MOLECULAR BASIS OF THE COLIBACTIN BIOSYNTHETIC PATHWAY: A STRUCTURAL BIOLOGY APPROACH

By

NAGA SANDHYA GUNTAKA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2018
To my Mom and Dad
ACKNOWLEDGMENTS

I am greatly appreciative of the guidance that my dissertation advisor Professor Steven D. Bruner has provided to me over the last five years. Steve believed in me when I wanted to completely change my field of research and join his lab with no background in structural biology. The first few months in Steve’s group were extremely exciting and equally challenging as I dove into natural product research. I will never forget the courage, guidance and support Steve provided throughout my graduate career, especially in stressful times where he allowed me to choose and work on multiple projects. Steve’s understanding nature, and willingness to assist my career goals has allowed me to experience the best of the graduate school. One simply could not wish for a better or friendlier supervisor.

I also thank my committee members Profs. Nicole Horenstein, Aaron Aponick, Rebecca Butcher, and Yousong Ding for their valuable suggestions. I thank my collaborators Profs. Jason Crawford, Seth Herzon and Dr. Alan Healy from Yale University for providing synthetic compounds used in the projects discussed in Chapter 2 and 4. Their invaluable suggestions and assistance have been vital for my research. I extend my thanks to Prof. Christian Jobin from UF College of Medicine and Prof. Zhongwu Guo from UF Department of Chemistry for their collaboration on other projects. I gratefully acknowledge the funding towards my PhD from UF Department of Chemistry and the National Science Foundation (Grant 1411991 to S.D.B).

I have been lucky to have wonderful colleagues and friends who made my years at graduate school more memorable. I am grateful to my former colleagues, especially Wei-Hung Chen for teaching me basic molecular biology techniques. A warm thanks to Brian MacTavish, Sasha Zagulyaeva, Prabhanshu Tripathi, Qiang Li, Ariel Li, Krishna
Patel, and Madhushi Ratnayake for their support and for all the fun times in and out of the lab.

I owe so much to many people beyond the few listed here. I wish warm thanks to Jessica, Erica, Mayra, Kim, and Clari for sharing the good memories. I’d like to thank Purna for his support and friendship. I am grateful to Sridhar for seeding the thought in my mind to pursue research in the United States. I’m thankful for my friendship with JD for helping me smile and find hope and courage during difficult times. Thank you for being part of my journey and for being there for me when needed the most. Undoubtedly, I owe a deep respect and gratitude to all my teachers and friends who have guided me in choosing the right path throughout my life and career. I’m grateful for Prof. Cory Emal, my previous graduate advisor at Eastern Michigan University for his extended support and guidance throughout these years.

Finally, to the ultimate sources of love, support and inspiration: my family. I am incredibly thankful to my sister Naga Swapna Kallam and brother-in-law Ravindra Reddy Kallam, for their invaluable love and care. Thank you so much for not only supporting my choices but also for sharing my responsibilities back home and making my journey a lot easier. I’m thankful to my grandparents, cousins, and other family members for their support. Finally, my deep sense of gratitude to my parents: Jyothi Kumari and Satya Harinadha Reddy Guntaka, for their incredible love and support. Despite societal pressures and rough times, you have been put through all these years, you endured all the pain and allowed me to reach my goals. I can’t thank you enough for all the freedom you gave me, and I couldn’t have done this without your encouragement, unconditional love, and support.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................. 4
LIST OF TABLES ................................................................. 9
LIST OF FIGURES ............................................................... 10
LIST OF ABBREVIATIONS ..................................................... 13
ABSTRACT ................................................................................... 16

CHAPTER

1 INTRODUCTION ........................................................................ 18
  1.1 Role of Bacteria in Human Health and Disease ...................... 18
  1.2 Role of E. coli “pks island” in CRC Pathogenesis .................. 19
  1.3 Colibactin Biosynthetic Pathway ........................................ 22
  1.4 Isolation, Synthesis and Structural Characterization of (Pre)Colibactins ...... 34
  1.5 Summary ............................................................................ 35

2 STRUCTURE AND FUNCTIONAL ANALYSIS OF CLBQ, AN UNUSUAL INTERMEDIATE-RELEASING THIOESTERASE FROM THE COLIBACTIN BIOSYNTHETIC PATHWAY ................................................................. 38
  2.1 Thioesterase Enzymes of NRPS/PKS Systems ....................... 38
  2.2 ClbQ is a TEII Enzyme in the Colibactin Biosynthetic Pathway ................................................................. 40
  2.3 Biochemical Characterization of ClbQ .................................. 42
    2.3.1 ClbQ-Mediated Release of AM Unit from the Colibactin Biosynthetic Pathway ................................................................. 43
    2.3.2 Substrate Specificity of ClbQ .......................................... 45
    2.3.3 Selectivity of ClbQ toward Carrier Proteins in Colibactin Pathway ................................................................. 48
  2.4 Crystallization, Structure Determination of ClbQ ................... 51
    2.4.1 Expression, Purification and Crystallization .................... 51
    2.4.2 Crystal Structure of ClbQ ............................................. 53
    2.4.3 Conformational Changes of ClbQ .................................... 58
    2.4.4 Structural Homology of ClbQ ......................................... 60
  2.5 Conclusions and Future Work .............................................. 62
  2.6 Experimental Procedures .................................................. 62
    2.6.1 Expression Plasmids .................................................... 62
    2.6.2 Protein Expression and Purification .................................. 63
    2.6.3 Crystallization ............................................................ 64
    2.6.4 Data Collection and Structure Determination .................... 65
    2.6.5 Acylation of Carrier Protein Domains ............................. 65
    2.6.6 Acyl-ACP Hydrolysis .................................................. 66
2.6.7 AM-ClbE Hydrolysis Assays ................................................................. 66
2.6.8 Precolibactin SNAC Hydrolysis Assays .................................................. 67
2.6.9 Acyl-SNAC Hydrolysis versus Precolibactin SNAC Hydrolysis ................. 67
2.6.10 Synthesis of Acyl-SNACs .................................................................. 68
   2.6.10.1 S-(2-acetamidoethyl) ethanethioate (Acetyl SNAC) ....................... 68
   2.6.10.2 S-(2-acetamidoethyl) dodecanethioate (Lauroyl SNAC) ............... 68
   2.6.10.3 S-(2-acetamidoethyl) tetradecanethioate (Myristoyl SNAC) ......... 69
2.6.11 Synthesis of Precolibactin SNACs ...................................................... 69
   2.6.11.1 S-(2-acetamidoethyl) (R)-4-amino-4-oxo-2-tetradecanamidobutanethioate (N-myristoyl-D-asparagine-SNAC; 20) . 69
   2.6.11.2 S-(2-acetamidoethyl)-2'-(3-(1-(S)-6-((R)-4-amino-4-oxo-2-tetradecanamido butanamido)-3-oxoheptanamido) cyclopropyl)-3-oxopropaaamido) methyl]-[2,4'-bithiazole]-4-carbothioate (Linear Precolibactin SNAC; 26) .......................................................... 70
   2.6.11.3 S-(2-acetamidoethyl)2'-(4'-(S)-3-((R)-4-amino-4-oxo-2-tetradecanamido butanamido) butyl)-3',6'-dioxo-2',3'-dihydrospiro[cyclopropane-1,1'-pyrrolo[3,4-c]pyridin]-5'((6'H)-yl) methyl]-[2,4'-bithiazole]-4-carbothioate (Cyclic Precolibactin SNAC; 27) .......................................................... 71
2.6.12 Catalog of NMR Data of Acyl-SNACs and Precolibactin SNACs .......... 72

3 STRUCTURAL STUDIES OF THE S-ADENOSY-L-METHIONINE (SAM) ACTIVATING ADENYLATION DOMAIN OF NRPS MODULE CLBH IN THE COLIBACTIN PATHWAY .................................................................. 83

3.1 Adenylation Domain of NRPS System ....................................................... 83
3.2 ClbH, a Unique NRPS Module in the Colibactin Pathway ......................... 85
3.3 Cloning, Expression and Crystallization of ClbH C-A2-PCP Domain .......... 89
3.4 Future Directions ..................................................................................... 93
3.5 Experimental Procedures ....................................................................... 94
   3.5.1 Cloning of ClbH Constructs ............................................................... 95
   3.5.2 Protein Expression and Purification of ClbH C-A2-PCP ....................... 95
   3.5.3 Crystallization of ClbH C-A2-PCP ...................................................... 96

4 EXAMINING THE ROLE OF CLBL IN THE COLIBACTIN BIOSYNTHETIC PATHWAY .................................................................................. 98

4.1 The Uncharacterized Amidase, ClbL ......................................................... 98
4.2 Preliminary Crystallization Data of ClbL ................................................... 99
4.3 Optimization of ClbL Crystals ................................................................. 100
4.4 Site-Directed Mutagenesis of ClbL Active Site Residues .......................... 101
4.5. Proposed Roles of ClbL in the Colibactin Biosynthetic Pathway .......... 103
4.6. Structural Homology Study of ClbL ...................................................... 106
4.6. FAAH Activity of ClbL ......................................................................... 110
4.7 Experimental Procedures ....................................................................... 112
   4.7.1 Cloning of ClbL and ClbL K80A Mutant ........................................... 112
   4.7.2 Expression and Purification of ClbL and ClbL K80A Mutant ............... 113
4.7.3 Crystallization of ClbL and ClbL K80A .......................................................... 113
4.7.4 Data Collection and Structure Determination .................................................. 114
4.7.5 ClbL Fatty Acid Amide Hydrolysis Assay ....................................................... 115

LIST OF REFERENCES .................................................................................................. 116

BIOGRAPHICAL SKETCH ........................................................................................... 126
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Crystallographic data of ClbQ.</td>
<td>55</td>
</tr>
<tr>
<td>2-2</td>
<td>Primers used in the study.</td>
<td>63</td>
</tr>
<tr>
<td>3-1</td>
<td>Primers used in the study</td>
<td>95</td>
</tr>
<tr>
<td>4-1</td>
<td>NCBI BLAST analysis of ClbL.</td>
<td>98</td>
</tr>
<tr>
<td>4-2</td>
<td>Primers used in the study.</td>
<td>112</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Schematic of minimal NRPS initiation, chain extension and termination module.</td>
<td>23</td>
</tr>
<tr>
<td>1-2</td>
<td>Schematic of minimal PKS initiation, chain extension and termination modules.</td>
<td>24</td>
</tr>
<tr>
<td>1-3</td>
<td>The <em>pks</em> gene island encoding the hybrid NRPS/PKS system.</td>
<td>25</td>
</tr>
<tr>
<td>1-4</td>
<td>Proposed biosynthetic pathway and chemical structures of reported precolibactin molecules.</td>
<td>27</td>
</tr>
<tr>
<td>1-5</td>
<td>Aminomalonyl (AM) synthesis by accessory enzymes.</td>
<td>30</td>
</tr>
<tr>
<td>1-6</td>
<td>Proposed mechanistic model for the production of precolibactin metabolites.</td>
<td>34</td>
</tr>
<tr>
<td>2-1</td>
<td>Role of Type I and Type II thioesterase enzymes in NRPS/PKS.</td>
<td>38</td>
</tr>
<tr>
<td>2-2</td>
<td>Sequence homology analysis of ClbQ.</td>
<td>41</td>
</tr>
<tr>
<td>2-3</td>
<td>Effect of ClbQ on <em>clb</em> pathway.</td>
<td>42</td>
</tr>
<tr>
<td>2-4</td>
<td>Proposed hypothesis on the role of ClbQ off-loading AM unit.</td>
<td>43</td>
</tr>
<tr>
<td>2-5</td>
<td>SDS-PAGE analysis of proteins used in the <em>in vitro</em> assays.</td>
<td>44</td>
</tr>
<tr>
<td>2-6</td>
<td>Hydrolysis of AM-ClbE with the ClbQ.</td>
<td>44</td>
</tr>
<tr>
<td>2-7</td>
<td>Hydrolysis of precolibactin thioester SNAC derivatives with ClbQ.</td>
<td>46</td>
</tr>
<tr>
<td>2-8</td>
<td>Acyl-ACP hydrolysis by ClbQ.</td>
<td>49</td>
</tr>
<tr>
<td>2-9</td>
<td>Sequence alignment of carrier protein domains in the colibactin biosynthetic pathway.</td>
<td>50</td>
</tr>
<tr>
<td>2-10</td>
<td>Crystallization of ClbQ.</td>
<td>52</td>
</tr>
<tr>
<td>2-11</td>
<td>Crystal structure of ClbQ.</td>
<td>54</td>
</tr>
<tr>
<td>2-12</td>
<td>Conformational flexibility of ClbQ.</td>
<td>57</td>
</tr>
<tr>
<td>2-13</td>
<td>Interactions of βME with Cys49 of each monomer of ClbQ.</td>
<td>58</td>
</tr>
<tr>
<td>2-14</td>
<td>Structural homology analysis of ClbQ.</td>
<td>60</td>
</tr>
<tr>
<td>2-15</td>
<td>Electrostatic potential map of the active site of ClbQ (chain A).</td>
<td>61</td>
</tr>
</tbody>
</table>
1H NMR spectrum of the Acetyl SNAC. ................................................................. 72
13C NMR spectrum of the Acetyl SNAC............................................................... 73
1H NMR spectrum of the Lauryl SNAC. ............................................................... 74
13C NMR spectrum of the Lauryl SNAC. ............................................................... 75
1H NMR spectrum of the Myristoyl SNAC. ............................................................ 76
13C NMR spectrum of the Myristoyl SNAC. ............................................................ 77
1H NMR spectrum of the N-myristoyl-D-asparagine-SNAC; 20. ......................... 78
1H NMR spectrum of the Linear Precolibactin SNAC; 26................................. 79
13C NMR spectrum of the Linear Precolibactin SNAC; 26. ............................... 80
1H NMR spectrum of the Cyclic Precolibactin SNAC; 27................................. 81
13C NMR spectrum of the Cyclic Precolibactin SNAC; 27................................. 82
Adenylation domain reaction. ............................................................................. 83
Proposed ways to incorporate ACC in the colibactin biosynthesis. ..................... 87
ClbI-dependent cyclopropanation of SAM activated by ClbH. ......................... 88
Schematic of ClbH expression constructs used for cloning. ............................... 91
Size exclusion chromatography of ClbH-C-A2-PCP protein after Ni-NTA resin purification. ........................................................................................................ 92
Crystal hits of apo- ClbH C-A2-PCP in various conditions. ............................... 93
Adenylation specific inhibitors for ClbH. ............................................................. 94
Unpublished crystal structure of ClbL as a crystallographic homodimer............. 100
Crystal structure of ClbL as a crystallographic monomer. .................................. 101
SDS-PAGE analysis of expression of ClbL mutants at varying conditions. ...... 102
ClbL function as a GatA homologue. ................................................................. 103
Possible role of ClbL in modification of colibactin structure. ......................... 104
ClbL-dependent release of ammonia alters pH of the epithelial microenvironment and favors colonization by pathogenic bacteria. ............... 106
Structural alignment of ClbL with humanized rat FAAH-1.......................... 107

Structural alignment of X-ray crystal structure of ClbL with predicted structural model from SWISS-MODEL server.................................................... 108

Structural homology of ClbL................................................................. 109

FAAH activity of ClbL using DepNA as a substrate.................................. 110

Possible fatty-acid amide substrates for ClbL......................................... 111
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenylation</td>
</tr>
<tr>
<td>ACC</td>
<td>1-Aminocyclopropane-1-Carboxylic acid</td>
</tr>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>ACP</td>
<td>Acyl Carrier Protein</td>
</tr>
<tr>
<td>AM</td>
<td>Amino Malonyl</td>
</tr>
<tr>
<td>AMP</td>
<td>Adenosine monophosphate</td>
</tr>
<tr>
<td>AMT</td>
<td>Aminotransferase</td>
</tr>
<tr>
<td>APS</td>
<td>Advanced Photon Source</td>
</tr>
<tr>
<td>AT</td>
<td>Acyl Transferase</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>AVS</td>
<td>Adenosine vinyl sulfonamide</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>BGC</td>
<td>Biosynthetic Gene Cluster</td>
</tr>
<tr>
<td>C</td>
<td>Condensation</td>
</tr>
<tr>
<td>CoA</td>
<td>Coenzyme A</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal Cancer</td>
</tr>
<tr>
<td>Cy</td>
<td>Cyclization</td>
</tr>
<tr>
<td>DePNA</td>
<td>Decanoyl p-Nitroaniline</td>
</tr>
<tr>
<td>DH</td>
<td>Dehydratase</td>
</tr>
<tr>
<td>DMAP</td>
<td>4-Dimethylaminopyridine</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl Sulfoxide</td>
</tr>
<tr>
<td>DSB</td>
<td>Double strand break</td>
</tr>
<tr>
<td>DTNB</td>
<td>5,5'-dithiobis (2-nitro-benzoic acid)</td>
</tr>
<tr>
<td>E</td>
<td>Epimerization</td>
</tr>
</tbody>
</table>
EcN  
*E. coli* Nissle 1917

EDCI  
1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide

EIC  
Extracted-Ion Chromatography

ER  
Enoyl reductase

ESI-MS  
Electro Spray Ionization-Mass Spectrometry

FAAH  
Fatty acid amide hydrolase

FASs  
Fatty acid Synthases

FPLC  
Fast Protein Liquid Chromatography

Fur  
Ferric uptake regulator

GABA  
gamma-Aminobutyric acid

HPLC  
High Performance Liquid Chromatography

IBD  
Inflammatory Bowel Disease

IPTG  
Isopropyl β-D-thiogalactopyranoside

KR  
Keto Reductase

KS  
Keto Synthase

LC-MS  
Liquid Chromatography Mass Spectrometry

LS-CAT  
Life Sciences Collaborative Access Team

M  
Methylation

MALDI-TOF  
Matrix-assisted Laser Desorption/Ionization-Time of Flight

MATE  
Multidrug and toxic compound extrusion

MLP  
MbtH-like protein

MS  
Mass Spectrometry

NMDA  
*N*-Methyl-D-Aspartate

NMR  
Nuclear Magnetic Resonance

NRPs  
Nonribosomal Peptides
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRPS</td>
<td>Nonribosomal Peptide Synthetase</td>
</tr>
<tr>
<td>Ox</td>
<td>Oxidation</td>
</tr>
<tr>
<td>PCP</td>
<td>Peptidyl Carrier Protein</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PEG</td>
<td>Polyethylene glycol</td>
</tr>
<tr>
<td>PKs</td>
<td>Polyketides</td>
</tr>
<tr>
<td>PKS</td>
<td>Polyketide Synthase</td>
</tr>
<tr>
<td>PLP</td>
<td>Pyridoxal phosphate</td>
</tr>
<tr>
<td>Ppant</td>
<td>Phosphopantetheine</td>
</tr>
<tr>
<td>PPTase</td>
<td>Ppant transferase</td>
</tr>
<tr>
<td>R</td>
<td>Reduction</td>
</tr>
<tr>
<td>SAM</td>
<td>S-adenosyl-L-Methionine</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulphate</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>SNAC</td>
<td>S-N-Acetyl Cysteamine</td>
</tr>
<tr>
<td>T</td>
<td>Thiolation</td>
</tr>
<tr>
<td>TE</td>
<td>Thioesterase</td>
</tr>
<tr>
<td>TEI</td>
<td>Thioesterase Type I</td>
</tr>
<tr>
<td>TEII</td>
<td>Thioesterase Type II</td>
</tr>
<tr>
<td>β-ME</td>
<td>β-Mercaptoethanol</td>
</tr>
</tbody>
</table>
MOLECULAR BASIS OF THE COLIBACTIN BIOSYNTHETIC PATHWAY: A STRUCTURAL BIOLOGY APPROACH

By

Naga Sandhya Guntaka

May 2018

Chair: Steven D. Bruner
Major: Chemistry

Small molecule microbial secondary metabolites, by regulating host-microbe interactions, play an important role in all aspects of disease etiology and treatment. Colibactin is a secondary metabolite linked to the progression and pathogenesis of colorectal cancer (CRC) and inflammatory bowel disease (IBD) by inducing DNA damage in host cells. The chemical details of colibactin and the biosynthetic pathway are emerging but clearly are unusual and noncanonical. Despite recent successes in understanding the complex colibactin synthetic machinery and function of genes involved in the biosynthesis, the structure and mechanism of action of colibactin remain elusive. Additionally, proposed gene functions and biosynthetic machinery of colibactin pathway, including the predicted structures, need to be validated by NMR or MS-independent methods. Therefore, we are applying a structural approach of key gene products involved in its biosynthesis and genotoxic activity.

Chapter 2 discusses a key aspect of colibactin biosynthesis, the occurrence of multiple metabolites, and the biosynthetic rationale for this through structural and biochemical studies of a thioesterase enzyme, ClbQ. Chapter 3 discusses structural studies of the adenylation domain of the ClbH protein and Chapter 4 discusses
preliminary structural characterization of an unannotated enzyme, ClbL, providing information on its proposed role in the colibactin biosynthesis and mechanism.
CHAPTER 1
INTRODUCTION

1.1 Role of Bacteria in Human Health and Disease

Human microbiota, a collection of trillions of resident microbes living on or in the human body play major roles in human health and disease.\(^1\,^3\) An average human body, consisting of about ten trillion cells, has approximately ten times that number of microorganisms in the gut.\(^4\) Gut microbiota perform metabolic activity which is equal to that of a virtual organ, including functions beneficial to the host such as training the immune system, preventing growth of harmful species, regulating the development of the gut, producing vitamins and hormones for the host, and fermenting unused energy substrates.\(^3\) While the intestinal microbial community exhibits host beneficial functions in a eubiosis homeostasis state, research suggests that dysbiosis\(^5\) resulting from an imbalance of symbionts (health-promoting bacteria), commensals (permanent residents of no benefit or detriment to the host), and pathobionts (pathogens) can be directly correlated to human diseases.\(^3\,^6\) Thus, the gut microbiome, encoding at least two orders of magnitude more genes than the human genome, plays a vital role in maintaining host health and has a profound effect on human diseases, including obesity, diabetes, cardiovascular diseases, inflammatory bowel diseases (IBD), colorectal cancer (CRC) and others.\(^3\,^6\,^7\) Many of these functions are likely carried through small molecule secondary metabolites including terpenes, saccharides, alkaloids, macrolide polyketides (PKs) and nonribosomal peptides (NRPs), which are produced by the microbial biosynthetic gene clusters (BGCs). These secondary metabolites exert effects on both the microbial community structure and host physiology by regulating various metabolic process.\(^7\)
Although the correlation of bacteria to human health and disease is being studied extensively, the contributions of secondary metabolites to human health and regulatory chemical signaling process remain largely unexplored. For example, select gut bacteria are correlated at the nucleotide-sequence level to colitis-associated CRC; however, the biosynthetic pathway of “colibactin”, a small molecule produced by these select gut microbes remains elusive. In this study, we explore the molecular basis of colibactin biosynthetic pathway.

1.2 Role of E. coli “pks island” in CRC Pathogenesis

CRC, also referred to as colon cancer is the second leading cause of cancer-related death, and is the third most common form of cancer in both men and women in the United States. One key environmental factor linked to CRC development is the gut microbiota. Human gut microbiota encompasses $10^{14}$ bacteria and $10^{16}$ bacterial genes, which play a major role in regulating the stability of the intestine. A wealth of studies conducted in both human and mouse models suggest the role of microbial dysbiosis and alterations of host-microbiota interactions in CRC pathogenesis. With such a diverse population inhabiting the human gut, identifying different species of microbes and differentiating between host beneficial and host pathogenic bacteria is challenging. However, with the recent advancements in gene sequencing techniques, analysis of deep sequencing of the 16S rRNA within the luminal compartment of the intestine of CRC patients resulted in a higher prevalence of various bacterial strains including Enterococcus, Escherichia/Shigella, Klebsiella, Streptococcus, and Peptostreptococcus. In this technique, universal PCR primers are used to target conserved regions of 16S rRNA gene, a highly conserved component of the
transcriptional machinery of all DNA-based life forms, which allows the identification and classification of microbes.\textsuperscript{13}

*Escherichia coli* (*E. coli*), a consistent member of the human intestinal microbiota, is a versatile organism that produces a diverse array of virulence factors manipulating basic host cell functions.\textsuperscript{14} Certain strains of mucosa-associated pathogenic *E. coli* are more frequently identified in colon tissue from CRC patients compared to controls.\textsuperscript{15} The role of certain strains of *E. coli* in CRC pathogenesis is evidenced by the ability of adherent-invasive *E. coli* NC101 to promote double-stranded DNA breaks and enhance tumor development in the azoxymethane (AOM)-treated interleukin-10 (IL-10 -/-) mouse model of colitis-associated CRC.\textsuperscript{16} Further research studies conducted on two other strains of *E. coli*, including 11G5 and CCR20, have shown the role of *E. coli* in the development and pathogenesis of both IBD and CRC.\textsuperscript{8,17}

Among the four major phylogenetic groups of *E. coli* (A, B1, B2 and D), the B2 group can survive longer in the colon and comprises about 30-50\% of the strains isolated from healthy human feces.\textsuperscript{18,19} In 2006, Oswald and co-workers reported that 34\% of the commensal B2 *E. coli* strains carry a conserved 54-kilobase pathogenic gene island named the “pks island” which has hallmarks of horizontal gene transfer.\textsuperscript{8} The genomic analysis revealed that the pks island encodes a BGC that synthesizes a hybrid polyketide-peptide secondary metabolite named “colibactin”, and the biosynthetic assembly-line is discussed in detail below. Further studies indicated that pks+ *E. coli* strain both induce megalocytosis and block mitosis in cultured eukaryotic cells. The cytopathic effect of pks+ *E. coli* was shown to be contact-dependent and was not observed when cells were separated by porous 0.2 µm semipermeable insert. This
unusual pathogenicity requiring live cell-cell contact can be attributed to two factors. First, colibactin produced by the *pks* island is a highly unstable small molecule that degrades rapidly after release from the producing cell. This is further corroborated by the observation that the transcription of *pks* genes is contact-independent, suggesting that colibactin is constitutively produced but may be highly unstable in solution. Second, the colibactin might modify a specific protein, limiting the identification by current techniques used for small molecule analysis, such as mass spectrometry (MS).

Further investigations on various *E. coli* strains including pathogenic (IHE3034, SP15, and CFT073) and commensal (Nissle 1917) strains suggest that the cytotoxic effect of the *pks* island is mediated via DNA double-stranded breaks *in vitro*. In the same study, laboratory *E. coli* strain DH10B hosting a bacterial artificial chromosome (BAC) with the *pks* island triggered megalocytosis, whereas DH10B with an empty BAC vector did not. The evidence of colibactin inducing DNA double-stranded breaks (DSBs) *in vitro* and promoting intestinal tumorigenesis is also supported by the *pks* gene deletion experiments that resulted in prevention of adenocarcinoma in preclinical models. In 2010, *in vivo* evidence of DNA DSBs induced by *pks*+ *E. coli* was reported, suggesting common *E. coli* strains may contribute to cellular transformation and sporadic CRC tumorigenesis. Additional studies revealed that the *pks* island encoding colibactin modulates host immunity, exacerbates lymphopenia leading to sepsis, and contributes to pathogenicity of inflammation-induced CRC and senescence-induced tumor growth.

Mounting evidence of the role of gut bacteria in CRC pathogenesis and the shift of *E. coli* populations from the declining phylogenetic group A to the newly dominant
phylogenetic group B2, which possess the genotoxic pks island, could potentially represent a health concern. Remarkably, the probiotic strain of *E. coli* Nissle 1917 (EcN), which is used in the treatment of IBD, also harbors the pks island. Deletion of the pks island in EcN attenuated DNA damage as well as its probiotic activity in eukaryotic cells. Since EcN lacks defined virulence factors, the carcinogenic role of the pks island in closely related strains highlights the necessity to decipher the mechanism and structure of colibactin and other metabolites in the pathway. Recent findings suggest that the pks island is not limited to *E. coli*, but is also found in genomes of other proteobacteria, including sponge, honeybee-associated *Pseudovibrio* FO-BEG1, *Frischella perrara*, *Citrobacter koseri*, and *Klebsiella pneumonia*, thus indicating that colibactin might mediate evolutionarily conserved interactions with animal hosts and a broad range of symbionts. Overall, all these biological studies implicate the role of gut bacteria, pks, and colibactin, in the pathogenesis of CRC. Thus, there is a great need to elucidate the molecular basis of colibactin biosynthetic pathway to not only understand the synthesis of complex biological molecules by the gut bacteria involved in CRC pathogenesis but also to identify potential therapeutic targets for CRC treatment. Current understanding of the colibactin biosynthetic gene cluster is discussed below.

**1.3 Colibactin Biosynthetic Pathway**

Colibactin is a hybrid polyketide-peptide secondary metabolite that is encoded by the polyketide synthase (pks) genomic island in *E. coli* and other species of *Enterococci*. It is a very complex, highly unstable molecule synthesized by a hybrid polyketide synthetase (PKS) and non-ribosomal peptide synthetase (NRPS) system, which challenges the structure elucidation and prediction of mechanism of action. A
A brief review on NRPS and PKS systems is provided for better understanding of the colibactin biosynthetic pathway.

Both modular NRPS and PKS systems are among the largest and most complicated multimodular megasynthetase enzymes often found in bacteria and fungi. A recent review article by Weissman outlines the detailed insights and current knowledge on the structural biology of these biosynthetic megaenzymes. In brief, each of the NRPS and PKS systems are organized into multiple units called ‘modules’ containing independently folding protein domains joined together by “linker” regions, and operate in assembly-line fashion to construct polymeric chains and tailor their functionalities. Each domain and assembly of domains (module) catalyze region-specific and stereo-specific reactions to produce wide variety of medically important natural products with diverse structures and chemistries.

Figure 1-1. Schematic of minimal NRPS initiation, chain extension and termination module.
NRPS systems utilize both natural and unnatural amino acid building blocks to produce diverse polypeptides. In the minimal NRPS initiation and chain extension modules (Figure 1-1), biosynthesis begins with activation of a specific amino acid building block as its aminoacyl adenylate by an adenylation (A) domain and transfer to a phosphopantethienylated peptidyl carrier protein (PCP). The condensation (C) domain then catalyzes peptide bond formation using the two aminoacyl-PCPs as substrates. These series of reactions are repeated until the full-length peptide is synthesized. Release of the peptide is catalyzed by a fourth domain, a thioesterase (TE), that is often found at the C-terminus of the NRPS. Additionally, domains such as epimerization (E), cyclization (Cy), oxidation (Ox), reduction (R), and N-methylation (M) are occasionally found in these systems, adding structural diversity among various NRPs.

![Schematic of minimal PKS initiation, chain extension and termination modules.](image)

Figure 1-2. Schematic of minimal PKS initiation, chain extension and termination modules.

Similarly, PKSs perform analogous chain extension and modification cycles (Figure 1-2). PKs are derived from acetyl- or malonyl-CoA as starter units and malonyl-
or methylmalonyl-CoA as elongating units. Typical PKS systems contain
acyltransferase (AT), acyl carrier protein (ACP), ketosynthase (KS) and thioesterase
(TE). The carrier proteins in both NRPS and PKS systems are often referred as
thiolation (T) domains. Additionally, domains such as dehydratase (DH), enoyl
reductase (ER) and keto reductase (KR) are occasionally found in these systems,
contributing to the structural diversity seen among PKs.

NRPS and PKS are also found to work collaboratively in hybrid NRPS-PKS
systems, contributing to the broad substrate diversity and producing virtually limitless
and complex hybrid peptide polyketide compounds, including colibactin.

Figure 1-3. The pks gene island encoding the hybrid NRPS/PKS system.

Genetic and functional analyses of the E. coli pks island indicated that it encodes
due NRPSs (clbN,H,J), three PKSs (clbC,I,O), two hybrid NRPS/PKS (clbB,K), and
additional accessory enzymes (clbA,D-G,L,P,M,R,S) (Figure 1-3). Mutation of individual
genes revealed that all components of the assembly line, except a Na⁺/drug antiporter
(clbM) and a resistance gene (clbS) were required for genotoxicity. Although efforts
from multiple labs to isolate colibactin have been unsuccessful, the research attempts
have highlighted the role of the individual enzymes in the biosynthetic assembly line as
discussed in detail below. Chemical structures of all the precolibactin metabolites (1-19) discussed in the following paragraphs are included in Figure 1-4.

The colibactin assembly line encodes three NRPS enzymes: ClbN, ClbH and ClbJ. ClbN encodes set of C-A-PCP-E domains, where the A domain activates L-Asn, which is epimerized to D-Asn (E-domain), and is incorporated into the precolibactin metabolite 1. Balskus and co-workers conducted thorough bioinformatics analyses and \textit{in vitro} assays which revealed the role of the C-domain. This was found to connect the fatty acyl chain to D-Asn, further supporting the structure of \(N\)-myristoyl-D-Asn, compound 1. ClbH is a unique enzyme component of the colibactin NRPS machinery with a rare A-C-A-PCP domain arrangement as per homology predictions. In contrast to the canonical NRPS architecture, ClbH lacks a PCP domain downstream of the A domain at the N-terminus (ClbH-A\(_1\)). Instead, ClbH-A\(_1\) is directly connected to a C domain and activated L-Ser. Piel and co-workers have shown that the L-Ser activating ClbH-A\(_1\) domain, together with accessory enzymes ClbDEF, produce a rare 2-aminomalonyl (AM) unit. On the other hand, the ClbH-A\(_2\) domain lacks conserved active site residues, suggesting it does not activate an \(\alpha\)-amino acid. Contrary to the feeding studies that suggested ClbH-A\(_2\) uses L-Met or the L-Met-derived aminoacid 1-aminocyclopropane-1-carboxylic acid (ACC) as a building block, \textit{in vitro} assays performed by Balskus and co-workers supported ClbH-A\(_2\) activation of S-adenosyl-L-methionine (SAM). Thorough details on ClbH will be discussed in Chapter 3. ClbJ is a straightforward two module NRPS enzyme encoding C-A\(_1\)-PCP\(_1\)-Cy-A\(_2\)-PCP\(_2\) domains, where A\(_1\) and A\(_2\) domains are predicted to activate L-Gly and L-Cys, respectively, and the cyclization domain catalyzes a thiazoline ring formation.
Figure 1-4. Proposed biosynthetic pathway and chemical structures of reported precolibactin molecules.
The \textit{pks} island encodes three PKSs containing an unusual mixture of \textit{cis}- and \textit{trans}-AT PKSs. A \textit{cis}-AT type PKS module contains an AT domain within the same polypeptide as KS and ACP domains, whereas the \textit{trans}-AT PKS utilizes an AT that is not a part of the same polypeptide as the KS and ACP domains.\textsuperscript{38} The \textit{trans}-AT (AT\textsuperscript{r}) PKSs, ClbC and ClbO encode a set of KS-AT\textsuperscript{r}-ACP domains. ClbC was previously shown to incorporate malonyl-CoA extender unit to generate metabolite 4,\textsuperscript{39} and \textit{in vitro} assays also predicted the utilization of the AM unit by the trans-AT PKS modules ClbC and ClbO.\textsuperscript{40} However, the function and resulting metabolite(s) structures related to ClbO remains unclear. ClbI encodes a \textit{cis}-AT containing KS-AT-ACP domains and predicted to activate and incorporate malonyl-CoA.\textsuperscript{41} While the AT domain possesses a conserved malonyl-CoA activating signature motif, HAFH, and an active site serine, the ClbI KS domain lacks a conserved active site cysteine and instead has a serine residue (S178), suggesting ClbI cannot accept an intermediate from an upstream assembly line. Very recently, it was proposed that S178 of the ClbI KS domain could function as a nucleophile to tether the intermediate to ClbI or act as a general base to promote cyclization of SAM-derived ClbH-tethered intermediate, to provide metabolite 6.\textsuperscript{35,39} This observation is consistent with the report suggesting that decarboxylation and cyclization of the intermediate produced by ClbI is abolished when KS and ACP domains are inactivated, whereas inactivation of ClbI-AT only partially reduced the production of metabolite 6.\textsuperscript{36} However, further studies are needed to confirm the role of ClbI KS and S178 in cyclopropane formation on the assembly line.

Two hybrid NRPS/PKS enzymes, ClbB and ClbK, are encoded in the \textit{pks} gene cluster. ClbB is a two-module, hybrid NRPS-PKS encoding C-A-PCP-KS-AT-KR-DH-
ER-ACP domains. In the ClbB cis-AT PKS module, the domain arrangement is unusual, as it is usually encoded in the order of KS-AT-DH-ER-KR-ACP, and the only other known example of this type I PKS modular arrangement is TubD from the tubulysin biosynthetic pathway.\(^{42}\) The A domain of ClbB NRPS module C-A-PCP was shown to incorporate L-alanine and the C-domain connecting N-myristoyl-D-Asn chain to produce metabolite 2 which is modified by the PKS domains of ClbB to produce metabolite 3.\(^{30}\) ClbK contains a trans-AT type-hybrid NRPS/PKS module encoding KS-AT*-ACP-Cy-A-Ox-T. The A domain of ClbK activates and incorporates L-Cys. The Ox domain in ClbK was suggested to oxidize the thiazoline formed by the Cy domain to provide thiazole ring.\(^{37}\) Later reports, based on domain-targeted metabolomics, to establish general timing events in bithiazole formation, has revealed that the oxidation of the first thiazoline introduced by ClbJ is catalyzed by the Ox domain of ClbK, and that the thiazole product is a preferred substrate for further assembly line processing events.\(^{36}\) Earlier studies suggested the module skipping of PKS domains (KS-AT*-ACP) in ClbK to yield the metabolites 17-19.\(^{34,43,44}\) Recent reports suggested that the AM unit generated by accessory enzymes can be incorporated into ClbK products, leading to the production of metabolites 14-16.\(^{45}\) The latter observation is further corroborated by Crawford and co-workers to experimentally support points of divergence in a complex biosynthetic pathway and that the module skipping does not require non-elongating transfer through ClbK PKS domains.\(^{36}\)

The pks cluster also encode a dedicated set of enzymes ClbD, E, F, and G to generate an uncommon AM extender unit which is predicted to be utilized by the trans-AT PKS modules ClbO, ClbC, and ClbK.\(^{32,40}\) The in vitro characterization of these
enzymes by Piel and co-workers suggested that the AM biosynthesis in the colibactin pathway parallels that of the zwittermycin pathway,

which also possesses close homology to ClbD-G and provided the biosynthetic rationale of the 2-AM unit (Figure 1-5). The ClbH A₁ domain activates L-Ser providing Ser-AMP to the phosphopantetheine (Ppant) arm of the standalone ACP enzyme, ClbE. The side chain alcohol of serine is then sequentially oxidized by the dehydrogenases ClbD and ClbF, producing the AM extender unit tethered to ClbE. Later studies by Li Zha and co-workers suggested the transfer of AM-ClbE to the trans-AT enzyme, ClbG, followed by transfer of AM-ClbG to the ACP domains of ClbC, O, and K, further suggesting that the AM unit could be utilized multiple times in the colibactin biosynthetic pathway.

However, in these in vitro characterization studies, inherent instability (perhaps from spontaneous decarboxylation) resulted in a glycyl-ClbE intermediate instead of an AM-ClbE, requiring further experiments to support this observation.

Figure 1-5. Aminomalonyl (AM) synthesis by accessory enzymes.

Additional tailoring enzymes ClbA, Q, L, and P are present and involved in the biosynthesis of colibactin. ClbA is a Ppant-transferase homologous to Sfp from Bacillus subtilis and utilizes CoA to transform the pks enzymes to their corresponding holo or
active form. Deletion of the \textit{clbA} gene was shown to abolish colibactin genotoxicity and the \textit{clbA} mutant has been used in the previous biological studies to generate \textit{pks}⁻ strains as negative controls in assays with wildtype \textit{pks}⁺ strains.\textsuperscript{8} However, a previous report by Oswald and co-workers suggested that in a subset of virulent \textit{E. coli}, the \textit{pks} island is associated with the high pathogenicity island (HPI) that encodes for the biosynthesis of the siderophore yersinibactin.\textsuperscript{48} It was shown that \textit{clbA} can modify enzymes in the yersinibactin biosynthetic pathway suggesting a strong interplay between siderophores and colibactin biosynthesis, thus suggesting a complex regulatory interaction of various virulence features.\textsuperscript{49} The conclusions from earlier \textit{clbA} mutant studies in the probiotic strain \textit{EcN} might have resulted from the impairment of both colibactin and yersinibactin biosynthesis affecting the virulence and require further validation. Two recent reports suggest that the ferric uptake regulator (Fur) and the small regulatory non-coding RNA RyhB (two master regulators of bacterial iron homeostasis) are involved in the regulation of \textit{clbA} transcription and the production of colibactin.\textsuperscript{50,51}

\textbf{ClbQ} is a thioesterase enzyme encoded in the colibactin pathway. Previous bioinformatic analysis suggested \textbf{ClbQ} is homologous to type II thioesterases, a class of enzymes that have been proposed to increase the efficiency of biosynthetic processes by performing a proof-reading or editing role. The role of \textbf{ClbQ} in colibactin biosynthesis is discussed to a greater detail in Chapter 2. \textbf{ClbL} is a predicted aminotransferase enzyme encoded in the \textit{pks} gene island and is one of the two enzymes (other is \textbf{ClbO}) in the gene cluster yet to be experimentally implicated in the colibactin biosynthetic
model. Our efforts to understand the role of ClbL in the colibactin biosynthesis and mechanism are outlined in Chapter 4.

ClbP is a membrane bound peptidase and indeed is the first enzyme whose structural and biochemical investigations revealed the unique prodrug resistance strategy of colibactin biosynthesis. The first insight into the in vivo prodrug resistance strategy was discovered and characterized by Bode and co-workers in 2011 while studying the biosynthesis of the antibiotic xenocoumacin.\textsuperscript{52} The close resemblance of several components of the colibactin and xenocoumacin pathways led to further investigation by Balskus and co-workers in 2012, revealing that colibactin biosynthesis involves a prodrug resistance strategy where an N-terminal prodrug scaffold (precolibactin) is assembled, transported into the periplasm, and finally cleaved to release the mature product.\textsuperscript{30} An in vivo evidence of the prodrug like mechanism of activation of colibactin was also reported by the Müller group.\textsuperscript{53} Following by this discovery, all the isolation studies have used clbP deletion strains (\textit{pks} \textit{\Delta} clbP) to promote the accumulation of precolibactin metabolites. Although this strategy has been helpful to isolate and elucidate several downstream metabolites of the colibactin pathway, recent studies have shown that this strategy has led to the production of non-genotoxic pyridone-based isolates that are derived from the diversion of linear biosynthetic intermediates toward alternative cyclization pathways.\textsuperscript{43,54,55} The structural studies of the periplasmic domain of ClbP (ClbP\textsubscript{pep}) revealed a negative electrostatic potential suggesting that precolibactin is positively charged.\textsuperscript{46} While the reported ClbP\textsubscript{pep} structure possesses catalytic active residues (S95, Y186 and K98), this construct was not able to restore genotoxicity in the complementation studies, contrary
to the full length ClbP, which contained the N-terminal signal sequence and all three transmembrane helices. This strongly warrants further structural and biochemical studies to understand the mode of action of full length ClbP.

In addition to the NRPS/PKS and accessory enzymes, pks cluster encodes for proteins involved in transport (ClbM), resistance (ClbS) and regulation (ClbR). ClbM is a key component involved in colibactin activity and transport. In 2016, Jarrod Moussa, a former graduate student of the Bruner laboratory, solved the X-ray crystal structure of ClbM. His work revealed that ClbM is a transmembrane, cation-coupled antiporter containing structural motif common to the recently identified multidrug and toxic compound extrusion (MATE) transporter family. It was also shown that the disruption of clbM attenuated pks+ E. coli-induced DNA damage and significantly decreased the DNA damage response in gnotobiotic IL10(-/-) mice, suggesting involvement in colibactin transport. In contrast, earlier studies showed that ClbM is not required for pks-associated genotoxicity, suggesting that another transporter encoded in the genome may compensate for ClbM and transport colibactin. Very recently, our group, in collaboration with Crawford and co-workers, reported the structure and function of the colibactin resistance protein ClbS. It was shown that ClbS confers self-resistance through cyclopropane hydrolase activity. The cyclopropane ring in the reported precolibactin metabolites has been shown to be essential for genotoxic effects in vitro, and the ClbS-catalyzed ring-opening activity suggested a means for the bacteria to circumvent self-induced toxicity. ClbR is a LuxR-like transcriptional regulator protein which is suspected to be a regulator of colibactin biosynthesis. LuxR family proteins are key players in quorum sensing and coordinates the expression of a variety of genes,
including those encoding virulence factors and antibiotics biosynthesis, motility, nodulation, plasmid transfer, bioluminescence, and biofilm formation. Recent studies have shown that ClbR is not involved in the iron-dependent regulation of \textit{clbA}. However, the exact role of ClbR in colibactin biosynthesis and genotoxic activity remains to be elucidated.

1.4 Isolation, Synthesis and Structural Characterization of (Pre)Colibactins

The discovery and characterization of the periplasmic peptidase, ClbP, provided a foundation for all subsequent metabolite isolation efforts. As mentioned earlier, the isolation studies conducted by multiple labs have employed \textit{clbP} deletion strains (\textit{pks}\textasteriskcentered clbP, \textit{pks}\textasteriskcentered clbP\textasteriskcentered clbG, or \textit{pks}\textasteriskcentered clbP\textasteriskcentered clbQ) to promote accumulation of more stable precolibactin metabolites within the cells, allowing isolation of metabolites from bacterial extracts at higher yields. The reported precolibactin metabolites and proposed precolibactin biosynthetic pathway, representing current state of knowledge, is outlined in the Figure 1-4.

![Figure 1-6. Proposed mechanistic model for the production of precolibactin metabolites.](image)

The two most structurally complex known precolibactins structures are precolibactin C (19) and precolibactin-886 (16). However, the DNA damaging activity of these metabolites was reported to be modest.
suggested that pyridone cyclization in the metabolites isolated from $\textit{pks} \Delta \textit{clbP} \textit{E. coli}$ is an artifact of isolation procedures.$^{43,44}$ Very recently, Herzon and co-workers provided a molecular level explanation for the metabolites produced in the presence and absence of ClbP by using synthetic standards.$^{55}$ Together with the previous reports from Crawford lab,$^{54}$ it was shown that wild-type $\textit{pks}^+ \textit{E. coli}$ produce electrophilic unsaturated imines after ClbP deacylation of corresponding linear precursors. It was also shown that deletion of ClbP, leading to the persistence of the terminal $N$-myristoyl-D-Asn side chain diverts the intermediates toward off-pathway pyridone derivatives resulting in non-genotoxicity, whereas the unsaturated imines produced in the presence of ClbP are potent DNA-damaging agents (Figure 1-6).$^{55}$

1.5 Summary

Almost twelve years have elapsed since the identification of the $\textit{pks}$ island and extensive attempts were made in multiple laboratories to identify underlying pathway of colibactin and its mechanism. While tremendous progress has been made in the past three years resulting in a decent understanding of the one of the nature’s most complicated biosynthetic pathways, several challenging issues related to the $\textit{pks}$ cluster and its potentially causal link to CRC remain unresolved. The elusiveness of the colibactin structure and mechanism can be attributed to several factors.

First, the isolation efforts are impeded by the extremely low yield of colibactin produced by wildtype strains or $\textit{clbP}$ deletion strains. For example, only 2.8 mg of precolibactin-886 from a 1000L fermentation$^{45}$ and 0.5 mg of precolibactin C from 48L fermentation$^{40}$ of mutant strains were isolated, suggesting that chemical synthesis of these complex derivatives may be the only feasible way to explore the mechanism
hypotheses. Second, the structural characterization is thwarted by the intrinsic instability of colibactin, which is evident from various isolation and synthetic studies where spontaneous cyclization and cyclodehydration of metabolites were reported.\textsuperscript{43,54} Third, the role of ClbL and ClbO remains unidentified and although precolibactin-886 is the most complex known precolibactin metabolite, its structure does not account for all the enzymes encoded in the pathway. In addition, while precolibactin-886 includes an AM unit produced by accessory enzymes ClbD-G which is shown to be indispensable for colibactin cytopathic effects, it was shown that ClbG can transfer AM unit to several PKS enzymes in the cluster.\textsuperscript{40} Therefore, further studies are needed to draw firm conclusions and address whether final colibactin structure contains one or several AM units and how it influences the colibactin activity.

Fourth, the identification is limited by the contact-dependent cytopathic activity of \textit{pks} island, suggesting the compounds do not passively diffuse across the cell membrane. This further suggests the observed probiotic/genotoxic effects of colibactin might dependent on the mechanism of transport to eukaryotic cell and to their nuclei. While there is a possibility that intestinal mucosal colonization of \textit{pks}\textsuperscript{+} \textit{E. coli} and the reduced mucosa in patients with intestinal inflammation allows greater cell-to-cell contact to promote carcinogenesis, such cell-to-cell contact may not be required for probiotic effects in EcN, where outer membrane vesicles undergo clathrin-mediated endocytosis.\textsuperscript{59} Further investigations are required to address these hypotheses. An alternate hypothesis for probiotic/genotoxic effects of colibactin is that the rich chemistry encoded in the colibactin biosynthetic pathway could enable the production of different metabolites or undergo different modes of cyclization to produce compounds with
various activities. Very recently, Cenac and co-workers identified an analgesic liopopetide, \( C_{12}\text{AsnGABA}OH \), produced by the probiotic EcN strain.\(^6\) The production of this molecule is shown to be dependent on three \( pks \) genes, \( clbANB \), and the addition of \( C_{12}\text{AsnOH} \) enhances GABA diffusion across the epithelial barrier and subsequently act on sensory neurons to exhibit analgesic properties. This work not only illustrated the colibactin’s probiotic/carcinogenic paradox, but also highlighted how the colibactin NRPS-PKS pathway represents a rich source of unusual assembly-line synthesis, encoding molecules with different properties. Further studies need to be focused on the regulation and mechanism of release of multiple metabolites from the biosynthetic pathway contributing to varying bioactivity.

In conclusion, establishing casual links between microbial secondary metabolites and human health or disease states is the greatest challenge of human microbiome research. Nevertheless, progress made in deciphering the molecular basis of colibactin biosynthesis and mechanism of activity is a cohesive effort from multiple research labs integrating both advanced multi-omics (metagenomics, metabolomics, metatranscriptomics) and traditional isolation, synthetic chemistry, and structural biology tools. Our structural biology approach to study the mechanism of key biosynthetic enzymes in the colibactin pathway is discussed in Chapters 2-4.
CHAPTER 2
STRUCTURE AND FUNCTIONAL ANALYSIS OF CLBQ, AN UNUSUAL INTERMEDIATE-RELEASING THIOESTERASE FROM THE COLIBACTIN BIOSYNTHETIC PATHWAY

2.1 Thioesterase Enzymes of NRPS/PKS Systems

TEs are diverse class of enzymes belonging to α/β hydrolase family encoded in the fatty acid synthases (FASs), PKSs and NRPSs and clade into 25 different families based on sequence alignments in the ThYme database. In the modular NRPS/PKS assembly lines, pathway intermediates that are tethered via a thioester bond, to the Ppant arms of carrier protein (ACP/PCP) domains throughout the chain assembly are often cleaved by TEs. TEs in NRPS/PKS are mainly classified into type I (TEI) and type II (TEII) as shown in Figure 2-1.

Figure 2-1. Role of Type I and Type II thioesterase enzymes in NRPS/PKS.

TEs are usually encoded in the C-terminus of the megasynthases. Recent review articles highlight current knowledge on the diversity, structure and mechanisms of various TEs in NRPS/PKS pathways as well as role of TEs as potential arbitrators in offloading. Unlike TEIs that are covalently attached to the terminal module in the multienzyme complex and remove only the final product, TEIIs are not covalently linked and can remove intermediates from any module in the megasynthetase complex. Some NRPS/PKS and hybrid system possesses encode more than one TEs, each with a varying specificity to the pathway intermediates and products. Many TE sequences have been reported in the literature, yet the prediction of TE selectivity and chemistry substantially lags behind other NRPS/PKS domains.

TEIIs have a corrective (editing or proof-reading) role (Figure 2-I), removing undesirable substrates and aberrant intermediates interrupting the biosynthetic pathway. Besides the editing role, they also participate in selection of starter units incorporated in to the pathway, release of intermediates or final products, regulate the yield of products and influence the overall performance of the megasynthetases. Deletion of TEIIs in NRPS/PKS biosynthetic pathways commonly result in decrease in the yield of final product. Two models have been reported to account of the activity of TEIIs, a high specificity model where TEIIs hydrolyze only aberrant units and a low specificity model where the hydrolases act on both ‘correct’ and ‘incorrect’ intermediates with the ‘correct’ released at a slower rate. Deletion of TEIIs in NRPS/PKS biosynthetic pathways have shown dramatic decrease in the yield of final product. Although a large number of TEIIs have been identified, only four TEIIs have been structurally characterized: RifR from the rifamycin hybrid NRPS/PKS pathway, RedJ
from the prodiginine hybrid NRPS/PKS,\textsuperscript{70} SrfD from the surfactin NRPS,\textsuperscript{72} and the human TEII involved in fatty acid synthesis.\textsuperscript{73} Collectively these structures have revealed that unique structural conformations of TEIIs enable selective interaction with individual enzyme domains within the multidomain systems. Also, a highly flexible lid region in each structure controls the size and shape of the active site, regulating the enzyme-based selectivity for various substrates. Therefore, determining the substrate selectivity based on sequence analysis alone is not yet feasible. Additional structural information will more broadly aid in defining the selectivities and roles of TEs in a specific NRPS/PKS pathway.

2.2 ClbQ is a TEII Enzyme in the Colibactin Biosynthetic Pathway
Colibactin gene island encodes a hybrid NRPS-PKS pathway and several intermediate precolibactin metabolites are reported.\textsuperscript{31,34,36,45} Insights gained from prior isolation studies, along with gene deletion experiments, suggest that the varied biological activities (pathogenic and probiotic) reported for colibactin might result from a mixture of compounds produced through the hybrid NRPS-PKS pathway.\textsuperscript{32,74} Therefore, understanding the off-loading mechanism of such biosynthetic intermediates is essential to gain insights into the biosynthesis and biological activities of pathway metabolites. However, the colibactin pathway unusually does not contain a terminal, releasing domain as part of the assembly-line. One candidate TE in the \textit{pks} gene cluster is the standalone gene product, ClbQ. Sequence homology analysis suggests that ClbQ belongs to the family of editing TEIIs (Figure 2-2).\textsuperscript{45}

Prior studies have demonstrated the importance of \textit{clbQ} in the colibactin pathway. In 2014, two independent studies conducted by Bringer and Bonnet labs have showed that deletion of \textit{clbQ} in \textit{E. coli} abolish the cytotoxic effects of \textit{clb}\textsuperscript{+} \textit{E. coli}.\textsuperscript{17,75}
2016, a co-authored publication by Qian and Moore labs has suggested a unique function of ClbQ in off-loading pathway intermediates rather than removing aberrant intermediates.\(^{45}\) In this study, *in vitro* assays on the substrate specificity of ClbQ against precolibactin N-acetyl cysteamine (SNAC) thioester derivatives (Figure 2-3) revealed that the TE readily hydrolyzed the early stage intermediate derivatives 21 and 22 to generate the corresponding precolibactins 2 and 3, respectively. However, hydrolysis of cyclized precolibactin derivatives 23, 24 or 25 was not observed.

**Figure 2-2.** Sequence homology analysis of ClbQ.

In addition, characterization of a clbPQ double mutant of *clb+ E. coli* allowed the identification of the late-stage precolibactin candidate 16, containing a 2-aminomalonyl unit (AM) (pathway A, Figure 2-3b). However, in these studies, a clbP single mutant directed the synthesis to pathway B to accumulate compound 19 (pathway B, Figure 2-3b).\(^{45}\) Overall, these studies suggest that ClbQ plays a key role in directing or controlling the flux of colibactin products and thus regulating biological activities of the colibactin pathway in host-microbe interactions. However, the structural and molecular basis of the unique TE activity of ClbQ remains unclear. Therefore, we carried further
investigation on each of the reported biochemical observations to provide more insights in to role of ClbQ in the colibactin biosynthesis and mechanism through structural and biochemical studies.

Figure 2-3. Effect of ClbQ on clb pathway. A) Reported specificity of ClbQ against precolibactin- SNAC- thioester derivatives. B) Overview of the effect of ClbQ on production of precolibactin metabolites.

2.3 Biochemical Characterization of ClbQ

For biochemical characterization of ClbQ, we investigated the role of ClbQ in the reported accumulation of AM containing compound 16 in the ΔclbPQ strain. In addition, we also conducted additional biochemical studies to understand the substrate specificity of ClbQ against early stage and late stage precolibactin intermediates.
2.3.1 ClbQ-Mediated Release of AM Unit from the Colibactin Biosynthetic Pathway

The hypothesized scheme for the role of ClbQ in the accumulation of compound 16 in a ΔclbPQ strain is shown in Figure 2-4. In this model, we hypothesized ClbQ catalyzed off-loading on the AM unit from the assembly line directing to pathway B providing metabolite 19, whereas in the absence of ClbQ in ΔclbPQ strain, AM is transferred to ClbK module via ClbG and leads to production of metabolite 16.

Figure 2-4. Proposed hypothesis on the role of ClbQ off-loading AM unit.

To test ClbQ-mediated release of AM unit, the proteins ClbD, ClbE with C-terminal hexa-histidine tag, ClbF, ClbQ, ClbQS78A, Sfp (Ppant Transferase), are cloned, expressed and purified (Figure 2-5). The in vitro reconstitution of AM-ClbE was generated using previously reported protocol\textsuperscript{40} and was tested against ClbQ and its active site mutant ClbQS78A. Consistent with our hypothesis, ClbQ hydrolyzed seryl-
ClbE and AM-ClbE (observed as glycy1-ClbE) suggesting ClbQ mediated off-loading of AM unit in the colibactin pathway (Figure 2-6).

Figure 2-5. SDS-PAGE analysis of proteins used in the *in vitro* assays.

Figure 2-6. Hydrolysis of AM-ClbE with the ClbQ. [M+H]+ detected for Apo-ClbE, 10539.6 (I); Holo-ClbE, 10880.7 (II); Seryl-ClbE, 10967.1 (III); AM-ClbE (detected as Glycyl ClbE) 10936.6 (VI). Panel IV and VII display the respective hydrolysis of Seryl-ClbE and AM-ClbE (glycyl ClbE) to the Holo-ClbE catalyzed by ClbQ and not by ClbQS78A mutant (Panel V and VIII).
2.3.2 Substrate Specificity of ClbQ

As introduced in section 2.2, previously reported \textit{in vitro} studies on the substrate specificity of ClbQ against precolibactin \textit{N}-acetyl cysteamine (SNAC) thioester derivatives (Figure 2-2) revealed that the TE readily hydrolyzed the early stage intermediate derivatives 21 and 22 to generate the corresponding precolibactins 2 and 3, respectively.\(^{45}\) However, hydrolysis of the late-stage, cyclized precolibactin derivatives 23, 24 or 25 was not observed. To note, this study did not include linear late-stage linear intermediates, suggesting one or a combination of the following possibilities. First, the cyclized compounds (pyridone-containing) 23, 24, and 25 are not the actual substrates for ClbQ, and cyclization is an artifact of the \textit{in vitro} isolation procedures. Second, ClbQ discriminates alternate substrates based on chain length, as has been observed for other TEIIs.\(^{70}\) Third, the substrate specificity is a result of ClbQ’s selectivity toward various ACP/PCP domains in the pathway. To address these possibilities, we conducted studies using both wild-type ClbQ and a S78A active site mutant, in which the catalytic serine residue was substituted with alanine.

To test the first possibility, we collaborated with Crawford lab at Yale university who provided us compounds 26 and 27 (SNAC derivatives of 18 and 19) as well as compound 20 (SNAC derivative of 1) for our assay studies. We tested the thioester hydrolysis activities in the presence of ClbQ for all these derivatives (Figure 2-7). Using an LC-MS based assay, we observed hydrolysis of compound 20 by wild type ClbQ, but we did not observe hydrolysis of 26 or 27 (30 min incubation). Given the poor solubility of late stage intermediates in the assay buffers, we repeated the assays for 18 h and observed slow hydrolysis of linear substrate 26 but no hydrolysis of the cyclized substrate 27 (Figure 2-7). To confirm that thioester hydrolysis of 26 is specifically
catalyzed by ClbQ, the enzymatic reaction was compared with an active site ClbQ mutant and a buffer control.

![Image]

**Figure 2-7. Hydrolysis of precolibactin thioester SNAC derivatives with ClbQ.** A) LC-MS based assay showing hydrolysis of a linear precolibactin SNAC thioester (26) catalyzed by WT ClbQ and partial thioester hydrolysis by the S78A mutant. B) Extracted ion chromatograms (EIC) EIC+ = 467.30 ± 0.01 (top) and 398.33 ± 0.01 (bottom), corresponding to 26 and the hydrolyzed and cyclized product precolibactin 19, respectively. C) LC-MS extracted ion chromatogram (EIC) traces displaying the hydrolysis of compound 20 catalyzed by recombinant ClbQ and not by ClbQ S78A (middle). EIC+ = 444.28 ± 0.01 and 343.28 ± 0.01, corresponding to SNAC thioester 20 and the hydrolyzed product precolibactin 1, respectively. D) LC-MS extracted ion chromatogram (EIC) traces displaying no hydrolysis of cyclized precolibactin 27. EIC+ = 449.28 ± 0.01 [M+H]+2, corresponding to SNAC thioester 27.

The LC-MS analysis data showed that the linear substrate 26 was weakly hydrolyzed by wild-type ClbQ, and the hydrolysis product 18 cyclized over time to compound 19. Spontaneous cyclodehydration of 18 to 19 was experimentally demonstrated in earlier studies.\(^\text{43}\) Partial hydrolysis of 26 at a decreased rate was
observed in the samples incubated with the ClbQ S78A mutant (approximately two-fold less compared to WT), suggesting that the substrate 26 may be capable of binding in the active site. In place of an active site serine residue, bound water can act as a nucleophile, allowing the hydrolysis of 26 albeit at a slower rate (Figure 2-7). Similar observations have been reported for other TEII active-site serine mutants.\textsuperscript{71} The ClbQ S78A mutant showed only minor or no hydrolysis of SNAC derivatives 20, 26, and 27 (Figure 2-7).

Our results support that ClbQ is a low specificity TEII enzyme with a stronger preference for early-stage intermediates. In addition, our observations of ClbQ hydrolysis of late-stage linear precolibactin intermediates support our hypothesis that late-stage linear precolibactin intermediates, and not the stable cyclized precolibactin intermediates, are indeed colibactin biosynthetic intermediates.

Additionally, we synthesized various acyl-SNACs (acetyl, lauroyl, myristoyl) to test the ability of ClbQ to hydrolyze common acyl-SNACs of varying chain lengths and to understand whether ClbQ performs a generic editing function (common to TEIIs) or is specific to early-stage precolibactin metabolites. Using the previously reported DTNB-based spectrophotometric assay,\textsuperscript{70} the activity of ClbQ against acetyl-, lauroyl-, and myristoyl- SNACs, as well as precolibactin SNACs 20, 26, 27, were compared. The assays revealed that ClbQ has thioesterase activity with model straight-chain acyl-SNACs, and the rates of hydrolysis were greater than that measured for precolibactin-SNACs. However, the slow hydrolysis rate of lauryl-SNAC, and precolibactin-SNAC thioesters 20 and 26 compared to acetyl-SNAC, indicate that these are relatively poorer substrates for ClbQ. The catalytic efficiencies ($k_{cat}/K_m$) of acetyl-, lauroyl-, myristoyl- 20-, 26-,
and 26-SNACs are 27.3, 7.3, 38.8, 5.2, and 1.6 M$^{-1}$ sec$^{-1}$, respectively. Additionally, we did not observe hydrolysis of cyclized precolibactin SNAC derivative 27 as anticipated from our LC-MS data and previous reports, suggesting that these compounds are not the optimal substrates.

2.3.3 Selectivity of ClbQ toward Carrier Proteins in Colibactin Pathway

ClbQ hydrolysis of specific precolibactin SNAC derivatives posed the question as to whether the thioesterase is selective toward carrier protein domains in the colibactin pathway. ClbQ shares high structural homology to RedJ, a type II TE that was shown to have specificity for ACPs in the prodiginine pathway. Therefore, we investigated the possibility of ClbQ discrimination between ACP and PCP domains. There are a total of twelve carrier protein domains (5-ACPs/7-PCPs) in the colibactin assembly line; all of these domains are embedded in multi-module enzymes, except for the stand alone, ClbE (Figure 1-3). Our initial strategy was to utilize the standalone ClbE carrier protein in acyl-ACP assays. However, our efforts to load various acyl substrates on ClbE (with N-terminal hexa-histidine tag) were unsuccessful (data not shown) with ClbE only accepting acetyl-CoA and not lauroyl- or myristoyl-CoA using the Ppant transferase (PPTase), Sfp. This is an interesting observation given that ClbE has a unique role in the colibactin pathway acting as an in trans carrier protein to the first ClbH adenylation domain, which is further processed to a 2-aminomalonyl unit (AM) by accessory enzymes ClbDEF. Accepting only short-chain acyl-CoA substrates but not longer ones suggests that ClbE unusually displays strict specificity toward small amino acids like L-Ser. Mass analysis by MALDI-TOF (Matrix-assisted laser desorption/ ionization-Time of Flight) did show that ClbQ was able to hydrolyze acetyl-ClbE consistent with its editing role (Figure 2-8).
Unfortunately, attempts to solubly express the colibactin gene cluster specific ACP/PCPs of ClbB, ClbK, and ClbO as discrete, standalone domains were unsuccessful under the conditions of our experiments (data not shown). Given the difficulty in soluble expression of individual in cis ACP/PCPs from the colibactin gene cluster, we chose an alternative approach. We tested the ability of ClbQ to recognize and hydrolyze acyl-CoA tethered to a model standalone carrier domain, FscF, a well behaved carrier protein from the fuscachelin NRPS biosynthetic pathway. We tested the hydrolysis of various acyl-FscFs by ClbQ and observed all conjugates were substrates (Figure 2-8). This suggests that ClbQ may lack selectivity toward specific
colibactin synth( et) ase carrier protein domains, implying that the differences in the hydrolysis of precolibactin SNAC derivatives is due primarily to the substrate chain-length specificity of ClbQ.

Figure 2-9. Sequence alignment of carrier protein domains in the colibactin biosynthetic pathway. Predicted structure of ClbE by is shown at the bottom of the sequence.

To further substantiate these results, we conducted a sequence homology analysis of the twelve carrier protein domains in the colibactin pathway to examine differences in the regions that have previously been reported to be essential for TE binding (Figure 2-9). Sequence alignment was generated with MUSCLE and ESPript 3.02 was used to generate the secondary structure alignment using EntF (PDB code
3TEJ) as a PCP model. Consistent with our assay results, no significant differences between the ACP/PCP domains in the TE binding region were apparent. However, consistent with its specialized role, ClbE showed the most differences as compared to the other carrier protein domains.

We further carried out secondary structure predictions of all twelve carrier protein domains using JPred4.3. Predicted structure of ClbE is manually depicted at the bottom in Figure 2-9. In general, PCP/ACP domains adopt the expected four-helix structure, in which residues in $\alpha_3$ and $\alpha_4$ helices interact with partner domains. Our prediction showed that all carrier protein domains in the colibactin pathway similarly possess the canonical four-helix structure, except ClbE, in which $\alpha_3$ is interestingly replaced with a beta strand (Figure 2-9). We speculate that this atypical change in overall predicted structural topology likely plays a critical role in ClbE’s unique activity with ClbH-A$_1$ recognition and AM production.

2.4 Crystallization, Structure Determination of ClbQ

To gain better understanding of the role of ClbQ in colibactin biosynthesis, we performed protein crystallization studies with a hope to understand unique structural features of ClbQ off-loading various intermediates from the colibactin biosynthetic pathway.

2.4.1 Expression, Purification and Crystallization

ClbQ protein was expressed, purified and subjected to various crystallization screens. ClbQ was cloned from a BAC containing the pks island and was ligated into a pET28a vector. The expression vector pET28 ClbQ encoding the N-terminal hexa histidine tag was transformed into E. coli BL21 (DE3) for overexpression. After testing
varying growth and expression conditions, induction at 18°C provided larger quantities of stable soluble protein.

ClbQ was purified by Ni-NTA resin, followed by Hi-TrapQ, and then size exclusion chromatography. Crystal screening was conducted in 96-well sitting drop plates using commercial screens: Wizard Classic I and II, Hampton Peg/ion HT screen, Hampton Crystal Screen HT, JCSG Core I Suite, JCSG Core II Suite. Initial crystal hits were found in one of the two conditions (20% PEG 6000, 0.1 M Bicine pH 8.5) was further optimized using hanging drop and macro seeding techniques provided single diffraction quality crystals (Figure 2-10). However, these crystals exhibited severe twinning (twin law= h, -k, -h-l) allowing an incomplete structure with electron density missing for amino acid residues 110-148 and 183-200, this including the active site aspartate, D186 (data not shown).

Figure 2-10. Crystallization of ClbQ.

Co-crystallization with synthetic precolibactin substrates produced several crystal hits in various conditions and with significantly less twinning, suggesting ClbQ is more structured in the presence of ligands. Several data sets were collected (data not shown) to obtain protein-ligand complexes from crystals soaked or co-crystallized with precolibactin derivatives (9, 18, 19, 20, 26, 27). Data for precolibactin-ClbQ complexes
were collected to high resolution, but no clear density for the bound ligand was apparent. However, the density for the disordered loops and missing regions from the native crystals was observed in the co-crystals, allowing a complete model (residues 2-240) of ClbQ, suggesting that transient ligand binding aids in crystal growth and formation.

2.4.2 Crystal Structure of ClbQ

ClbQ was crystallized in the space group P2₁ with cell dimensions of a= 45.0 Å, b = 95.0 Å, c = 68.4 Å, β = 109.2°. The overall structure contains two independent monomers in the asymmetric unit (Figure 2-11) and was isotropically refined to a resolution of 1.98 Å with a final a R<sub>free</sub> of 21% (Table 2-1). The ClbQ structure reveals, as expected, a canonical α/β hydrolase core domain (Ser2-Pro110 and Ile165-His240), and a cap domain (Asp111-Thr164) (Figure 2-11A). Five α helices of the core domain surround a six-stranded parallel β-sheet, and a flexible cap domain is composed of helices α'4-α'5, lid loop 1 and lid loop 2 inserted between β4 and β5 of the core. The active site of ClbQ shows a classic α/β hydrolase signature sequence Gly76-His77-Ser78-Xaa79-Gly80 of the “nucleophilic elbow” between strands β4 and the α3 helix. The catalytic triad of ClbQ comprising residues Ser78, Asp186, His215 (Figure 2-11B), are on loop regions, following strands β3(Ser), β5(Asp), and β6(His) of the α/β-hydrolase fold.

The catalytic Ser78 in both monomers of ClbQ is a conformational outlier based on Ramachandran analysis, as commonly observed in the structures of α/β hydrolase fold enzymes. In addition, several significant changes in the residues adjacent to the catalytic site as well as hydrogen bond positioning of residues in the catalytic region
were observed in the ClbQ structure. His77 of the signature motif forms a hydrogen bond with the backbone carbonyl of the catalytic His215, stabilizing its alignment within the triad (Figure 2-11B).

Figure 2-11. Crystal structure of ClbQ. A) The asymmetric unit with two monomers of ClbQ, βME is shown as orange sticks. B) The active site of ClbQ. The catalytic triad Ser78, Asp 186, His 215 are shown as gold sticks, the oxyanion hole (Ser12 and Leu79) and surrounding residues active site are colored salmon. C) Superimposition of the chain A and chain B monomers. The core regions of both domains (cyan) is constant and the flexible lid domains are represented (chain A, green and chain B, orange). Sidechains of the residues on the flexible lid domain are labelled in the expanded view and the movement/orientations between the two monomers are depicted with blue arrows. The catalytic triad is shown as pink sticks.
<table>
<thead>
<tr>
<th>Crystallographic data of ClbQ.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
</tr>
<tr>
<td>X-ray source</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
</tr>
<tr>
<td>Space group</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
</tr>
<tr>
<td>α, β, γ (deg)</td>
</tr>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>Total reflections</td>
</tr>
<tr>
<td>Unique reflections</td>
</tr>
<tr>
<td>Completeness %</td>
</tr>
<tr>
<td>Multiplicity</td>
</tr>
<tr>
<td>&lt;I/σ(I)&gt;</td>
</tr>
<tr>
<td>Rmerge (%)</td>
</tr>
<tr>
<td>Rmeas (%)</td>
</tr>
<tr>
<td>CC1/2</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
</tr>
<tr>
<td>Rwork / Rfree (%)</td>
</tr>
<tr>
<td>Non-H atoms</td>
</tr>
<tr>
<td>protein atoms</td>
</tr>
<tr>
<td>ligand atoms</td>
</tr>
<tr>
<td>water atoms</td>
</tr>
<tr>
<td>RMS deviations</td>
</tr>
<tr>
<td>bonds (Å)</td>
</tr>
<tr>
<td>angles (deg)</td>
</tr>
<tr>
<td>Ramachandran</td>
</tr>
<tr>
<td>favored (%)</td>
</tr>
<tr>
<td>outliers (%)</td>
</tr>
<tr>
<td>Average B-factors (Å²)</td>
</tr>
<tr>
<td>protein</td>
</tr>
<tr>
<td>water</td>
</tr>
<tr>
<td>ligands</td>
</tr>
<tr>
<td>PDB ID</td>
</tr>
</tbody>
</table>

*Values for the outer shell are given in parentheses

The variable amino acid Xaa79 of the signature sequence is a methionine in most TEIIIs. However, in ClbQ it is a leucine. The backbone amides of Leu79 and Ser12 (Met95 and Ala29 respectively in RifR)\(^{71}\) are reported to be essential in stabilizing the
tetrahedral intermediate in the oxyanion hole, which is occupied by a water molecule in our structure.

The catalytic histidine (His215 in ClbQ) in most TEII s is present in the motif: GXHF (where X= G/N/D). In ClbQ, a significant change is observed in this signature motif containing AADHF. An additional alanine residue results in a longer loop region (indicated by purple arrow in the Figure 2-12A) and presumably contributing to greater flexibility of the catalytic histidine in the active site.

We also observed clear electron density for a $\beta$-mercaptoethanol ($\beta$ME) molecule forming a disulfide bond with Cys49 of each monomer. The $\beta$ME was added to the dialysis buffer and gel filtration buffer during the protein purification process. The $\beta$ME molecule attached to Cys49 of chain B is located at the dimer interface and makes extensive contacts with chain A (Figure 2-13A), while the $\beta$ME molecule attached to chain A points away from chain B (Figure 2-13B).

As compared with other TEII s, significant changes around the Cys49 region was observed in ClbQ. Additional differences between ClbQ and TEII s include position 157. While this position is conserved as Ala or Ser in other TEII s, it is an Asn in ClbQ with the side chain interacting with the backbone amide of Cys49, and in turn, the amide NH forms a hydrogen bond with the carbonyl of Leu50. This arrangement could possibly stabilize the structural conformation of the loop region between $\beta$2 and $\alpha$2 (Figure 2-13B). A cysteine at position 49 is not present in classic TEII enzymes, and structural changes around this region are unique to ClbQ.
Figure 2-12. Conformational flexibility of ClbQ. A) Alignment of ClbQ (chain A, cyan) with RedJ (salmon, PDB code 3QMV, chain B) and RifR (grey, PDB code 3FLB, chain A). The extended loop region in ClbQ formed by the motif PAADH is indicated (purple arrow). The catalytic triad is labeled. B) Modelling of a Ppant arm (green) into ClbQ active site based on the EntF PCP (light pink)-TE (wheat) didomain structure (PDB code 3TEJ) with ClbQ (chain A, blue/cyan). C) Surface diagram of the two monomers of ClbQ. The substrate binding site is indicated by a red circle and Ppant entrance is indicated by red arrow. D) Cartoon and surface representative of the active site entrance channel. The narrow tunnel-like entrance channel of chain A (circled and open) and chain B (closed, largely by His115 and Glu116).
2.4.3 Conformational Changes of ClbQ

The two monomers in the asymmetric unit are highly similar in structure (RMSD\textsubscript{all atoms} = 1.5 Å), with observed conformational changes around the cap domain (RMSD\textsubscript{111-164} = 2.9 Å) (Figure 2-12C), suggesting flexibility of the lid region. Similar to other TEIIs, lid loop 1 is associated with the size and shape of the substrate chamber, whereas lid loop 2 contributes to the Ppant entrance tunnel. In chain A, lid loop 1 is positioned toward the active site and the lid loop 2 is in an open conformation with respect to the chain B structure, where lid loop 1 is positioned away from the active site and lid loop 2 is in a closed conformation (Figure 2-12C).

In order to model a structural basis for a ClbQ-carrier protein domain interaction, the structure of EntF PCP-TE (PDB code 3TEJ) was aligned to ClbQ providing a
proposed protein/protein interaction surface along with a basis for the positioning of the Ppant arm into the TE active site (Figure 2-12B). We previously reported the structure of the PCP-TE didomain from the *E. coli* EntF NRPS with a conjugated Ppant based inhibitor providing a structural basis for TE/PCP interactions. As common with carrier protein domains, the PCP is a four-helical domain with the conserved Ser covalently linked to the Ppant arm. Structural alignment suggests the position of Ser78 of ClbQ is in a catalytically relevant form and provides insights for the observed structural differences between the two monomers. The alignment and interaction of the two ClbQ monomers is similar to the interaction of the two domains in the EntF PCP-TE structure. Contact residues of ClbQ suggest that chain B acts as a PCP mimic to chain A, leading to the movement of flexible lid loop 2 to an open conformation. The backbone amides of residues Tyr20 and Lys21 of the $\alpha_1$ helix of chain A interact with the $\beta$ME molecule attached to the Cys49 of chain B, while the residues of $\alpha_5$ and loop region between $\beta_2$ and $\alpha_2$ of chain A make contacts with the residues of $\alpha_2$ and $\alpha_3$ of chain B (Figure 2-13A).

Movement of lid loop 1 affects the size and shape of the substrate chamber in both monomers. The sidechains of residues Ala113 and Glu116 of lid loop 1 are pointed inwards to the active site, forming a narrow and longer substrate chamber in chain A. In contrast, these residues are in an opposite configuration forming a wider, short substrate chamber in chain B (Figure 2-12C and Figure 2-12C). The Ppant entrance site and general access to the catalytic triad were affected by the movement of flexible lid loop 2 (Figure 2-12D). The sidechains of residues Val139 and Asp140 are pointed to the active site covering the Ppant entrance in chain B, whereas these residues are
pointed outward opening the Ppant entrance and forming a narrow tunnel-like entrance in chain A. This active site entrance tunnel is blocked by the sidechains of residues His115 and G116 in chain B (Figure 2-12D).

2.4.4 Structural Homology of ClbQ

![Diagram of ClbQ structure]

Figure 2-14. Structural homology analysis of ClbQ. Structure-based sequence alignment of ClbQ with RedJ (PDB code 3QMV), RifR (PDB code 3FLB), Human TEII (4XJV). Catalytic residues are designated with an asterisk.

A structural homology search using the DALI server reveals that ClbQ shares the highest structural similarity (Z-score of 29.9) with RedJ (PDB code 3QMV). The RMSD is 1.7 Å for 227 aligned αC atoms between the N-terminal domain of RedJ and
ClbQ. ClbQ also shares high structural similarity with RifR (Z-score of 27.0, PDB code 3FLB, RMSD of 2.3 Å for 223 aligned αC atoms). Interestingly, ClbQ also shares significant structural homology to the recently published crystal structures of human TEII (Z score of 22.4, PDB code 4XJV) (Figure 2-14). ClbQ has equal similarities with RedJ and RifR in the core region; however, the flexible lid region of chain A in ClbQ is more similar to RedJ than RifR (Figure 2-12A).

In addition, from our crystal structure the putative substrate binding pocket of ClbQ reveals a general negatively charged surface (Figure 2-15) similar to the previously reported ClbP structure, providing additional evidence that fully mature (pre)colibactins are positively charged due to the use of AM substrates. Our findings on structural and biochemical characterization of ClbQ are reported in the cited publication.

Figure 2-15. Electrostatic potential map of the active site of ClbQ (chain A). Negative potential is depicted by red and positive potential by blue. Negatively charged substrate binding pocket and active site suggest the ligands, fully mature precolibactins, could potentially be positively charged compounds due to use of amino malonate substrates.
2.5 Conclusions and Future Work

This work has detailed the biochemical characterization and structure of the thioesterase ClbQ. We highlighted the role of ClbQ off-loading the AM unit from the assembly line through our *in vitro* biochemical studies. Our results provide the explanation for the observed accumulation of AM-containing compound 16 in a \( \Delta clbP \Delta clbQ \) *pks* mutant. As a slow acting TEII enzyme, ClbQ was shown to process late-stage linear intermediates in the colibactin NRPS/PKS pathway, suggesting that increased production of 16 is dependent on the lifetime of a bound AM unit. In addition, we provided additional evidence that the inactive pyridone derivatives are not the actual substrates for ClbQ and are likely produced by a facile double cyclodehydration route during the fermentation and isolation process. From the crystal structure, we highlighted unique structural features of ClbQ contributing to the pathway intermediate off-loading activity.

While we continue our efforts to co-crystalize ClbQ with reported precolibactins, our future work is focused on understanding the factors regulating *clbQ* gene expression and how it contributes to the production of metabolites with varying biological activities.

2.6 Experimental Procedures

Unless otherwise noted, all chemicals and general reagents were purchased from Sigma-Aldrich or Fischer Scientific.

2.6.1 Expression Plasmids

The genes encoding the proteins ClbQ, ClbE, ClbH, ClbD, and ClbF were amplified by PCR from a bacterial artificial chromosome (BAC) harboring the *pks* genomic DNA\(^8\) using the primers listed in Table 2-2. Amplified fragments were digested
with appropriate restriction enzymes and were purified using agarose gel electrophoresis. The digests were ligated into respective linearized expression vectors (pET28a or pET30b) using T4 DNA ligase and transformed into competent *E. coli* TOP10 cells. All plasmids were verified by DNA sequencing. A ClbQS78A mutant was generated using the quickchange site-directed mutagenesis kit (Stratagene). The Sfp containing plasmid was obtained using a previously reported protocol.

### Table 2-2. Primers used in the study.

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>pET28a-clbQ</td>
<td>ClbQ-nHis_fwd_NdeI</td>
<td>5'-GCGCATATGAGTAATATCAGTTTGTATT-3'</td>
</tr>
<tr>
<td>pET28a-clbQ</td>
<td>ClbQ-nHis_rev_Xhol</td>
<td>5'-GCGGAAGCTTTTACCTACTATTTCGAGTGATT-3'</td>
</tr>
<tr>
<td>pET28a-clbQS78A</td>
<td>ClbQS78A_QC_fwd</td>
<td>5'-TTTTGGGATGCCTCCGGAAGGATC-3'</td>
</tr>
<tr>
<td>pET28a-clbQS78A</td>
<td>ClbQS78A_QC_rev</td>
<td>5'-ATGGCCGAGTCTCCACCT-3'</td>
</tr>
<tr>
<td>pET28a-clbE</td>
<td>ClbE-nHis_fwd_NdeI</td>
<td>5'-GCGCATATGAAAAAGCAAGATGAAAGGCG-3'</td>
</tr>
<tr>
<td>pET28a-clbE</td>
<td>ClbE-cHis_rev_Xhol</td>
<td>5'-GCGGAAGCTTTTACTCGAGGCGTGGTCGTC-3</td>
</tr>
<tr>
<td>pET28a-fscF</td>
<td>FscF-nHis_fwd_NdeI</td>
<td>5'-GCGCATATGACCAACGACGATCGTACTC-3'</td>
</tr>
<tr>
<td>pET28a-fscF</td>
<td>FscF-nHis_rev_HindIII</td>
<td>5'-GCGGAGCTTTAGGGCTGGTGAGG-3'</td>
</tr>
<tr>
<td>pET28a-clbHA1</td>
<td>ClbHA1_nHis_fwd_EcoRI</td>
<td>5'-GGGTGCGAATTCTAGGAAGCAAGG-3'</td>
</tr>
<tr>
<td>pET28a-clbHA1</td>
<td>ClbHA1-nHis_rev_Xhol</td>
<td>5'-TACTTGCTCGAGTAACGTAATCGGCGC-3'</td>
</tr>
<tr>
<td>pET30b-clbE</td>
<td>ClbE-cHis_fwd_NdeI</td>
<td>5'-GCAGCGCATATGAAAAAGCAAGATGAAAGGCG-3'</td>
</tr>
<tr>
<td>pET30b-clbE</td>
<td>ClbE-cHis_rev_Xhol</td>
<td>5'-ATACTTGCTCGAGTAACGTAATCGG-3'</td>
</tr>
<tr>
<td>pET30b-clbF</td>
<td>ClbF-cHis_fwd_NdeI</td>
<td>5'-GCGCAGCTATGAAAAAGCAAGATGAAAGGCG-3'</td>
</tr>
<tr>
<td>pET30b-clbF</td>
<td>ClbF-cHis_rev_Xhol</td>
<td>5'-ATTAATCTCGAGTCCAGGCGTGGTGTCG-3'</td>
</tr>
<tr>
<td>pET30b-clbD</td>
<td>ClbD-cHis_fwd_NdeI</td>
<td>5'-GCTTGGGATATGAAACTCGCGGCTGGATAAG-3'</td>
</tr>
<tr>
<td>pET30b-clbD</td>
<td>ClbD-cHis_rev_NdeI</td>
<td>5'-ATA TTA CTC GAGCTGCTCGGC GTACGTC-3'</td>
</tr>
</tbody>
</table>

*Mutations are underlined*

### 2.6.2 Protein Expression and Purification

Unless otherwise stated, all the proteins were expressed and purified using the following protocol. Expression vectors were transformed into *E. coli* BL21(DE3) and grown in 2L of LB medium containing 50 µg mL\(^{-1}\) kanamycin at 37 °C and 155 rpm to an optical density of 0.5, equilibrated to 18 °C, induced with 0.2 mM IPTG and expression continued at 18 °C for ~ 18h. Cells were harvested and suspended in 25mL of lysis buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl) and lysed with a microfluidizer and the lysate was clarified by centrifugation at 14000g for 40min. The supernatant was
incubated with 1mL of Ni-NTA resin (Qiagen) for 45 mins at 4 °C and washed with one column volume of lysis buffer then one column volume of wash buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 25 mM imidazole). The tagged protein was eluted with elution buffer (20mM Tris-HCl pH 7.5, 500mM NaCl, 250mM imidazole) and dialyzed overnight in the buffer containing 20mM Tris-HCl pH 7.5, 100mM NaCl, 1mM βME, 10% glycerol. The protein was further purified by gel filtration chromatography (HiLoad 16/60 SuperDex-200 column, AKTA FPLC System, GE Healthcare) with 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1mM βME. The fractions containing protein, as detected by SDS-PAGE analysis were pooled, concentrated and used for crystallization experiments.

FscF was also expressed and purified as described above except that it was induced at 37 °C with 0.25 mM IPTG and allowed to grow at 37 °C for ~ 5h before harvesting and purification. Sfp was expressed and purified following the reported protocol. Proteins ClbH, ClbD, ClbF and ClbE (c-term histag) were expressed and purified according to the reported protocol.

2.6.3 Crystallization

Purified ClbQ was concentrated to 5 mg mL⁻¹ and incubated with 2 mM precolibactin SNAC (27) for 2 h at 37°C. Co-crystallization trials were conducted using the hanging drop vapor diffusion method with 2 μL drops containing equal volumes of protein stock and reservoir solution (30% PEG-4000 and 0.1 M Tris-HCl, pH 8.5, 0.2 M lithium sulfate monohydrate) at 20°C. Crystals formed in ~12 hours and were harvested in loops, cryoprotected with corresponding well solution supplemented with 20% glycerol, and frozen in liquid nitrogen for data collection.
2.6.4 Data Collection and Structure Determination

Diffraction data were collected on beamline 21-ID-G of the Life Sciences Collaborative Access Team (LS-CAT) facility at the Advanced Photon Source (APS), Argonne National Laboratory (Argonne, IL). Data were collected at 100° K with a wavelength of 0.9767 Å (1 Å=0.1 nm), integrated, merged and scaled using XDS to a resolution of 1.98 Å in space group P121, with two protein molecules per asymmetric unit. A sequence homology search indicated that the RedJ (32%) and RifR (30% identity) TEs had the highest sequence identity. We tested these α/β hydrolase domains (PDB codes 3FLB, 3QMV), along with the human TEII (4XJV; 25% identity) as molecular replacement search models with PHASER. Initial refinement of the diffraction data (ClbQ crystallized with 27) using RifR as the initial search model provided a partial structure of ClbQ containing residues 2-110, 150-182, 204-240. Missing residues were built manually, and the final atomic model containing residues 2-240 was completed by several rebuilding and refinement cycles using PHENIX.REFINE and COOT. A standard distance restraint between S\(^{\text{ME}}\) and SCys was added (2.05 Å) during the refinement. Water molecules were placed in the structure based on manual inspection of the 2mFo-DFc and mFo-DFc electron density maps and the refined co-ordinates have been deposited in the PDB (accession code 5UGZ). Statistics on data collection and atomic structure refinement are given in Table 2-1. PyMOL was used for structural illustrations.

2.6.5 Acylation of Carrier Protein Domains

Sfp from *Bacillus subtilis* was used for the conversion of apo-ACP to corresponding holo-ACP (acyl-ACPs) using previously reported procedure. Briefly,
each reaction mixture of 100 µL contained 50µM apo-ACP, 150µM acyl-CoA, 2 µM Sfp in Tris HCl buffer, pH 7.5 and 0. 5mM TCEP and 2.5mM MgCl2 and was incubated for 2 h at 37 °C. Acyl-ACP formation was confirmed by MALDI and ESI-LC-MS. The reaction mixtures were then filtered using Microcon YM-3 filter (Millipore) to remove excess acyl-CoA substrate, and exchange the buffer to 100mM Tris HCl, pH 7.5, 20mmM NaCl and concentrate the Acyl-ACP product to 90 µL.

2.6.6 Acyl-ACP Hydrolysis

An LC-MS and MALDI-TOF based assay was used to determine the hydrolysis of Acyl-ACP catalyzed by ClbQ. In a standard assay reaction, 10 µL of the concentrated Acyl-ACP product was incubated with 5 µL of 1 µM ClbQ or ClbQS78A mutant or no enzyme for 15 min at 37 °C. Reactions were quenched with 10 µL of 10% formic acid and the loss of acyl-ACP and formation of holo-ACP were analyzed by LC-MS. Samples for LC-MS analysis were injected via auto sampler and mass analysis was performed using an Agilent 6130 quadruple LC-MS with an Agilent Zorbax SB-C18 1.8 µm (2.1 × 50 mm) column operated at a flow rate of 0.2 mL/min. The running method was 0-2 min (H2O/0.1% formic acid), 2-22 min (0-100% acetonitrile (ACN)/0.1% formic acid). Retention times were as follows: Holo-FscF (FscF-Ppant), 13.74 min; Acetyl-FscF, 13.84 min; Lauroyl-FscF, 14.36 min; Myristoyl-FscF, 14.41 min.

ClbE-Acylation and acyl-ClbE hydrolysis catalyzed by ClbQ and its mutant was confirmed by MALDI-TOF MS-based assay using benchtop MicroFlex LT mass spectrometer (Bruker Daltonics) following previously reported method.73

2.6.7 AM-ClbE Hydrolysis Assays

An in vitro reconstitution of AM-ClbE was generated using previously reported protocol.40 Each 40µL of the reaction mixture containing seryl-ClbE or AM-ClbE were
incubated with 1 µM of ClbQ or ClbQ S78A mutant or no enzyme for 15 min at 37°C. Reactions were quenched with 10 µL of 10% formic acid and the loss of acyl-ACP and formation of holo-ACP were analyzed by LC-MS. Samples for LC-MS analysis were injected via auto sampler and mass analysis was performed using an Agilent 6130 quadrupole LC-MS with an Agilent Zorbax SB-C18 1.8 µm (2.1 × 50 mm) column operated at a flow rate of 0.2 mL/min. The running method was 0-2 min (H₂O/0.1% formic acid), 2-50 min (0-100% acetonitrile (ACN)/0.1% formic acid). Deconvolution of protein raw mass spectra was performed using Agilent MassHunter software (version B.06.00). The masses detected correspond to [M+H]⁺ ions.

2.6.8 Precolibactin SNAC Hydrolysis Assays

An LC-MS based assay was used to determine the TE activity of ClbQ towards precolibactin SNAC thioesters using previously reported method.⁴⁵,⁶⁹ Each assay was carried out in triplicate. In a standard 100 µL assay reaction containing 50 mM Tris HCl pH 8.0, 80 µM substrate (in DMSO) was incubated with 4 µM ClbQ, or ClbQ S78A mutant or no enzyme at 37°C for 30 min and 18 h. Each reaction mixture was processed and analyzed according to a previously reported method.⁴⁵

2.6.9 Acyl-SNAC Hydrolysis versus Precolibactin SNAC Hydrolysis

The DTNB (5,5'- dithiobis (2- nitro-benzoic acid) based continuous spectrophotometric assay was used to examine substrate specificity of ClbQ with various acyl-SNACs and precolibactin SNACs using previously reported protocol.⁷⁰ In brief, each experiment (in triplicates) was carried out at 37 °C in a 96- well flat bottom microtiter plate. Each reaction mixture containing varying of concentrations of acyl-SNACs was added to 100µM DTNB in reaction buffer (100 mM Tris HCl pH 7.4; 20 mM NaCl) and incubated with 100nM ClbQ, or ClbQS78A or no enzyme. All data points
were collected in triplicates and the amount of free thiol released is measured at 412nM. Data analysis and calculation of \(k_{cat}\) and \(K_M\) values were performed using GraFit 4.012 (Middlesex, UK).

2.6.10 Synthesis of Acyl-SNACs

Acyl-SNACs are synthesized according to reported protocol for acetyl-SNAC. A solution of corresponding acid (1 mmol) was dissolved in DCM and cooled to 0° C for ten minutes. 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDCI, 1 mmol), 4-Dimethylaminopyridine (DMAP, 1 mmol) and \(N\)-acetylcysteamine (1mmol) were added and the mixture was left stirring for ~16 h at room temperature. The reaction mixture was concentrated under reduced pressure, and saturated aqueous ammonium chloride was added and extracted with ethyl acetate. Organic fractions were collected, dried with anhydrous potassium sulfate, filtered and concentrated under reduced pressure. The crude reaction mixture was further purified by silica gel chromatography with ethyl acetate to provide the desired SNAC thioester derivative. The purity of the compounds was confirmed by LCMS and NMR spectra.

2.6.10.1 \(S\)-(2-acetamidoethyl) ethanethioate (Acetyl SNAC)

\(^1\)H NMR (CDCl\(_3\), 500 MHz): \(\delta\) 6.38 (bs, 1H), 3.45 (m, 2H), 3.05 (t, \(J=6.47\) Hz, 2H), 2.37 (s, 3H), 2.03 (s, 3H). \(^{13}\)C NMR (CDCl\(_3\), 125 MHz): \(\delta\) 196.3, 170.8, 39.78, 30.63, 28.65, 22.93. ESIMS (m/z): [M+H]\(^+\) calcd. for C\(_6\)H\(_{11}\)NO\(_2\)S 162.05, found 162.05. Yield 75%.

2.6.10.2 \(S\)-(2-acetamidoethyl) dodecanethioate (Lauroyl SNAC)

\(^1\)H NMR (CDCl\(_3\), 500 MHz): \(\delta\) 6.25 (bs, 1H), 3.45 (m, 2H), 3.05 (t, \(J=6.49\) Hz, 2H), 2.57 (t, \(J=7.59\) Hz, 2H), 1.99 (s, 3H), 1.65 (m, 2H), 1.33-1.26 (m, 16H), 0.88 (t, \(J=6.56\) Hz 3H). \(^{13}\)C NMR (CDCl\(_3\), 125 MHz): \(\delta\) 200.2, 170.5, 44.1, 39.8, 31.9, 29.6, 29.4,
29.3, 29.2, 28.9, 28.4, 25.7, 23.1, 22.7, 14.1. ESIMS (m/z): [M+H]+ calcd. for C_{16}H_{31}NO_{2}S 302.21, found 302.21. Yield 90%

2.6.10.3 S-(2-acetamidoethyl) tetradeCANethioate (Myristoyl SNAC)

\(^1\)H NMR (CDCl\(_3\), 500 MHz): δ 6.25 (bs, 1H), 3.45 (m, 2H), 3.05 (t, J=6.47 Hz, 2H), 2.59 (t, J= 7.59 Hz, 2H), 2.04 (s, 3H), 1.68 (m, 2H), 1.35-1.28 (m, 21H), 0.99 (t, J= 6.83 Hz, 3H). \(^{13}\)C NMR (CDCl\(_3\), 125 MHz): δ 200.4, 170.4, 44.2, 40.1, 31.9, 29.7, 29.6, 29.6, 29.4, 29.3, 29.2, 28.9, 28.3, 25.7, 22.7, 14.1. ESIMS (m/z): [M+H]+ calcd. for C_{18}H_{35}NO_{2}S 330.24, found 330.24. Yield 94%.

2.6.11 Synthesis of Precolibactin SNACs

Our collaborators Eric Trautman and Dr. Alan Healy from Crawford lab at Yale university synthesized Precolibactin SNACs 20, 26 and 27 from previously reported acid derivatives 1, 34 18 and 1943 respectively.

2.6.11.1 S-(2-acetamidoethyl) (R)-4-amino-4-oxo-2-tetradeCANamidobutanethioate (N-myristoyl-D-asparagine-SNAC; 20)

N-myristoyl-D-asparagine (1) (25 mg, 1.0 equivalent) was dissolved in 5 mL DCM under nitrogen. The solution was cooled to 0 °C on ice. N-acetylcycteamine (19 μL, 1.2 equivalents), EDCI (14 mg, 1.2 equivalents), and DMAP (2 mg, 0.2 equivalents) were added to the solution. The reaction was stirred overnight, letting warm to room temperature. The solvent was removed in vacuo and the resulting thioester was purified via preparative HPLC (column: Agilent Polaris 5 C18-A 250 × 21.2mm) with a water: acetonitrile gradient. The gradient used was as follows: 10-95% acetonitrile over 50 min, then hold at 95% acetonitrile for 10 min using a flowrate of 10 mL/min. The product was eluted at 44.3 minutes. Removal of solvent by lyophilization provided the product 20 as a white solid (6.5 mg, 20%). \(^1\)H NMR (500 MHz, CD\(_3\)OD): δ4.87 (1 H), 53.31 (2H), 53.00
(2H, t), δ2.78 (1H, dd), δ2.68 (1H, dd), δ1.92 (3H, s), δ1.64 (2H, t), δ1.29 (22H, m),
δ0.90 (3H, t). ESIMS (m/z): [M+H]^+ calcd. for C_{22}H_{41}N_{3}O_{4}S 444.28, found 444.28.

2.6.11.2 S-(2-acetamidoethyl)-2'-(3-1-((S)-6-((R)-4-amino-4-oxo-2-tetradecanamido
butanamido)-3-oxoheptanamido) cyclopropyl)-3-oxopropamido) methyl)-[2,4'-
bithiazole]-4-carbothioate (Linear Precolibactin SNAC; 26)

PyBOP (50.0 mg, 96.1 μmol, 2.00 equiv.) and N,N-diisopropylethylamine (25.1
μL, 0.14 mmol, 3.00 equiv.) were added sequentially to a solution of 18 (40.0 mg, 48.1
μmol, 1 equiv.) in N,N-dimethyl formamide (500 μL) at 23 ºC. The reaction mixture was
stirred for 15 mins before the addition of N-acetyl cysteamine (7.67 μL, 72.1 μmol, 1.50
equiv.). The reaction mixture was stirred for a further 16 h at 23 ºC. The product mixture
was diluted with aqueous citric acid solution (5%, 5.0 mL). The resulting precipitate was
isolated by filtration. The solid was washed with H₂O (2 × 5.0 mL), ice-cold methanol
(5.0 mL) and dried in vacuo to provide the linear precolibactin SNAC 26 as a white solid
(31.0 mg, 69%). H NMR (600 MHz, DMSO-d₆): δ 8.99 (t, J = 6.0 Hz, 1H), 8.84 (bs, 1H),
8.55 (s, 1H), 8.27 (s, 1H), 8.14 (t, J = 5.7 Hz, 1H), 7.90 (d, J = 7.7 Hz, 1H), 7.53 (d, J =
8.6 Hz, 1H), 7.27 (bs, 1H), 6.87 (bs, 1H), 4.61 (d, J = 5.9 Hz, 2H), 4.43 (q, J = 7.2 Hz,
1H), 3.75 - 3.64 (m, 1H), 3.63 (s, 2H), 3.35 (s, 2H), 3.31 - 3.25 (m, 2H), 3.08 (t, J = 6.7
Hz, 2H), 2.57 - 2.49 (m, 2H), 2.42 (dd, J = 15.1, 5.6 Hz, 1H), 2.36 (dd, J = 15.3, 7.8 Hz,
1H), 2.09 (t, J = 7.5 Hz, 2H), 1.81 (s, 3H), 1.67 - 1.55 (m, 1H), 1.57 - 1.47 (m, 1H), 1.49
- 1.42 (m, 2H), 1.40 (q, J = 3.1 Hz, 2H), 1.22 (bs, 20H), 1.06 (q, J = 3.4 Hz, 2H), 0.99 (d,
J = 6.6 Hz, 3H), 0.84 (t, J = 6.8 Hz, 3H). C NMR (151 MHz, DMSO-d₆): δ 204.8, 204.2,
185.2, 172.2, 171.6, 171.4, 170.5, 169.3, 168.1, 167.1, 162.6, 152.3, 146.6, 125.4,
118.9, 50.2, 49.9, 46.5, 43.6, 40.6, 40.5, 39.1, 38.1, 37.4, 35.2, 31.3, 29.7, 29.09, 29.07,
29.03, 28.98, 28.87, 28.73, 28.67, 28.1, 25.2, 22.6, 22.1, 20.6, 19.4, 14.0. ESIMS (m/z):
[M+H]^+ calcd. for C_{43}H_{65}N_{9}O_{9}S_{3} 933.40, found 933.40.
2.6.11.3 S-(2-acetamidoethyl)2'-((4'-(S)-3-((R)-4-amino-4-oxo-2-tetradecanamido butanamido) butyl)-3',6'-dioxo-2',3'-dihydrospiro[cyclopropane-1,1'-pyrrolo[3,4-c] pyridin]-5'(6'H)-yl) methyl)-[2,4'-bithiazole]-4-carbothioate (Cyclic Precolibactin SNAC; 27)

PyBOP (11.8 mg, 22.6 μmol, 2.00 equiv.) and N,N-diisopropylethylamine (5.91 μL, 74.2 μmol, 3.00 equiv.) were added sequentially to a solution of 19 (9.0 mg, 11.3 μmol, 1 equiv.) in N,N-dimethylformamide (300 μL) at 23 °C. The reaction mixture was stirred for 15 mins before the addition of N-acetyl cysteamine (1.80 μL, 17.0 μmol, 1.50 equiv.). The reaction mixture was stirred for a further 16 h at 23 °C. The product mixture was diluted with aqueous citric acid solution (5%, 3.0 mL). The resulting precipitate was isolated by filtration. The solid was washed with H₂O (2 × 3.0 mL) and dried in vacuo to provide the precolibactin SNAC 27 as a white solid (7.2 mg, 71%).

¹H NMR (600 MHz, DMSO-d₆) δ 8.52 (s, 1H), 8.49 (bs, 1H), 8.28 (s, 1H), 8.13 (t, J = 5.7 Hz, 1H), 7.83 (d, J = 7.9 Hz, 1H), 7.67 (d, J = 8.0 Hz, 1H), 7.24 (bs, 1H), 6.81 (bs, 1H), 6.17 (s, 1H), 5.60 (d, J = 15.8 Hz, 1H), 5.54 (d, J = 15.9 Hz, 1H), 4.48 (app q, J = 7.1 Hz, 1H), 3.87 (app hept, J = 6.5 Hz, 1H), 3.49 - 3.37 (m, 1H), 3.37 - 3.34 (m, 1H), 3.28 (dt, J = 12.7, 6.0 Hz, 2H), 3.07 (t, J = 6.7 Hz, 2H), 2.49 - 2.44 (m, 1H), 2.35 (dd, J = 15.2, 7.2 Hz, 1H), 2.04 - 1.97 (m, 2H), 1.80 (s, 3H), 1.77 - 1.68 (m, 2H), 1.54 - 1.39 (m, 2H), 1.40 - 1.36 (m, 2H), 1.36 - 1.33 (m, 2H), 1.27 - 1.11 (m, 20H), 1.07 (d, J = 6.5 Hz, 3H), 0.84 (t, J = 6.9 Hz, 3H).

¹³C NMR (151 MHz, DMSO-d₆) δ 185.1, 172.1, 171.5, 170.4, 169.3, 167.1, 166.7, 162.4, 161.8, 159.8, 153.1, 152.3, 146.3, 125.6, 119.6, 109.6, 103.3, 49.9, 44.7, 44.4, 40.1, 38.1, 37.3, 35.4, 35.2, 31.3, 29.06, 29.04, 29.01, 28.96, 28.86, 28.7, 28.6, 28.0, 25.2, 22.5, 22.1, 20.3, 15.2, 14.0. ESIMS (m/z): [M+H]+ calcd. for C₄₃H₆₁N₈O₇S₃ 897.38, found 897.38.

71
2.6.12 Catalog of NMR Data of Acyl-SNACs and Precolibactin SNACs

Figure 2-17. $^1$H NMR spectrum of the Acetyl SNAC.
Figure 2-18. $^{13}$C NMR spectrum of the Acetyl SNAC.
Figure 2-19. $^1$H NMR spectrum of the Lauryl SNAC.
Figure 2-20. $^{13}$C NMR spectrum of the Lauryl SNAC.
Figure 2-21. $^1$H NMR spectrum of the Myristoyl SNAC.
Figure 2-22. $^{13}$C NMR spectrum of the Myristoyl SNAC.
Figure 2-23. $^1$H NMR spectrum of the N-myristoyl-D-asparagine-SNAC; 20.
Figure 2-24. $^1$H NMR spectrum of the Linear Precolibactin SNAC; 26.
Figure 2-25. $^{13}$C NMR spectrum of the Linear Precolibactin SNAC; 26.
Figure 2-26. $^1$H NMR spectrum of the Cyclic Precolibactin SNAC; 27.
Figure 2-27. $^{13}$C NMR spectrum of the Cyclic Precolibactin SNAC; 27.
CHAPTER 3
STRUCTURAL STUDIES OF THE S-ADENOSY-L-METHIONINE (SAM) ACTIVATING ADENYLATION DOMAIN OF NRPS MODULE CLBH IN THE COLIBACTIN PATHWAY

3.1 Adenylation Domain of NRPS System

Adenylation (A) domains, generally referred to as “gate keepers” of NRPs, initiate and introduce the first level of substrate specificity in NRPS systems. By selecting specific amino acid building blocks, they dictate the peptide sequence. The A domains catalyze a two-step reactions to activate and load each amino acid onto the peptide chain (Figure 3-1).

**Figure 3-1. Adenylation domain reaction.**

The first step is an ATP-dependent reaction that involves the activation of the carboxylate group of an amino acid or aryl acid monomer substrate to its respective amino (or aryl-) acyl-AMP intermediate releasing pyrophosphate. The second step involves the transfer of the intermediate, formed in the first step onto the 4′-phosphopantetheine (4′-PP or Ppant) arm of the adjacent carrier (PCP) domain, and thus forming an acyl thioester and releasing AMP. A domain crystal structures, in both apo- and holo-forms, as well as bi- and multi-domain A-T, and C-A-PCP-TE have been reported in different NRPS systems, showing a common subdomain architecture.\(^{29,90,91}\) Crystallographic snapshots of the A domains in different catalytic states reveal that the arrangement of the C-terminal subunit during the catalytic cycle cause the A domain to
adopt several distinct conformations to conduct the two partial reactions (adenylation and thiolation) within the same active site located between the subdomains.\textsuperscript{90,91} Also, detailed structural analysis and sequence alignment resulted in the disclosure of the so-called “specificity conferring code”, or Stachelhaus motif, the ten core amino acid residues responsible for substrate selection.\textsuperscript{33} Although the structures of A domains are diverse, there is significant structural homology among A domains, and this code is often used by prediction programs for \textit{in silico} analysis to predict peptides assembled by uncharacterized NRPS systems identified via genome mining. Additionally, the spectroscopic assay of radiolabeled pyrophosphate exchange assay measuring inorganic pyrophosphate released by the adenylation reaction can be employed to determine the predicted substrate specificity.\textsuperscript{92}

As an additional means to incorporate structural diversity in NRPs, A domains that are commonly disrupted in the conserved signature sequences located in C-terminal domain by the auxiliary enzymes, including M, KR, and Ox, have been observed in various NRPS systems.\textsuperscript{93} These interrupted A-domains have been found to be catalyze two distinct chemical reactions. While the mechanistic and structural details of these interrupted A domains remain to be resolved, the existence of these bifunctional A domains provides another viable means of controlling novel NRPS biosynthesis which needs to be explored. In addition, many NRPS A domains also require an auxiliary protein belonging to MbtH-like protein (MLP) superfamily for solubility or for activation of amino acids substrate by adenylation domains.\textsuperscript{93,94} A very recent report highlighted that the improper NPS+MLP interaction can negatively impact
not only amino acid solubility but also aminoacyl-S-PCP formation, along with the overall function of the NRPS.\textsuperscript{95}

### 3.2 ClbH, a Unique NRPS Module in the Colibactin Pathway

Recent experiments have demonstrated that reported colibactin compounds exert their DNA-damaging activity via alkylation of DNA by cyclopropane ring opening.\textsuperscript{55} This cyclopropane ring was previously shown to be incorporated by ClbH, an NRPS enzyme utilizing 1-aminocyclopropane-1-carboxylic acid (ACC) unit.\textsuperscript{36,37} However, the biosynthetic rationale for incorporation of ACC unit is unknown. In plants, ethylene biosynthesis involves an ACC intermediate derived from methionine via a pyridoxal phosphate (PLP)-dependent mechanism.\textsuperscript{96,97} Cyclopropane formation in other biosynthetic pathways has been suggested to involve S-adenosyl-L-methionine (SAM)-dependent alkylation (Jawsamycin\textsuperscript{98}), Favoriskii-like rearrangement (ambruticin\textsuperscript{99}), intramolecular nucleophilic ring opening of a lactone (hormaomycin\textsuperscript{100}), or a chlorination and NADPH-mediated ER domain (Curacin A\textsuperscript{101}). However, \textit{E. coli} is not a known producer of ACC and the \textit{pks} gene cluster does not encode a PLP-dependent enzyme, or a halogenase as in known pathways, suggesting ClbH pathway utilizes a novel chemistry to incorporate ACC in the assembly line.

Homology studies suggested ClbH is a unique enzyme component of the colibactin NRPS biosynthetic machinery with a rare A\textsubscript{1}-C-A\textsubscript{2}-PCP domain arrangement. In contrast to the canonical NRPS architecture, ClbH lacks a PCP domain downstream of the A domain at the N-terminus of ClbH (ClbH-A\textsubscript{1}). Instead, ClbH-A\textsubscript{1} is directly connected to a C domain and this unique arrangement of ClbH was confirmed by resequencing studies conducted by Brachmann, O. \textit{et al.}\textsuperscript{32} Previous studies from multiple groups including our own study (Chapter 2) have confirmed the role of ClbH-A\textsubscript{1}
activating L-Ser which is further processed by accessory enzymes ClbDEF to produce the AM unit (Figure 1-5).\textsuperscript{32,40,84} However, the ClbH A\textsubscript{2} domain lacks conserved “Stachelhaus motif” residues, suggesting it does not activate an $\alpha$-amino acid. This unusual feature of ClbH A\textsubscript{2} together with ClbH non-canonical arrangement and role in the incorporation of ACC unit has attracted attention from various researchers.

Earlier research by Piel and co-workers employed traditional ATP-$[^{32}\text{P}]$ pyrophosphate (PP\textsubscript{i}) exchange assay to investigate the amino acid specificities of ClbH-A\textsubscript{2} domain, using both proteinogenic and non-proteinogenic substrates including ACC and SAM.\textsuperscript{32} ClbH-A\textsubscript{2} was shown to be inactive against all the tested substrates. However, the authors used an excised ClbH-A\textsubscript{2} lacking the partnering PCP and C domains for the PP\textsubscript{i} exchange assays. Previous studies on A domains have strongly suggested that the disruption of linker region between A-PCP or C-A domains in a multidomain NRPS module could influence the activity of individual domains, resulting in their inactivity.\textsuperscript{102,103}

Later, based on radiolabeled amino acid feeding experiments, the Crawford group proposed ACC, and SAM together with L-Met are potential substrates for either ClbH-A\textsubscript{1} or ClbH-A\textsubscript{2}, to account for the presence of ACC in isolated metabolites.\textsuperscript{28} The PP\textsubscript{i} exchange assay suggested ClbH-A\textsubscript{1} displayed significant specificity for L-Ser and a mild specificity for ACC. Based on gene deletion experiments, it has been speculated that ClbH-A\textsubscript{1} may participate in ACC. However, this speculation is limited by the technique used as the gene deletion method eliminates the entire NRPS module and one cannot predict the role of individual enzyme contributions to the observed activity. To overcome this issue, Crawford lab used domain targeted metabolomics approach to
individually inactivate all the ClbH domains and measure the production of clb-dependent metabolites.\(^{36}\) It was shown that the inactivation of any of the individual C-A\(_2\)-PCP domains abolished the production of cyclopropane-containing metabolites, whereas inactivation of ClbH-A\(_1\) has no effect the production. The results from this work eliminated ClbH-A\(_1\) as a possible ACC supplier, suggesting a unique way of incorporation or formation of the ACC unit by ClbH C-A\(_2\)-PCP domain.

Figure 3-2. Proposed ways to incorporate ACC in the colibactin biosynthesis.

The three possible ways to incorporate ACC in the colibactin biosynthesis is illustrated in Figure 2-29. While it has been clear that ClbH-A\(_1\) is dispensable for cyclopropane ring formation, eliminating proposed pathways A and B in Figure 2-29, direct activation and loading of SAM has been recently re-evaluated by the Balskus
In contrary to an excised ClbHA₂ used in the prior study by Piel and co-workers, the Balskus group used a full-length ClbH and C-A₂-PCP construct for the PP₁ exchange assay to evaluate substrate specificity. It was shown that full length ClbH activated L-Ser (~100% rate) and SAM (~20% rate) whereas ClbH C-A₂-PCP activated SAM (~20% rate), and neither of these constructs activate L-Met, ACC, or SAM degradation products L-homocysteine or adenosine. This result suggests that SAM is the substrate for ClbH-A2 and further highlights the role of partnering C and PCP domains critical for A-domain activity. Further in vitro studies revealed that SAM is activated by ClbH-A₂ and then loaded on to the partnering PCP domain. In addition, it was shown that the following PKS module, ClbI, is essential for cyclopropane formation, and proposed that the S178 of ClbI could act as a general base to promote cyclization of ClbH-tethered SAM-derived intermediate. This observation is consistent with the isolation of an isomeric γ-lactone 6a (Figure 3-3) in the absence of ClbI PKS module, suggesting enzymes beyond ClbH are likely involved in cyclopropane formation.

Figure 3-3. ClbI-dependent cyclopropanation of SAM activated by ClbH.

Further bioinformatic analyses revealed key features of the ClbH- A₂ domain. Sequence homology suggests that ClbH-A₂ has modest sequence identity (30-35%) to
structurally characterized A domains. Lacking a conserved Asp residue in the core A4 motif, and an expanded specificity pocket compared to the structurally characterized Phe activating GrsA A-domain, suggest a different arrangement of the ClbH- A₂ active site. Together, previous studies confirmed the role of ClbH-A₂ incorporating the SAM-derived cyclopropane ring in the colibactin pathway. However, the biosynthetic rationale for transformation of L-Met to SAM and precise timing of the cyclopropane ring formation on the assembly line still need to be validated. Therefore, there is a greater need to decipher structural and mechanistic information of the C-A₂-PCP domain to understand SAM recognition and activation mechanisms by ClbH-A₂.

In addition, since NRPSs do not have direct homologs in humans, these proteins are potential orthogonal therapeutic targets. However, the A domains in NRPS uses similar chemistry to ubiquitous aminoacyl- tRNA synthetases and thus represent a challenge to design selective inhibitors for the target protein. All other A domains in the colibactin NRPS pathway have common counterparts in humans except ClbH-A₂. The unique NRPS domain architecture and unusual SAM specificity of ClbH make it an ideal candidate for targeting colibactin biosynthesis. Our efforts to obtain structural information and mechanistic characterization of the ClbH C-A₂-PCP domain are outlined below.

3.3 Cloning, Expression and Crystallization of ClbH C-A₂-PCP Domain

The crystallization of the excised domain fragments of multi-domain synthetases is often challenging in terms of protein solubility and stability. The major challenge is to choose the N- and C- terminal boundaries of discrete domains within multi-enzymes. The standard strategy is to locate the extremities within the linker regions, which are presumably unstructured, on either side of the conserved domains. Since the transition
from functional domains to the linker region is associated with a clear drop in sequence conservation, these linker regions can be hypothesized by sequence analysis. Also, the inter-domain sequences can be revealed by limited proteolysis of intact subunits coupled with N-terminal sequencing. Since the precise boundaries can influence both the stability and activity of the resulting protein constructs, cloning multiple constructs using several terminals tend to be a successful approach. Therefore, we decided to clone multiple constructs with varying N and C- terminus: ClbH-A2; a di-domain construct, ClbH A2-PCP; and a tri-domain construct ClbH C-A2-PCP (Figure 3-4).

The full-length ClbH protein contains 1599 amino acids, and the boundaries of individual domains and linker regions of ClbH are predicted using a NRPS/PKS predictor. Our initial attempts to clone and express soluble ClbH-A2PCP protein using two different constructs (1 and 2, Figure 3-4) were unsuccessful. For both these protein constructs, the overexpression was tested at various conditions, including combination of different induction temperatures (30 °C, 24 °C, 18 °C, 16 °C), different IPTG concentrations (0 µM, 100 µM, 150 µM and 250 µM), and using various vectors (pET28, pET30, pMal-C5x). While almost all the tested conditions resulted in the overexpression of protein in an insoluble form, overexpression of partially soluble protein is observed in the non-induced (0 µM IPTG) cultures of E. coli (BL21-pET30a-ClbH- A2PCP, Construct 1) grown at 16 °C for 72hrs. Because of the low yield and poor stability of this construct, further crystallization studies were not carried out. Meanwhile, our preliminary efforts to express the excised A2 domain (construct 3, Figure 3-4) without the partnering PCP domain resulted in a low yield of protein. However, the expression of a stable protein is difficult to reproduce and therefore discontinued for further crystallization experiments.
Figure 3-4. Schematic of ClbH expression constructs used for cloning.

Since the linker region and the partnering domains plays a crucial role in A domain activity, we decided to clone a tri-domain construct C-A$_2$-PCP with varying N- and C- termini for crystallization studies. While we were continuing our efforts to clone and express C-A$_2$-PCP domain (construct 4, Figure 3-4), the Balskus group reported soluble protein constructs of ClbH C-A$_2$-PCP (construct 5) and a full length ClbH (construct 6). We used a similar approach for cloning and expression of constructs 5 and 6 (Figure 3-4) as per the reported protocol, to obtain soluble protein for our crystallization experiments.

The full length ClbH from the *pks* island is cloned to a pET28a vector with a N-terminal hexahistidine tag. This construct is used as a template to clone ClbH-C-A$_2$-PCP construct (aa 486-1561) to a pET30a vector providing the pET30-*clbH*-CA2PCP plasmid with an N-terminal hexahistidine tag. The plasmid was transformed to BL21 (DE3) strains and expressed, and purified according to the reported protocol. After elution, further purification of ClbH-C-A$_2$-PCP protein (~ 129 KDa) through size exclusion chromatography was carried out and the purity of the samples was analyzed by SDS-PAGE (Figure 3-5). The protein-containing fractions were collected, concentrated to 4 mg/mL, and subjected to further crystallization experiments.
Initial crystallization screening experiments of ClbH C-A₂-PCP were performed in 96-well plates at 20 °C using a sitting drop vapor diffusion method with commercial matrix conditions, including Wizard (Rigaku) and Crystal screens (Hampton). As shown in Figure 3-6, multiple initial hits were found in various conditions for the apo-ClbH C-A₂-PCP protein purified in the absence of βME (β-mercaptoethanol).

Figure 3-5. Size exclusion chromatography of ClbH-C-A₂-PCP protein after Ni-NTA resin purification. Protein-containing fractions were analyzed by SDS-PAGE.

However, following optimization trials using a different batch of protein in the presence of βME were unsuccessful. In addition, we have also performed preliminary sitting drop 96-well screening plates with Wizard (Rigaku) and Crystal screens (Hampton) for holo-ClbH C-A₂-PCPs; where PCP is loaded with a Ppant arm and SAM-Ppant. The Ppant loading reactions on to ClbH C-A₂-PCP are carried out using the
reported protocol. Currently, we are optimizing conditions and testing various crystallization conditions.

Figure 3-6. Crystal hits of apo- ClbH C-A2-PCP in various conditions.

3.4 Future Directions

While waiting to collect the diffraction data of the initial crystal hits, we continue our efforts to repeat the crystallization hits of ClbH C-A2-PCP. Our goal of obtaining apo, holo, and SAM loaded ClbH C-A2-PCP crystal structure is to understand the mechanism of SAM recognition and activation. In addition, we would also like to test whether ClbH-A2 activity is dependent on MLP proteins. The pks cluster does not encode any MLP protein; however previous studies showed that the ability of non-cognate MLPs influence the function of NRPs synthesis. Earlier studies on the ClbH-A2-dependent
activation of SAM suggested low reaction rate (~20% rate). As MLPs influence the activation of amino acid substrates, we will investigate the possibility of MLP dependent activity of ClbH-A2 domain in the activation and incorporation of SAM in the colibactin pathway.

Several classes of NRPS A domain-specific inhibitors, mostly mechanistic-based inhibitors, have been developed. Non-hydrolysable sulfomoyl-based inhibitors mimicking the aminoacyl-AMP intermediate of the reaction are one successful example of A domain inhibitors (Figure 3-7). In addition, non-hydrolysable vinylsulfonamide inhibitors have been reported as probes to study A-PCP interaction in NRPS modules. Given the difficulty of protecting the reactive methyl thioadenosine group in SAM during the inhibitor synthesis process, we will synthesize an ACC-sulfamoyl inhibitor or ACC-AVS compounds as alternative probes to decipher ClbH-A2 activity and mechanism of recognition via co-crystallization studies.

![ACC-sulfamoyl inhibitor and ACC-adenosine vinylsulfonamide (AVS) inhibitor](image)

Figure 3-7. Adenylation specific inhibitors for ClbH.

### 3.5 Experimental Procedures

Unless otherwise noted, all chemicals and general reagents were purchased from Sigma-Aldrich or Fischer Scientific.
3.5.1 Cloning of ClbH Constructs

The genes encoding the full length ClbH protein (construct 6) was amplified by PCR from a bacterial artificial chromosome (BAC) harboring the *pks* genomic DNA and the gene encoding the ClbH C-A2-PCP protein (construct 5) was amplified from a full-length pET-28-ClbH plasmid as a template using the primers listed in Table 3-1. Amplified fragments were digested with appropriate restriction enzymes and were purified using agarose gel electrophoresis and QIAquick gel extraction kit. The digests were ligated in to respective linearized expression vectors (pET28a or pET 30a) using T4 DNA ligase and transformed to competent *E. coli* TOP10 cells. All plasmids were verified by DNA sequencing. Sfp plasmid was obtained from a previously reported protocol.47

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>pET28a-clbH-FL</td>
<td>ClbH-nHis_fwd_EcoRI</td>
<td>5'-CGGTGCGAATTCAATGGAGAAGCAAG-3'</td>
</tr>
<tr>
<td>pET28a-clbH-FL</td>
<td>ClbH-nHis_rev_SacI</td>
<td>5'-GTAAGCGAGCTCTACGTGAAGACTG-3'</td>
</tr>
<tr>
<td>pET30-clbH-CA2PCP</td>
<td>ClbHCA2T-cHisfwd_Ndel</td>
<td>5'-ATAATAGAATTCACTGCTGCAGAAC-3'</td>
</tr>
<tr>
<td>pET30-clbH-CA2PCP</td>
<td>ClbHCA2T-cHisrev_SacI</td>
<td>5'-ATTTAGAGACGCCTCGTGACTGTC-3'</td>
</tr>
</tbody>
</table>

3.5.2 Protein Expression and Purification of ClbH C-A2-PCP

Unless otherwise stated, all the proteins are expressed and purified using the following protocol. The expression vector pET30-clbH-CA2PCP was transformed into *E. coli* BL21(DE3) and grown in 6L of LB medium containing 50 µg mL⁻¹ kanamycin at 37 °C and 155 rpm to an optical density of 0.5, equilibrated to 18 °C, induced with 0.2 mM IPTG and expressed continued at 16 °C for ~18h. Cells were harvested and suspended in the 25mL of lysis buffer (50 mM Tris-HCl pH 8.5, 500 mM NaCl, 10mM MgCl₂, 10% glycerol) and lysed with a microfluidizer, and the lysate was clarified by centrifugation at 14000g for 40min. The supernatant was incubated with 1mL of Ni-NTA resin (Qiagen)
for 45 mins at 4 °C and washed with one column volume of lysis buffer followed by one column volume of wash buffer (50 mM Tris-HCl pH 8.5, 500 mM NaCl, 10 mM MgCl₂, 10% glycerol, 25 mM imidazole). The tagged protein was eluted with elution buffer (50 mM Tris-HCl pH 8.5, 500 mM NaCl, 10 mM MgCl₂, 10% glycerol, 250 mM imidazole) and dialyzed overnight in the buffer containing 50 mM Tris-HCl pH 8.5, 100 mM NaCl, 1 mM βME, 10% glycerol. The protein was further purified by gel filtration chromatography (HiLoad 16/60 Superdex-200 column, AKTA FPLC System, GE Healthcare) with 50 mM Tris-HCl pH 8.5, 100 mM NaCl, 1 mM βME. However, for the first batch of purification, βME was not added in dialysis and gel filtration buffers, providing a broader peak elution. The fractions containing protein, as detected by SDS-PAGE analysis were pooled, concentrated, and used for crystallization experiments. Sfp was expressed and purified following the reported protocol. In vitro reconstitution reactions of holo-CIbH C-A₂-PCP and SAM-CIbH C-A₂-PCP were carried out according to reported protocol by using the purified apo-protein incubated with Sfp, CoA, ATP and SAM. After the reaction, the proteins were further purified by gel filtration chromatography and used for further protein crystallization experiments.

### 3.5.3 Crystallization of CIbH C-A₂-PCP

Purified apo-CIbH C-A₂-PCP, holo-CIbH C-A₂-PCP, SAM-CIbH C-A₂-PCP were concentrated to 4 mg mL⁻¹ and crystallization trials were conducted using 96-well sitting drop vapor diffusion method, with 1 μL drops containing equal volumes of protein stock and reservoir solution of the commercial matrix screens (Wizard (Rigaku) and Crystal screens (Hampton)) at 20°C. Only the apo-CIbH C-A₂-PCP protein was crystallized in
various conditions as shown in Figure 3-6. Further optimization experiments are currently being conducted.
CHAPTER 4
EXAMINING THE ROLE OF CLBL IN THE COLIBACTIN BIOSYNTHETIC PATHWAY

4.1 The Uncharacterized Amidase, ClbL

ClbL is one of the five genes found in the pks gene island to be upregulated in CRC mouse models\(^\text{16}\) and is one of two enzymes yet to be characterized in the colibactin biosynthetic model. Previous gene deletion studies suggested ClbL is needed for cytopathic effects.\(^8\) In addition, studies have showed that ClbL is not critical for compound 1 (N-myristoyl-D-Asn) identification and precolibactin formation.\(^{16,36}\) Based on NCBI BLAST analysis and sequence homology studies, ClbL is predicted to be an amidase and the closest homologues include an amidase from \textit{Mycobacterium tuberculosis} and \textit{Pseudomonas aeruginosa} PA01, followed by fatty acid amide hydrolase (FAAH)-2 as shown in Table 4-1.

<table>
<thead>
<tr>
<th>Description</th>
<th>Accession</th>
<th>Max Score</th>
<th>Total Score</th>
<th>Query Cover</th>
<th>E value</th>
<th>Identity%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amidase [\textit{Mycobacterium tuberculosis} H37RV]</td>
<td>NP_217691.1</td>
<td>210</td>
<td>210</td>
<td>96%</td>
<td>9e-61</td>
<td>36</td>
</tr>
<tr>
<td>Amidase [\textit{Pseudomonas aeruginosa} PA01]</td>
<td>NP_253032.1</td>
<td>162</td>
<td>162</td>
<td>95%</td>
<td>5e-43</td>
<td>31</td>
</tr>
<tr>
<td>Fatty-acid amide hydrolase 2-A [\textit{Danio rerio}]</td>
<td>NP_001002700.1</td>
<td>157</td>
<td>157</td>
<td>97%</td>
<td>7e-41</td>
<td>28</td>
</tr>
<tr>
<td>CG7910, isoform A [\textit{Drosophila melanogaster}]</td>
<td>NP_649765.1</td>
<td>144</td>
<td>144</td>
<td>98%</td>
<td>3e-36</td>
<td>27</td>
</tr>
<tr>
<td>Predicted: FAAH-2 isoform X1 [\textit{Homo sapiens}]</td>
<td>XP_005262039.1</td>
<td>142</td>
<td>142</td>
<td>97%</td>
<td>1e-35</td>
<td>27</td>
</tr>
</tbody>
</table>

Since the A domain of the NRPSs can utilize any amino acid necessary as building blocks, the presence of additional proteins such as amidotransferase (AMT) is very uncommon, and to date only the MycA protein from \textit{Bacillus subtilis} and zwittermicin-producing \textit{Bacillus cereus} are reported as examples.\(^{108,109}\) The AMT domain of MycA is shown to be responsible for the transfer of an amine to the \(\beta\)-position.
of the growing acyl chain, using PLP as a co-factor. The amidase ZmaL, from the zwittermycin biosynthesis, is initially hypothesized to release the growing peptide-polyketide molecule and form a final PCP domain of NRPS/ PKS. However, a later report from the same group confirmed that the release of the growing peptide-polyketide indeed occurs through an alternative oxidative cleavage mechanism. In the colibactin producing *pks* island, a standalone terminal TE domain, ClbQ, is present. While ClbL could possibly release the growing peptide, ClbQ was reported to process the precolibactins release in the colibactin pathway. The presence of multiple hydrolases in the single NRPS/PKS pathway is highly uncommon, leaving the role of ClbL in colibactin production and/or utilization elusive. Therefore, we aimed for a structural approach to investigate the structural and mechanistic role of ClbL as well as to identify its therapeutic potential for CRC treatment.

### 4.2 Preliminary Crystallization Data of ClbL

The ClbL protein (~54 kDa) encodes 487 amino acids. Preliminary studies of cloning, expression, and purification and crystallization of the ClbL protein were conducted by our former group members, Jarrod Mousa and Rachel Newsome. In brief, ClbL was cloned from the BAC containing the *pks* island and was ligated into pET28a (N-terminal hexahistidine tag) and a pET30a construct (C-terminal hexahistidine tag). After several attempts to express the protein in various strains of *E. coli* and at different temperatures, soluble protein was obtained from the overnight expression of C43 (DE3) cells containing a pET28a or pET30a construct by inducing with 0.1M IPTG at 25°C. The purified protein was subjected to crystallization screening by vapor diffusion method. Further crystal optimization studies provided crystal hits in
30% PEG3000 and 100mM CHES pH 9.5 conditions. Despite several unsuccessful attempts that were made to repeat/optimize the quality of these crystals which provided 2.9 Å diffraction data, preliminary structural information on ClbL was obtained (Figure 4-1). From this diffraction data, it was shown that ClbL structure consists of a crystallographic dimer, with two protein molecules per unit cell (space group C121). The electron density for a few amino acid residues (1-5, 47-61, 298-307, 330-372, 479-487) was missing due to poor diffraction pattern and quality of the crystals. However, the active site was visible and we identified the key active site residues as S155, S179, K80. The catalytic triad Ser-Ser-Lys is unique and highly conserved among the amidase family of enzymes, although the catalytic mechanism for acyl enzyme intermediate formation may vary between enzymes.

Figure 4-1. Unpublished crystal structure of ClbL as a crystallographic homodimer. The α-helices are depicted in cyan, β sheets in magenta, and loop regions are in salmon red.

4.3 Optimization of ClbL Crystals

Several attempts were made to repeat the crystallization of ClbL. Micro-seeding crystallization efforts using initial crystals as seeding stocks in the commercial buffer
0.1M CHES pH 9.5 and 30% PEG 3000, provided multiple small crystals of ClbL (Figure 4-2A). However, our attempts to obtain diffraction quality crystals have been unsuccessful to date. Meanwhile, we observed single crystals in previous hanging drop optimization trays that were ~ 8 months old (Figure 4-2B). These crystals also provided similar resolution (~ 3.0 Å) diffraction data, though they indexed into an alternate space group P3\(_1\)21. This ClbL structure was monomeric and contained one protein molecule per unit cell (Figure 4-2C). While the electron density for a few amino acid residues is still missing, the active site is visible. We are continuing our efforts to obtain a single diffraction quality crystal to solve the complete structure of the protein.

Figure 4-2. Crystal structure of ClbL as a crystallographic monomer. A) Crystal hits from microseeding experiments. B) Single crystals from optimization trays. C) cartoon representation of crystal structure of ClbL monomer in space group P3\(_1\)21. The active site residues are shown as sticks.

4.4 Site-Directed Mutagenesis of ClbL Active Site Residues

As we are continuing our efforts to optimize ClbL crystals, we have also tried to clone and express ClbL active site mutants for crystallization experiments and biological
assay studies. Six ClbL mutants were originally constructed: S179A, S155A, S155W, K80W, K80R, K80A. All of these mutants were cloned to pET-28a and pET-30a vectors and the expression levels were tested at varying induction temperatures (37°C, 25°C, and 18°C and in different cells including C43, BL21(DE3), and C41 (Figure 4-3). Interestingly, three of these mutants, S179A, S155A and S155W, resulted in proteins that either did not express or express only in inclusion bodies. No active protein could be recovered from cell lysates, suggesting these residues are crucial for protein folding. The K80 mutants K80W, K80R, K80A were expressed in various conditions, and only K80A mutant in pET30a variant, expressed at 25°C, provided a soluble protein and in levels comparable to those of wild-type ClbL.

![SDS-PAGE analysis of expression of ClbL mutants at varying conditions.](image)

Figure 4-3. SDS-PAGE analysis of expression of ClbL mutants at varying conditions.
4.5. Proposed Roles of ClbL in the Colibactin Biosynthetic Pathway

![Chemical diagram showing the proposed role of ClbL as a GatA homologue.]

Figure 4-4. ClbL function as a GatA homologue.

Based on our initial sequence homology studies, ClbL was predicted to be an amidase/amido-transferase enzyme, serving to convert an Asp/Glu to an Asn/Gln, similar to GatA. The GatA protein of the GatCAB complex is a structurally characterized amido-transferase enzyme generates ammonia from glutamine, which further reacts with Asp-tRNA\(^{Asn}\) to form Asn-tRNA\(^{Asp}\) (Figure 4-4).\textsuperscript{113} Interestingly, the \textit{pks} island is found within the \textit{asnW} tRNA locus in \textit{E. coli},\textsuperscript{20} and the ClbN A-domain utilizes the L-Asn moiety. However, previous gene deletion studies and domain-targeted metabolomics revealed that ClbL is dispensable for production of reported precolibactin metabolites, and ClbL does not complement the ClbQ activity.\textsuperscript{36}

While the previous studies suggest that ClbL is not involved in biosynthesis of the isolated precolibactin molecules, the current biosynthetic model does not include ClbO, and the final colibactin structure is yet unknown. In addition, an increased production of late stage intermediates was observed in ClbL active site mutant cells in the domain-targeted metabolomic study.\textsuperscript{36} Therefore, the ammonia generated by the amidase activity of ClbL might possibly be incorporated in the colibactin structure or perhaps
modifies the colibactin structure to provide a reactive amine for conjugation with other proteins (Figure 4-5).

To test the above hypotheses, we first determined the amidase activity of ClbL converting L-Asn to L-Asp. We incubated purified recombinant ClbL with Asn substrates; L-Asn and N-myristoyl-D-Asn (1), at concentrations up to 3mM. No amidase activity was detected for either substrates at the tested conditions. In contrast, our collaborators did observe the conversion of L-Asn to L-Asp, albeit at high concentrations of substrate (10mM) (J. Crawford, personal communication, February 16, 2018).

However, no transamidination product was observed when added the ethylester of an aminomalonyl-extended linear precolibactin to the reaction mixture, and rather the product was degraded. Current investigations are focused on a SNAC-thioester derivative of the same compound.
We also speculate that ClbL might indeed not involved directly in colibactin biosynthesis, rather it contributes to colibactin transport mechanism and/or in releasing toxic ammonia by pks* E. coli to effect host intestinal epithelial cell viability. This could possibly occur by two mechanisms. First, by allowing excessive ammonia permeation through simple diffusion or through N-methyl-D-aspartate (NMDA) channels as observed in Helicobacter pylori infections. While H. pylori and many viable intestinal organisms (bacteroides, bifidobacteria, clostridia, Proteus spp, and Klebsiella spp) produce ammonia through urease activity, E. coli (the dominant Gram-negative aerobic bacilli in the intestines of most subjects) does not possess such urease activity. Instead, E. coli releases ammonia by deamination of substrates other than urea, including L-Asn or L-Gln. A recent report highlighted GRIN2D (also known as NMDAR2D, NR2D, GluN2D), a subunit of a NMDA receptor and facilitator of cellular calcium influx, as a putative vascular endothelial target in CRC. GRIN2D expression was shown to be specific to CRC and targeted knockdown studies in vitro demonstrated a role for GRIN2D in endothelial function and angiogenesis. Thus, the ClbL dependent ammonia release might play a role in NMDA-dependent transport of colibactin/colibactin-producing bacteria (via endocytosis) to induce host-DNA damage.

Second, the ClbL-dependent release of ammonia might affect the pH of the environment of the epithelial cells, and the altered pH would benefit the survival of bacteria. These micro environmental changes can further aid in recruitment and survival of pathogenic bacteria near the site of infection. The basis for this hypothesis is from a recent clinical report from the Bonnet group. In this report, authors have suggested that the colonization of colonic mucosa by pathogenic E. coli could be involved in the
development of colorectal cancer (CRC), especially through the production of genotoxins such as colibactin and/or by interfering with the DNA mismatch repair (MMR) pathway that leads to microsatellite instability (MSI). The authors also suggested that based on the CRC phenotype, pks+E. coli could be a “driver” bacteria or “passenger” bacteria to favor colonization of pathogenic strains, leading to carcinogenesis. Thus, the ClbL-dependent ammonia release could possibly favor the colonization and survival of pathogenic/ pks+E. coli bacteria to release the genotoxin and damage the host DNA (Figure 4-6).

Figure 4-6. ClbL-dependent release of ammonia alters pH of the epithelial microenvironment and favors colonization by pathogenic bacteria.

4.6. Structural Homology Study of ClbL

While the above hypotheses for ClbL activity are based on sequence homology studies and literature reports, our structure-based homology study of ClbL provided intriguing results. The structure- based search of ClbL revealed significant homology to the fatty acid amide hydrolase (FAAH). FAAH is an enzyme responsible for catabolism of several primary- and secondary- fatty acid amide bioactive lipids and has been identified as a potential therapeutic drug target for the treatment of wide range of clinical disorders including pain, sleep and inflammation. A wealth of potent, selective, and efficacious inhibitors of FAAH have been reported, many of which have been utilized to
mechanistically and structurally characterize the enzyme.\textsuperscript{119} Unlike eukaryotic FAAHs, ClbL is not transmembrane anchored, however it possesses the canonical catalytic triad composed of the Ser-Ser-Lys motif.

Figure 4-7. Structural alignment of ClbL with humanized rat FAAH-1. A) Overall structural alignment of ClbL (cyan and magenta) with FAAH-1 (gray). B) Catalytic triad of ClbL, K80-S155-S179. Sidechains of active site residues are represented as sticks and interatomic distances (dashed lines) are noted in green for FAAH-1 and in magenta for ClbL.

Structural alignment of ClbL monomer with humanized rat FAAH-1 structure (PDB code 2wj1\textsuperscript{120}) revealed significant homology (Figure 4-7A). The co-crystal structure with the ligand (PDB code 2wj1) captured the active conformation of the enzyme.\textsuperscript{120} Alignment of active site residues of ClbL and FAAH-1 and measurement of interatomic distances suggests ClbL crystalized in a partially active conformation where the S155 takes another rotamer conformation (Figure 4-7B). Consistent with the observation from structural alignment study, S179 in ClbL is an active nucleophile and acts via a covalent mechanism by attacking and hydrolyzing the fatty acid amide bond through a tetrahedral intermediate.
In addition, we also compared our 3.0 Å X-ray crystal structure of ClbL to an automated protein structure homology-model (Figure 4-8) generated by the SWISS-MODEL server. Structural comparisons revealed that the missing regions in the X-ray crystal structure of ClbL are highly disordered loop regions, and few of which are covering the surface of the active site. We used the predicted ClbL model to compare structural homology with FAAH-1 (PDB code 2wj1\textsuperscript{120}), GatA (PDB code 4wj3\textsuperscript{122}) and an amidase enzyme (3al1\textsuperscript{123}). The PDB sequences of these enzymes were aligned with ClbL using MAFFT\textsuperscript{124} and the alignment file is used to generate a structural alignment model from ESPript 3.0\textsuperscript{79} (Figure 4-9).
Figure 4.9. Structural homology of ClbL. Secondary structure was predicted using SWISS-MODEL and the sequence alignment was generated with MAFFT and ESPRipt 3.0. Catalytic residues are depicted with asterisks.
4.6. FAAH Activity of ClbL

Based on structural homology studies, we conducted preliminary biochemical assays to test the ability of ClbL to hydrolyze a fatty acid amide substrate. We used decanoyl-p-nitroanilide (DepNA) as a substrate to confirm ClbL’s FAAH activity. Enzymes possessing FAAH activity, when incubated with the substrate DepNA, hydrolyze the substrate and release p-nitroaniline (ε=13,500, 410 nM, yellow) compound, allowing for the rapid measurement of FAAH activity using a 96-well plate spectrophotometer (Figure 4-10). Using the previously reported colorimetric assay method,\textsuperscript{125} the activity of ClbL and mutant K80A were compared. The assays revealed that ClbL has FAAH activity with the substrate DePNA ($k_{cat}/K_m = 1.35 \times 10^5 \text{ M}^{-1} \text{s}^{-1}$), whereas its active site mutant K80A shows no activity over baseline.

The colibactin biosynthetic pathway encodes fatty acid amide metabolites and we hypothesized these compounds as potential ligands for ClbL (Figure 4-11). While all the reported precolibactins possess an $N$-terminal fatty acid chain, we initially hypothesized that ClbL cleaves the fatty acid terminal of the precolibactins, and those fatty acids
could participate in quorum sensing or signaling pathways. Very recently, the Balskus group reported a \( \gamma \)-lactone derivative 6a, produced by the \( pks \) ClbNBCH enzymes.\(^{35}\) Fatty acid acyl-homoserine lactones (AHLs) produced by certain strains of bacteria are part of quorum sensing and AHL’s are often auto-inducers (AI’s) which are involved in gene regulation and microbe-microbe/microbe-host communication.\(^{58}\) We are currently investigating whether ClbL has activity against the substrate 6a and whether this molecule has any role in ClbR-dependent transcriptional regulation of \( pks \) genes.

Figure 4-11. Possible fatty-acid amide substrates for ClbL.

Additionally, a recent report has suggested the production of lipopeptides C12-Asn-\( \beta \)-aminobutyric acid (C12AsnBABAOH) and C12-Asn-\( \gamma \)-aminobutyric acid (C12AsnGABA.OH) in \( pks^+ \) EcN (\( E. \ coli \) Nissle 1917) probiotic strain.\(^{60}\) The synthesis of C12AsnGABA.OH, a visceral analgesic, was shown to be dependent on at least three \( pks \) enzymes, ClbANB, supporting the dual role of the \( pks \) island in EcN. While GABA itself was unable to diffuse through the epithelial barrier, the addition of the fatty acid amide C12AsnOH moiety confers to GABA the capacity to diffuse and subsequently to
act on sensory neurons. Considering this observation, we also expand our investigation to verify whether ClbL has any FAAH activity against C12AsnGABAOH. While previous study on the effect of the pks island on EcN probiotic activity utilized a clbA deletion mutant,\textsuperscript{25} it will be interesting to investigate whether late stage enzymes in the colibactin pathway are essential for probiotic activity or only involved in genotoxin synthesis.

Each of these possible hypotheses discussed above are currently being investigated to understand the role of uncharacterized ClbL in colibactin biosynthesis and activity. We continue our efforts to solve structural data and conduct biological assays to decipher the role of ClbL.

4.7 Experimental Procedures

4.7.1 Cloning of ClbL and ClbL K80A Mutant

The genes encoding the ClbL protein (construct 6) were amplified by PCR from a bacterial artificial chromosome (BAC) harboring the pks genomic DNA\textsuperscript{8} using the primers listed in Table 4-2. Amplified fragments were digested with appropriate restriction enzymes and were purified using agarose gel electrophoresis and QIAquick gel extraction kit. The digests were ligated in to respective linearized expression vectors (pET28a or pET30a) using T4 DNA ligase and transformed to competent *E. coli* TOP10 cells. All plasmids were verified by DNA sequencing. A ClbL K80A mutant was generated using the quickchange site-directed mutagenesis kit (Stratagene).

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Primer Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>pET30a-clbL</td>
<td>ClbL-FW-Ndel</td>
<td>5’-CGGAGCCCATATGAGTGAGCAGAGCTAT-3’</td>
</tr>
<tr>
<td>pET30a-clbL</td>
<td>ClbL-RV-XhoI</td>
<td>5’-GGCAGCGCTCGAGGTACCTTCCCGGTACCG-3’</td>
</tr>
<tr>
<td>pET28a-clbL</td>
<td>ClbL-FW-Ndel</td>
<td>5’-CGGAGCCCATATGAGTGAGCAGAGCTAT-3’</td>
</tr>
<tr>
<td>pET28a-clbL</td>
<td>ClbL-RV-XhoI</td>
<td>5’-GGCAGCGCTCGAGGTACCTTCCCGGTACCG-3’</td>
</tr>
<tr>
<td>pET30- clbL-K80A</td>
<td>ClbL-K80A-FW</td>
<td>5’-TTGACTGTGTTGGAGTGGTCTTGTTGACG-3’</td>
</tr>
<tr>
<td>pET30- clbL-K80A</td>
<td>ClbL-K80A-RV</td>
<td>5’-GTAATCCCGTGCAACAG-3’</td>
</tr>
</tbody>
</table>

*Mutations are underlined*
4.7.2 Expression and Purification of ClbL and ClbL K80A Mutant

Expression vectors pET30a-clbL and pET30a-clbL-K80A were transformed into E. coli C41(DE3) cells and grown in 6L of LB medium containing 50 µg mL⁻¹ kanamycin at 37 °C and 155 rpm to an optical density of 0.5, equilibrated to 25 °C, induced with 0.15 mM IPTG and expressed at 25 °C for ~ 16h. Cells were harvested and suspended in 25mL of lysis buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl), and lysed with a microfluidizer, and the lysate was clarified by centrifugation at 14000g for 40min. Protease inhibitors are added following the lysis to prevent degradation of protein. The supernatant was incubated with 1mL of Ni-NTA resin (Qiagen) for 45 mins at 4 °C and washed with one column volume of lysis buffer, followed by one column volume of wash buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 25 mM imidazole). The tagged protein was eluted with elution buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 250 mM imidazole) and dialyzed overnight in buffer containing 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM βME, 10% glycerol. The protein was concentrated and injected into a Hi-Trap Q column. Protein containing fractions are collected and further purified by gel filtration chromatography (HiLoad 16/60 SuperDex-200 column, AKTA FPLC System, GE Healthcare) with 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM βME, 5% glycerol. The fractions containing protein, as detected by SDS-PAGE analysis, were pooled, concentrated, and used for crystallization experiments. ClbL K80A was purified using the similar protocol.

4.7.3 Crystallization of ClbL and ClbL K80A

To obtain a diffraction quality crystal, we repeated the crystallization trials for ClbL and ClbL K80A. Purified ClbL and ClbL K80A were concentrated to 6.8 mg mL⁻¹.
and crystallization trials were conducted using the 96-well sitting drop vapor diffusion
method with 1 μL drops containing equal volumes of protein stock and reservoir solution
of the commercial matrix screens (Wizard (Rigaku), Crystal screens (Hampton), Index
HT (Hampton) and Peg/ion HT screen (Hampton)) at 20°C. In addition, sitting drop and
hanging drop optimization trials were conducted based on previous crystal hits in 30%
PEG3000 and 100mM CHES pH 9.5 conditions. Despite multiple attempts, rescreening
and co-crystallization experiments of ClbLK80A-precolibactin complexes were
unsuccessful. As our preliminary biochemical data suggests ClbL reactivity with the
substrate DePNA and L-Asn, we will attempt co-crystallization of ClbL and ClbL K80A
with these substrates to obtain diffraction quality crystal.

4.7.4 Data Collection and Structure Determination

Initial data collection and structure determination of ClbL (as a crystallographic
dimer) was reported by Jarrod Mousa. Second diffraction data of ClbL was collected on
the beamline 21-ID-G (June, 2015) of the Life Sciences Collaborative Access Team
(LS-CAT) facility at the Advanced Photon Source (APS), Argonne National Laboratory
(Argonne, IL). Data were collected at 100° K with a wavelength of 0.9786 Å (1 Å=0.1
nm), integrated, merged, and scaled using iMosflm to a resolution of 2.99 Å. ClbL was
crystallized in the space group P3_{1}21 with cell dimensions of a = 58.83 Å, b = 58.83 Å,
c = 214.73 Å, γ=120°. The overall structure contains one monomer in the asymmetric
unit and was isotropically refined to a resolution of 2.99 Å with a final a R_{free} of 39% and
R_{work} of 29%. Previous ClbL (dimer) structure was used as molecular replacement
search model with PHASER.\textsuperscript{56} Initial refinement of the diffraction data provided a partial
residues are being built manually, and refinement is being carried out using PHENIX.REFINE\textsuperscript{87} and COOT\textsuperscript{88}. PyMOL was used for structural illustrations.

4.7.5 ClbL Fatty Acid Amide Hydrolysis Assay

Initial velocity measurements were made at increasing concentration of decanoyl $p$-nitroaniline (DepNA) substrates (5-200 $\mu$M). Reaction was initiated by addition of 5 $\mu$M of purified recombinant ClbL and ClbLK80A proteins and the reaction was incubated at 37°C for 30 min. Data points are mean ± S.D. values of specific activity from triplicate assays from single batch of enzyme purification, and plots were generated by fitting the data points into Michaelis-Menten equation using prism software version 3.0.
LIST OF REFERENCES


BIOGRAPHICAL SKETCH

Naga Sandhya Guntaka was born in Gudavalli, Andhra Pradesh, India in 1987. She attended Sri Pragna Public School, where she completed her high school and attend a two-year college at Gowtham Residential College, where she majored in chemistry, mathematics, and physics. She began her undergraduate studies in 2004 at KVSR Siddhartha College of Pharmaceutical Sciences (SCOPS) in Vijayawada, India, where she obtained her bachelor’s degree in Pharmacy. During this time, she worked on pharmacokinetics research where she studied techniques to enhance dissolution rate of commercially available tablets. She received multiple awards while at SCOPS including Outstanding Undergraduate Student Award. She later worked as an Industrial Trainee at the Dr. Reddy’s Laboratories in Hyderabad, India in Summer 2008, where she gained analytical and technical skills related to Quality Control and Quality Assurance in the Formulation Unit. She then obtained her master’s degree in Pharmaceutical Chemistry from the Andhra University in Vizag, India in 2010 and worked as an Assistant Professor at the MNR College of Pharmacy, Telangana, India.

In 2011, Sandy moved to the Eastern Michigan University (EMU), Ypsilanti, MI, where she obtained her Master of Science degree in Chemistry. During this time, she worked with Professor Cory Emal, where she worked on a medicinal chemistry project and synthesized a library of small molecules that are used as inhibitors for the target enzyme. Sandy received multiple awards while at EMU, including ACS Huron Valley Section Outstanding Graduate Student Award, Ronald M. Scott Memorial Scholarship, EMU Chem. Dept. Outstanding Accomplishment in Chemistry Award, and ACS Travel Award. After obtaining her master’s degree from EMU, Sandy moved to the University of Florida (UF) to begin her doctorate degree in Chemistry, where she worked under
Professor Steven D. Bruner. While at UF, Sandy conducted research on the natural product biosynthesis of bacterial secondary metabolites, and gained interdisciplinary research skills in structural biology, analytical chemistry, and organic synthesis. Sandy worked on multiple collaborative projects with both external and internal experts in the field resulting in 2 publications, 1 of which is highlighted in this dissertation. Sandy has been invited as a speaker and has participated at several regional and national conferences including the Gordon Research Conference on Natural Products and Bioactive Compounds in 2017. She has received several awards during her time at UF, including Grinter Fellowship, William and Arlene Reugamer Fellowship, Gordon Research Conference Travel Award, two UF travel awards, and two best poster awards. Sandy will continue her career at the Intel Corporation, Hillsboro, OR, where she will work as a PTD Module & Integration Device Yield Engineer.