limited by ascertainment bias (21), the relative increase in rigidity associated with Asp is likely to position side chains more consistently to fulfil electrochemical activation functions (19). With respect to LHEs, residue position 184 is involved with coordinating the shared metal that promotes phosphoryl transfer (15). Therefore, the Asp of E184D variants could be more efficient at positioning the metal to draw charge from scissile phosphates, and/or facilitating protonation of the 3’ hydroxylate leaving group during cleavage events.

5.5 Nucleotide Identity in the Central Four (C4) Constrains LADLIDADG Homing Endonuclease (LHE) Cleavage Activity, and Base Pair Opening is Favorable to Catalysis

The cleavage preference experiments and genetic selection screens from Chapter 4 demonstrate that C4 sequences with high G-C content are poor substrates for I-LtrI and I-OnuI. Specifically, C4 targets with dinucleotide G-C steps at the -1 and +1 positions (i.e. NCCN, NCGN, NGCN, and NGGN) are not permissible to activity (Figure 4.8 and
Figure 4.10), even in the context of E184D (note that C4 sequences are underlined to distinguish from protein variants). This finding is consistent with previous studies (22, 23), and supports nucleotide preference in the C4 region being constrained by the thermodynamics of base pair interactions (23, 24), and their influence on substrate structure and divalent metal-binding within the LHE active site (22, 23).

While some LHEs demonstrate catalytic activity against G-C rich C4 substrates (22, 25), the evolution of this ability must have been driven by unique fitness landscapes, and differences between the nucleotide make-up of host genomes. Given that the free energy change required to deform and unstack G-C rich DNA can be as high as 3 kcal/mol (24, 26), these enzymes must impose unique forces on their substrate to induce the backbone deformation required for catalysis by most LHEs (15, 27). This is the case for I-PanMI where G-C base pairs are tolerated because protein-induced DNA bending causes distinctive base pair sliding in the C4 region (22). This is assumed to lower the energetic cost of unstacking and deforming G-C rich DNA (22) in manner that is not achievable by other LHEs, such as I-LtrI and I-OnuI.

An interesting observation from the structural analysis in Chapter 4 was the relationship between substrate permissibility and the extent of base pair opening in the C4 region of I-LtrI substrates (Figure 4.16, Figure 4.18). I-LtrI AEGE (wild-type), GEGE, and AEGD display significant cleavage activity against the cognate (ATAC) substrate, and were found to increase the opening angle of base pairs in the C4 region. In contrast, the I-LtrI variant AEAE demonstrates poor cleavage activity against ATAC, and actually decreased the opening angle of base pairs in the C4 region (Figure 4.16). Similarly, the I-LtrI variant AEGD induced greater base pair opening in the non-cognate C4 substrate AACC than the wild-type enzyme (Figure 4.18), and it was also measured to be ~3 times as efficient at cleaving AACC following kinetic analysis (Figure 4.13).

Previous studies have demonstrated that y-displacement and base-pair roll promote the catalysis of LHE substrates (22, 28), but there has been limited discussion of a function for base pair opening. Positive opening increases the width of the major groove, and consequentially leads to narrowing of the minor groove (29). Local narrowing of the
minor groove is a requirement for LHE cleavage activity because it is thought to bring scissile phosphates closer to the catalytic residues, and limit the extent of conformational change required between the sequential nicking reactions (30).

Studies on the endonuclease domain of ColE7 have suggested that base pair opening results from a sequence having sufficient DNA backbone deformation to promote catalysis (31). This conclusion was based on the observation that backbone atoms from preferred ColE7 substrates were deformed to a greater extent than non-preferred sequences following binding, and this deformation increased the opening angle of base pairs within the active site (31). I believe that preferred LHE C4 substrates also undergo protein-induced DNA bending and base pair opening in the LHE active site. Larger opening angles likely contribute to narrowing of the minor groove, and presumably increase access to the electrophilic scissile phosphates and increases their susceptibility to nucleophilic attack. Given that each LHE scaffold imposes unique binding and bending forces on their substrate (24, 28, 32), in silico predictions of base pair opening in C4 substrates are not currently feasible. However, these findings help explain the observation that LHEs prefer A-T rich C4 substrates (33) and pyrimidine-purine steps because these sequences are generally amenable to protein-induced DNA bending and deformation (28, 34).

5.6 Some LADLIDADG Homing Endonuclease (LHE) Family Members Make Direct Contacts to Central Four (C4) Base Pairs

Indirect, or shape-based, readout of substrates is generally accepted to be the sole determinant of nucleotide specificity in the C4 region of LHE target sites (22, 35, 36). This mechanism is influenced by protein-induced DNA bending, and the thermodynamics of base pairing and stacking interactions (37). While there is no doubt that these factors prevent most LHEs from cleaving G-C rich sequences, this mechanism does not restrict nucleotide identity sufficiently to explain the overwhelming preference for C at the +2 positions of I-LtrI and I-OnuI target sites (Figure 4.8, Figure 4.10). Other than disfavoring G-C dinucleotide steps at the centre of their recognition sequence, there was no better predictor of I-LtrI and I-OnuI C4 sequence permissibility than having a C
at +2. This observation was reproducible between the *in vitro* enrichment and depletion experiments (Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11), as well as the *in vivo* genetic selection screens (Figure 4.3, Figure 4.4).

Although indirect readout is elegant, one of the best pieces of evidence in support of this mechanism is the fact that no direct contacts to C4 base pairs have been described in the literature. For this reason, indirect readout by LHEs has become dogmatic, and researchers (including myself!) have investigated C4 specificity with little regard for the role of direct-base contacts. The structural analysis presented in Chapter 4 contradicts this dogma, and suggests that I-LtrI makes direct contacts to the A-T and C-G base pairs at positions +1 and +2 position of the C4 region (Figure 4.19). Analysis of the I-LtrI protein-cognate DNA interface predicts the amino groups of R311 form an energetically favourable hydrogen bond network with the 4’ oxygen in the respective nucleobases. Interestingly, these contacts are not predicted in a previously published structure of I-LtrI (PDB: 3R7P) because the R311 side chain in that duplex has an alternative rotamer conformation (Figure 4.19A). The main difference between these two structures is that we obtained a pre-cleavage complex with an intact double helix, while 3R7P is post-cleavage and contains a double-stranded break.

For this reason, I re-examined every available pre-cleavage LHE-substrate structure, and identified 6 other LHEs with predicted direct C4 contacts (Table 4.4). This included I-OnuI which was predicted to make energetically favorable contacts to the T-A and C-G base pairs at positions +1 and +2 through the side chain of K229. Interestingly, the exact same hydrogen bond network has been identified in other protein-DNA structures (PDB: ISVC) (38), and this interaction could explain why I-OnuI demonstrated activity against most C4 sequences that ended with T-C (i.e. NNTC) (Figure 4.10). To my surprise, the first structure of I-OnuI (39) also identified this interaction but there was no mention of the direct-base contacts within the manuscript text (Figure 5.2). These findings suggest that during substrate binding, and prior to cleavage, some LHEs discriminate between C4 sequences through direct-base contacts.
Intriguingly, not all pre-cleavage LHE structures were found to make direct contacts to C4 nucleotides. In some structures, the helical and loop regions that contain contacting residues in I-LtrI and I-OnuI were entirely absent (40), while some results were inconclusive because no density was observed for side chains that would be expected to contact DNA. This lack of electron density could be explained by the location of contacting residues in flexible loop regions, which are often poorly visualized in crystallographic structures. However, I believe these findings exemplify the diversity that is witnessed in the LHE family. As previously mentioned, some LHEs (e.g. I-SceI(41)) have evolved to fulfill biological functions within their host, while others (e.g. I-TnaI (42)) have been shown to contain maturase activity in addition to their ability to hydrolyze DNA. Therefore, it would not be surprising to discover that members of the LHE family have also evolved the capacity to contact bases of the C4 nucleotides.

Admittedly, the direct C4 base contacts described in this text require extensive experimental validation to prove that computational predictions are not artifacts from crystallography. However, Occam’s razor would suggest direct readout as the simplest explanation for many observations in this thesis including: (i) the strong nucleotide preference I have determined for the corresponding positions; (ii) similar hydrogen bonding networks being employed by other DNA-binding proteins to recognize the same substrates; (iii) the documentation of these contacts in analyses from previously published manuscripts; and (iv) the identification of predicted C4 contacts in 6 other LHE family members.
Figure 5.2: The I-OnuI contact map.

Visual representation of the I-OnuI (PDB: 3QQY) protein-DNA interface clearly shows two direct central four contacts. Modified with permission (Appendix S1.1) from Takeuchi et al. (2011).
5.7 Considerations for Re-Engineering LADLIDADG Homing Endonuclease (LHE) Cleavage Specificity and Final Thoughts

The results presented in this thesis identify amino acid interactions, C4 nucleotide restrictions, and interactions with C4 base pairs that are important for understanding LHE biology. In addition, these findings should also be considered when re-engineering native LHE scaffolds, or engineering novel DFCs:

1. Altering amino acid identity at the coevolving positions 28, 29, 183 and 184 should be avoided. Data in these studies suggest that maintaining residue combinations found naturally within the LHE family are optimal, while pairings that are not represented could negatively impact enzyme functionality. Therefore, if alternations to the network are required then mutation and functional selection should be performed to achieve optimal residue pairings.

2. If a native scaffold, or DFC, contains an AEGE coevolving network then including the E184D substitution will likely increase catalytic efficiency, and could expand the number of C4 substrates that are cleavage by the enzyme. E184D mutations may also increase targetability of GDAE variants but this has not been experimentally determined.

3. The engineering of DFCs should be limited to domains within each of the two major LHE clades (i.e. D29 or E29 LHEs). This is recommended because constraints imposed by this position have differentially affected LHE fitness landscapes, and presumably have led to structural divergence and incompatibility between these evolutionary distinct scaffolds.

4. Central four (C4) nucleotide preference is scaffold-dependent. For this reason, the cleavage preference of each LHE should be profiled using the product enrichment or substrate depletion assays that were described in Chapter 4 of this thesis. This
knowledge will increase the success of LHE re-engineering by appreciating the limitations of the respective scaffold’s targetability.

5. There is strong evidence that some LHEs (e.g. I-LtrI and I-OnuI) make direct-base contacts to the C4 region of their recognition site. Although these observations require experimental validation, the existence of direct C4 nucleotide contacts could facilitate re-engineering of LHE scaffolds towards genomic sequences that are not currently targetable.

LHEs represent a powerful class of genome-editing (GE) reagent that will maintain niche market applications regardless of the targeting power of other technologies such as CRISPR-Cas9. I hope the studies in this thesis facilitate LHE re-engineering towards genomic sequences of interest, and help to increase their applicability within the GE field.

Beyond GE, the sequence diversity of LHEs make these enzymes an ideal model system to apply and study covariation analyses. In addition, the DNA-binding and endonucleolytic activity of LHEs also provided endless opportunities to explore structure-function relationships, and facilitated experimental validation of covarying residues to prove amino acid coevolution. This enabled me to showcase a strategy to infer protein function and structure from genetic information, and emphasizes the need to expand these studies to other protein families.

I hope that one-day covariation analyses can be expanded to proteins within the human genome/proteome, and that the work within this thesis plays a small part in confirming the validity of these methods to identify biologically significant amino acid interactions.
5.8 References


160
Appendices

Appendix S 1: Copyright Permissions

S1.1: Copyright permission for Chapter 1 and 5

PNAS Permissions  
June 12, 2017 at 3:01 PM

To: Thomas McMurrough

From: Thomas McMurrough

Permission is granted for your use of the figure as described in your message. Please list a full citation for the PNAS article when re-using the material. Because this material published after 2008, a copyright note is not needed. There is no charge for this material, either. Let us know if you have any questions.

Best regards,

Executive Editor  
PNAS

To whom it may concern:

Please see attached PDF with a request to reprint a PNAS publication figure in my Doctoral thesis.

Sincerely,

Thomas McMurrough
S1. 2: No copyright permission required for Chapter 2

http://www.pnas.org/site/aboutpnas/rightperm.xhtml

Accessed: 2017-09-19

PNAS authors need **not obtain permission** for the following cases:

1. to use their original figures or tables in their future works;
2. to make copies of their articles for their own personal use, including classroom use, or for the personal use of colleagues, provided those copies are not for sale and are not distributed in a systematic way;
3. **to include their articles as part of their dissertations**; or
4. to use all or part of their articles in printed compilations of their own works. The full journal reference must be cited and, for articles published in volumes 90–105 (1993–2008), "Copyright (copyright year) National Academy of Sciences."
Appendix S 2: Supplemental Information for Chapter 2

S2 Supplemental Figures

Figure S2. 1: An alignment of 178 non-redundant single-chain LHEs was constructed as previously described (1), and hand-edited to remove potential maturase sequences.

A fasta formatted version of the multiple sequence alignment is provided as a separate file (FigureS2.1.pdf).
Figure S2. 2: A cladogram of single-chain LHEs generated from an unrooted maximum-likelihood tree made by PhyML.

Outer ring contains LHE GI numbers as well as the identity of residues at positions 28, 183, 29, and 184. The text is also colored according to the identity of the metal- binding residues at positions 29 and 184: E_E is gold, D_E is red, E_D is blue, and D_D is green.
Figure S2.3: Heat map of log2 effect size for all 1,600 I-LtrI variants tested using a liquid two-plasmid genetic selection.
Data are displayed in alphabetical order by the identity of amino acids at positions 28 and 183 (28_183). (A) Constructs containing A_A to F_Y (Left) and G_A to L_Y (Right). (B) Constructs containing M_A to R_Y (Left) and S_A to Y_Y (Right). Effect size was determined using a modification of the ANOVA-like differential analysis method outlined in Fernandes et al. (2).
Figure S2.4: Heat map of all I-HjeMI variants.
(A) Heat map of all I-HjeMI variants tested with a log2 effect size >3.0 and a false discovery rate <0.05. Data are displayed in alphabetical order by the identity of amino acids at positions 28 and 183 (28_183). (B) Heat map of log2 effect size for all 1,600 I-HjeMI variants.
Figure S2. 5: Example of slow-growth phenotypes displayed by I-LtrI and I-OnuI variants in the two-plasmid genetic selection.

(A) Selective plates containing the wild-type A28_G183:E29_E184 constructs (A_G:E_E) were grown for 16 h at 37 °C and displayed a uniform cell morphology. Variants A28_G183:E29D_E184D (A_G:D_D) and A28G_G183A:E29D_E184 (G_A:D_D) were grown for 24 h. Variants containing an E29D mutation often display a
nonuniform colony morphology (best illustrated by I-LtrI A_G:D_D). (B) Boxplots of time (min) for I-LtrI variants to reach midlog growth (A600 = 0.35) in selective media, with individual data points shown as dots. I-LtrI variants are indicated on the x axis with vector only (Vec1 and Vec2) and media without cells used as negative controls (n = 4 for all cultures).
A

Wild-type I-Ltrl (A_G:E_E)

Time (min)

60.90nM
121.7nM
243.4nM

Time (min)

486.9nM
973.7nM
1391nM

Linear Product (%)

Time (min)

Initial Rate (nM/min)

Concentration (nM)

Replicate 1
Replicate 2
Replicate 3
Figure S2. 6: (A) Example of a cleavage assay for wild-type I-LtrI (A_G:E_E) resolved on a 1% agarose Tris-borate EDTA gel. Seven time points (as well as a no-protein negative control) were collected over a 23-fold range of enzyme concentrations. Bands are as follows: nicked plasmid or open circle (OC), linear (Lin), and supercoiled (SC). Band intensities were measured using AlphaEaseFCTM, and the percentage cleavage was calculated by dividing the density of the linear band by the sum of the intensities of all three bands. The percentage cleavage for each enzyme concentration was plotted versus...
time using GraphPad Prism, and initial rates were determined using a one-phase association function. The rates from each time-course experiment were then plotted versus enzyme concentration, and a Michaelis–Menten fit was used to determine the parameters $k_{\text{cat}}^*$ and $K_M^*$. (B) Example of an I-LtrI variant (G_A:D_D) cleavage assay, labeled as in A. (C) One-phase association and Michaelis–Menten plots for I-LtrI variants A28G (G_G:E_E), G183A (A_A:E_E), E29D/E184D (A_G:D_D), A28S (S_G: E_E), and A28S/E184D (S_G:E_D). (D) Plot of $k_{\text{cat}}^*$ versus time to midlog growth ($A_{600} = 0.35$) in selective media for I-LtrI variants on a log10 scale. (E) Plot of $k_M^*$ versus time to midlog growth ($A_{600} = 0.35$) in selective media for I-LtrI variants on a log10 scale. Fit of the data to a linear regression model is shown by a black line, with the 95% confidence interval as a gray-shaded area.
S2 Supplemental Tables

Table S2. 1: Quartet residue frequencies (28_183:29_184) from the LHE multiple sequence alignment.

<table>
<thead>
<tr>
<th>Residue Positions</th>
<th>Metal-binding Residues</th>
<th>Frequency in LHE Multiple Sequence Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>28_183</td>
<td>29_184</td>
<td></td>
</tr>
<tr>
<td>A_G</td>
<td>E_D</td>
<td>21</td>
</tr>
<tr>
<td>A_G</td>
<td>E_E</td>
<td>30</td>
</tr>
<tr>
<td>A_G</td>
<td>D_D</td>
<td>4</td>
</tr>
<tr>
<td>G_A</td>
<td>D_D</td>
<td>28</td>
</tr>
<tr>
<td>G_A</td>
<td>E_E</td>
<td>15</td>
</tr>
<tr>
<td>G_A</td>
<td>D_E</td>
<td>52</td>
</tr>
<tr>
<td>G_G</td>
<td>E_E</td>
<td>7</td>
</tr>
<tr>
<td>G_G</td>
<td>D_E</td>
<td>1</td>
</tr>
<tr>
<td>G_G</td>
<td>D_D</td>
<td>2</td>
</tr>
<tr>
<td>G_R</td>
<td>D_E</td>
<td>1</td>
</tr>
<tr>
<td>G_S</td>
<td>D_E</td>
<td>3</td>
</tr>
<tr>
<td>G_S</td>
<td>D_D</td>
<td>8</td>
</tr>
<tr>
<td>G_T</td>
<td>D_D</td>
<td>1</td>
</tr>
<tr>
<td>G_V</td>
<td>D_D</td>
<td>2</td>
</tr>
<tr>
<td>S_G</td>
<td>D_D</td>
<td>3</td>
</tr>
</tbody>
</table>

Table S2. 2: Melting temperature (Tm) and enthalpy of denaturation ($\Delta$H) for I-LtrI and I-OnuI as determined by differential scanning calorimetry.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Residues (28_183:29_184)</th>
<th>Tm (°C)</th>
<th>$\Delta$H ($10^4$ cal/mole)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-LtrI</td>
<td>A_G:E_E (wt)</td>
<td>45.99 ± 0.046</td>
<td>2.451 ± 0.487</td>
</tr>
<tr>
<td></td>
<td>G_G:E_E</td>
<td>45.94 ± 0.15</td>
<td>2.952 ± 0.443</td>
</tr>
<tr>
<td></td>
<td>A_A:E_E</td>
<td>49.42 ± 0.048</td>
<td>2.692 ± 0.464</td>
</tr>
<tr>
<td></td>
<td>S_G:E_E</td>
<td>46.30 ± 0.02</td>
<td>3.489 ± 0.227</td>
</tr>
<tr>
<td></td>
<td>A_G:D_D</td>
<td>45.67 ± 0.027</td>
<td>3.472 ± 0.383</td>
</tr>
<tr>
<td>I-OnuI</td>
<td>A_G:E_E (wt)</td>
<td>42.12 ± 0.053</td>
<td>1.876 ± 0.308</td>
</tr>
<tr>
<td></td>
<td>G_G:E_E</td>
<td>41.40 ± 0.043</td>
<td>6.082 ± 0.619</td>
</tr>
<tr>
<td></td>
<td>A_A:E_E</td>
<td>45.53 ± 0.048</td>
<td>1.890 ± 0.302</td>
</tr>
<tr>
<td></td>
<td>S_G:E_E</td>
<td>40.52 ± 0.072</td>
<td>1.271 ± 0.297</td>
</tr>
</tbody>
</table>

Data include SE from cure-fitting for three replicates.
S2 Supplemental Datasets

Dataset S2.1: Output from the LHE covariation analysis.

Attached as “DatasetS2.1.xls”

Dataset S2.2: Survival percentages of individual coevolving network variants.

Attached as “DatasetS2.2.xls”

S2 Supplemental References


Appendix S 3: Supplemental Information for Chapter 4

S3 Supplemental Materials and Methods

Crystallography substrates

I-LtrI ATAC:

5’-GGTCTAAACGTCGATAGGAGCATTT-3’
5’-CAAATGCTCCTATACGACGTTTAGAC-3’

I-LtrI AACC:

5’-GGTCTAAACGTCGTTAGGAGCATTT-3’
5’-CAAATGCTCCTAACCAGCGTTTAGAC-3’

Crystallization procedures

Protein preparations were initially combined with hybridized substrate molecules in a 1:1.5 ratio (protein:substrate) in the presence of 30mM CaCl$_2$, and incubated for a minimum of 4 hours at 4°C to promote pre-cleavage complex formation. Broad crystallization screens were then performed using the hanging-drop method and commercially available crystallization solutions (Wizard II, Ragaku Reagents Inc., formally Emerald BioStructures).

Wild-type I-LtrI (AEGE) crystals in complex with the cognate (ATAC) substrate grew in a 1:1 ratio of protein (~5mg/mL) and precipitant solution (20%[w/v] PEG-2000 MME, 0.1M Tris-HCl, pH 7.0, and 30mM CaCl$_2$). I-LtrI variant (AEAE, GEGE, and AEGD) crystals in complex with the cognate (ATAC) substrate grew in a 1:1 ratio of protein (~5mg/mL) and precipitant solution (10%[w/v] PEG-8000, 0.1M CHES, pH 9.5, 0.2M NaCl, and 30mM CaCl$_2$). Wild-type I-LtrI (AEGE) crystals in complex with the non-cognate (AACC) substrate grew in a 1:1 ratio of protein (~5mg/mL) and precipitant solution (2.5M NaCl, Na/K phosphate, pH 6.2, and 30mM CaCl$_2$). I-LtrI E184D (AEGD) crystals in complex with the non-cognate (AACC) substrate grew in a 1:1 ratio of protein (~5mg/mL) and precipitant solution (20%[w/v] PEG-2000 MME, 0.1M Tris-HCl, pH
7.0, and 30mM CaCl$_2$). All droplets were equilibrated over 1.5 M ammonium sulfate at 18°C, and crystal growth was achieved within 7 days for all constructs.

**Diffraction data reduction and processing**

Diffraction data were collected at Beam 17-ID at the Advanced Photon Source of Argonne National Labs. Data were collected in half degree wedges with an exposure time of 0.0877 seconds/image. 1080 images representing 270 degrees of rotation or 1440 images representing 360 degrees comprise the data set. Images were indexed and integrated using iMOSFLM. Reflections were scaled and merged using the Aimless and Ctruncate modules from CCP4i (1,2).

Merged reflections were then used for molecular replacement in PHENIX. An existing I-LtrI structure (PDB: 3R7P) was used as a search model for all mutants. Corresponding mutations to the protein structure and changes in the core 4 cleavage site were done manually in coot. Models were refined manually in coot and using the refine module from Phenix until the Rfree and Rwork factors converged. The data collection and model refinement statistics are listed in Dataset S3.1.

**S3 Supplemental Datasets**

**Dataset S3.1: I-LtrI variant structural refinement information.**

Attached as “DatasetS3.1.xls”

**S3 Supplemental References**


Curriculum Vitae

Name: Thomas Alan McMurrough

Post-secondary Education and Degrees:
The University of Western Ontario
London, Ontario, Canada
2011-2017, Ph.D. Biochemistry

Honours and Awards:
Doctoral Excellence Research Award
The University of Western Ontario
2016, Value $10 000

Alexander Graham Bell Canada Graduate Scholarship
Natural Sciences and Engineering Research Council (NSERC)
Government of Canada
2015-2016, Value $70 000

Ontario Graduate Scholarship (declined)
Government of Ontario
2015, Value $15 000

Dr. Bishnu D. Sanwal Graduate Performance Award
The University of Western Ontario, Department of Biochemistry
2014, Value $1 000

Western Graduate Research Scholarship
The University of Western Ontario, Department of Biochemistry
2011-2016, Annual Value $7 800

Related Work Experience:
Teaching Assistant
The University of Western Ontario
Biochemistry and Molecular Biology (BIOCHEM 2280A)
2014-2016
Teaching Assistant
The University of Western Ontario
Biochemistry Laboratory (BIOCHEM 3380G)
2013-2015

Teaching Assistant
The University of Western Ontario
Macromolecular Informatics (BIOCHEM 4445F)
2013-2014

Publications:


Protein Structures Solved:


