2020

STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF CAZYME AND CAZYME-RELATED PROTEINS FROM: BACTEROIDES THETAIOTAOMICRON AND PORPHYROMONAS GINGIVALIS: TWO ABUNDANT COLONIZERS OF THE HUMAN MICROBIOME

James Stevenson
stev4510@mylaurier.ca

STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF CAZYME AND CAZYME-RELATED PROTEINS FROM: *BACTEROIDES THETAIOTAOMICRON* AND *PORPHYROMONAS GINGIVALIS*: TWO ABUNDANT COLONIZERS OF THE HUMAN MICROBIOME

by

James Eric Stevenson

BSc. Biochemistry and Biotechnology, Wilfrid Laurier University, 2017

THESIS

Submitted to the Department of Chemistry and Biochemistry

Faculty of Science

In partial fulfillment of the requirements for the

Master of Science in Chemistry

Wilfrid Laurier University

ABSTRACT

The human body consists of approximately 30 trillion cells, while non-human microbes that reside on and within the body outnumber human somatic cells by a factor of 1.3 – 2.3. The interplay between our cells and those of the colonizing microorganisms affect physiology in a multitude of ways, both beneficial and detrimental. Microbes found in the oral cavity, such as the Red Complex member *Porphyromonas gingivalis*, are associated with pathology, namely periodontal diseases including gum deterioration, tooth decay, and loss of underlying alveolar bone. At the other end of the gastrointestinal tract, microbes such as *Bacteroides thetaiotaomicron* are found in abundance in the distal intestine and form symbiotic relationships that are critical for the proper function of the mammalian colon. Carbohydrates are integral for the survival of both of these organisms; as sources of energy, physical anchors, and barriers against environmental stress. As such, these microbes present interesting models for the study of extremely important host-microbial relationships involving carbohydrates and pathologic or beneficial mechanisms in carbohydrate metabolism.

The successful expression and purification of four homologous putative galactosidases from *B. thetaiotaomicron*, BT2109, BT2918, BT2966, and BT3158 are reported, along with BT2857, which has been expressed and purified previously. Crystallization and X-ray diffraction analyses of three of these full-length galactosidases was achieved, along with two truncations of BT2857 and one of BT3158. Functional kinetic characterization and generation of Michaelis-Menten constants was conducted using the galactose analogs para-nitrophenyl-β-galactopyranoside and para-nitrophenyl-α-galactopyranoside. Thin layer-chromatographic analysis using a variety of galactose-containing oligo/poly-saccharides was used to analyze

degradative capability of these enzymes, along with nuclear magnetic resonance and mass spectrometric analysis to characterize the products of these degradative reactions.

The expression and purification of three proteins from *P. gingivalis*, PGN1176 (a short-chain fatty-acid dehydrogenase), PGN1459 (a putative DNA-defense and mutagenic protein), and PGN1461 (a putative spore-maturation protein, researched in collaboration with Katarina Mandić) are reported. Crystallization and preliminary X-ray diffraction of all three proteins was completed. Preliminary functional characterization of two proteins, PGN1459 and PGN1461, was conducted using a crude DNA-shift assay and thin-layer chromatographic analysis against a diverse array of (poly)saccharides, respectively.

Crystallization, diffraction, structural determination, and functional definition of each of these proteins is of critical importance in gaining insight into the roles these proteins play in either the symbiotic or pathogenic relationships between the human body and the resident organisms of the microbiome.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Michael Suits, for giving me the opportunity to work in his lab, as well as for the support and guidance he provided over the course of my research. I would also like to acknowledge Dr. Alicia Lammerts Van Buren for collaborating with me on the homologous *Bacteroides thetaiotaomicron* proteins. I would like to thank the members of my committee, Dr. Louise Dawe and Dr. Joel Weadge, for their advice, assistance, and encouragement throughout my project. Thank you to Dr. Geoff Horsman for the guidance and context he provided in researching the *Porphyromonas gingivalis* proteins. Thank you to Dr. Lillian DeBruin for her assistance and patience with me throughout my time at Laurier. Thank you to Joe Meissner for allowing me usage of undergraduate lab space and equipment through several experiments. Thank you to Maly Lab members Lana Hiscock and Zack Schroeder for all their guidance in product characterization. A big thanks to the members of Suits lab, both past and present. I would like to acknowledge Jonah Nechacov, Jacqueline Whiteside, and Ian Fernandez, for all the training and advice they provided me throughout all phases of my studies in the Suits Lab. Thanks to my colleague and friend Peter Nguyen both for all his help throughout my stay in the Suits Lab and for being up for anything outside of it. Special thanks to my collaborator and dear friend Katarina Mandić for all the physical, cerebral, and emotional sweat she poured into the PGN1461 project, as well as for being my rock throughout my time here both in and out of lab. Lastly, I would like to thank my parents for their unwavering support and belief in me throughout my endeavours here.

DECLARATION OF WORK PERFORMED

All data related to *Bacteroides thetaiotaomicron* presented and described in this thesis are the result of my own work.

Crystallization work on PGN1461 done in collaboration with Katarina Mandić.

TABLE OF CONTENTS

ABSTRACT i-ii

ACKNOWLEDGEMENTS iii

DECLARATION OF WORK PERFORMED iv

TABLE OF CONTENTS v-viii

LIST OF TABLES ix

LIST OF FIGURES x-xii

LIST OF ABBREVIATIONS xiii-xiv

CHAPTER 1. INTRODUCTION 1

1.1 Carbohydrates and Metabolic Chemistry 1

1.2 Carbohydrate-Active Enzymes and Carbohydrate-Binding Modules 2

1.3 The Human Microbiome 3

1.4 *Bacteroides thetaiotaomicron* and Gut Microflora 4

1.4.1 Polysaccharide Utilization Loci 5

1.4.2 Protein/Domains of Unknown Function 7

1.5 Periodontal Disease, *Porphyromonas gingivalis*, and the Red Complex 8

1.5.1 Putative Virulence Factors and Upregulation in Biofilm 10

1.6 Research Objectives 11

CHAPTER 2. EXPERIMENTAL TECHNIQUES 14

2.1 Bioinformatics, Construct Design, and Functional Hypothesis 14

2.2 Protein Expression and Solubility Screening 17

2.3 Purification of Proteins 18

2.4 Separation and Purity Analysis of Proteins through SDS-PAGE 20

2.5 Protein Concentration Determination through Spectral Analysis 21

2.6 Kinetic Activity Analysis of Enzymes through Spectroscopic Measurement of Nitrophenol Release from Substrate Analogs 22

2.7 Site-Directed Mutagenesis of Potential Active-Site Residues 24

2.8 Screening of Potential Carbohydrate Substrates through Thin-Layer Chromatography 26

2.9 Product Characterization Using Nuclear-Magnetic Resonance and Mass Spectrometry 27

2.10 Crystal Screening and Diffraction Optimization 28

CHAPTER 3. MATERIALS AND METHODOLOGY 30

3.1 Materials 30

3.2 Methodology and Experimental Design 32

3.2.1 Introduction into and Extraction of Recombinant Plasmids in *E. coli* Cells 32

3.2.2 Protein Expression and Initial Isolation 34

3.2.3 SDS-PAGE Analysis 36

3.2.4 Dialysis of Protein 37

3.2.5 Purification through Fast Protein Liquid Chromatography 37

3.2.6 Determination of Protein Concentration 38

3.2.7 Crystal Screening and Expansion Plating 40

3.2.8 Enzyme Activity Assays 41

3.2.9 Site-Directed Mutagenesis and Construct Truncation 44

CHAPTER 4. RESULTS 48

4.1.1 Structural Prediction for *Bacteroides thetaiotaomicron* Proteins 48

4.1.2 Model Prediction and Functional Hypothesis for *Porphyromonas gingivalis* Proteins 52

4.2.1 Expression and Purification of *Bacteroides thetaiotaomicron* Proteins 58

4.2.2 Expression and Purification of *Porphyromonas gingivalis* Proteins 64

4.3 Kinetic and Mechanistic Characterization of *Bacteroides thetaiotaomicron* Proteins 67

4.4 Crystallization and Diffraction of Selected Targets 77

4.5 Mutagenesis of BT2857C and Truncation of DUF4959 82

4.6 Preliminary Functional Analysis of PGN1459 87

CHAPTER 5. DISCUSSION 89

5.1 Structural and Functional Characterization of BT2857 and BT3158 89

5.2 Bioinformatics and Characterization of *P. gingivalis* Proteins 91

5.2.1 PGN1176 91

5.2.2 PGN1459 92

5.2.3 PGN1461 93

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS 96

6.1 *B. thetaiotaomicron* Proteins 96

6.2 *P. gingivalis* Proteins 97

REFERENCES 100

APPENDICES 109

Appendix 1: List of Bioinformatics Tools 109

Appendix 2: Amino Acid Sequences of Protein Constructs 110

Appendix 3: ProtParam Outputs for Protein Constructs 111

Appendix 4: List of Primers for Site-Directed Mutagenesis and Truncation 112

Appendix 5: Amino Acid Sequence Results for Mutant BT2857C and BT2857N Constructs 113

Appendix 6: pET21b(+) Vector with XhoI and NdeI Cut Sites Labelled for Insertion of Truncated BT2857 and BT3158 Constructs 114

Appendix 7: Crystal Diffraction Preliminary Statistics for Each Protein Construct 115

LIST OF TABLES

Table 1. Solution compositions and volumes used in the making of SDS-PAGE gels 36

Table 2. Typical Microplate Setup for pNP Michaelis-Menten Kinetic Assays 42

Table 3. Reagent volumes typically used in Phusion DNA Polymerase SOMA protocol for SDM of BT2857 constructs 44

Table 4. Typical thermal cycler temperature profile for SOMA mutagenesis of BT2857 constructs 44

Table 5. Reagent volumes used in PCR amplifications of DUF4959 in BT2857/3158 45

Table 6. Typical thermal cycler temperature profile for DUF4959 PCR amplifications 46

Table 7. Reagent volumes for DUF4959/pET21b(+) digestion reactions 46

Table 8. Reagent volumes for DUF4959/pET21b(+) ligation reactions 47

LIST OF FIGURES

Figure 1. Illustration of PULs of Interest 7

Figure 2. Example of the typical mechanism utilized by retaining glycoside hydrolases 22

Figure 3. Example of the typical mechanism utilized by inverting glycoside hydrolases 23

Figure 4. para-nitrophenyl-based substrate analogs used in the kinetic analysis of putative glycoside hydrolase enzymes 24

Figure 5. Domains of Unknown Function Spanning Proteins of Interest from BT 48

Figure 6. Phyre^2 Intensive Homology Models of *B. thetaiotaomicron* Target Proteins 50

Figure 7. RaptorX Homology Models of *B. thetaiotaomicron* Target Proteins 51

Figure 8. Illustrated BLASTp Domain Results for *Porphyromonas gingivalis* Target Proteins 52

Figure 9. TMPred Output for *Porphyromonas gingivalis* Target Protein PGN1461 54

Figure 10. Phyre^2 (left) and RaptorX (right) homology models of PGN1176 55

Figure 11. Phyre^2 (top) and RaptorX (bottom) homology models of PGN1459 56

Figure 12. Phyre^2 (left) and RaptorX (right) homology models of PGN1461 57

Figure 13. SDS-PAGE Gels for the IMAC Isolation of Full-Length *B. thetaiotaomicron* Proteins 59

Figure 14. SDS-PAGE Gels for the IMAC Isolation of Truncated and Truncated Mutant BT2857 Proteins 60

Figure 15. Chromatogram and SDS-PAGE Analysis of BT2109 FPLC Purification 61

Figure 16. Chromatogram and SDS-PAGE Analysis of BT2857 FPLC Purification 62

Figure 17. Chromatogram and SDS-PAGE Analysis of BT2857N FPLC Purification 62

Figure 18. Chromatogram and SDS-PAGE Analysis of BT2966 FPLC Purification 63

Figure 19. Chromatogram and SDS-PAGE Analysis of BT3158 FPLC Purification 63

Figure 20. SDS-PAGE Gels for the IMAC Isolation of *P. gingivalis* Proteins 65

Figure 21. Chromatogram of PGN1461 Cation-Exchange FPLC Purification Attempt 66

Figure 22. Chromatogram and SDS-PAGE Analysis of PGN1461 Anion-Exchange FPLC Purification 66

Figure 23. pH Optima Kinetic Assay of Full-Length and N/C-Terminal Truncations of BT2857 68

Figure 24. pH Optima Kinetic Assay of BT3158 68

Figure 25. Michaelis-Menten Analysis of BT2857 and BT2857 Truncations 69

Figure 26. Michaelis-Menten Analysis of BT3158 70

Figure 27. Michaelis-Menten Analysis of BT2857C Mutants 71

Figure 28. Preliminary Michaelis-Menten Analysis of BT2109 72

Figure 29. Thin-Layer Chromatography Analysis of BT2857, BT2857N, BT2857C Incubated with α-lactose 73

Figure 30. Thin-Layer Chromatography Analysis of BT3158 Incubated with α-lactose 73

Figure 31. C^{13} NMR of Lower Scrape from Lactose Digestion TLC 74

Figure 32. C^{13} NMR of Upper Scrape from Lactose Digestion TLC 75

Figure 33. 1D H^{1} NMR of Lower Scrape from Lactose Digestion TLC 75

Figure 34. 1D H^{1} NMR of Lower Scrape from Lactose Digestion TLC 76

Figure 35. HMBC NMR Spectra of Lower Scrape from Lactose Digestion TLC 76

Figure 36. Optimized Crystals and Example Diffraction Pattern for BT2857C 77

Figure 37. Initial Crystallization Hits and Hanging-Drop Optimization Attempt BT2857 78

Figure 38. Initial Crystallization Hit for BT2966 and Corresponding Diffraction Pattern 78

Figure 39. Initial Crystallization and Examples of Optimized Crystallization and Diffraction of BT3158 79

Figure 40. Crystallization of BT2857C H300A/D304A Mutant in Sitting Drop 80

Figure 41. Crystallization of PGN1176 and Sample Diffraction Pattern 80

Figure 42. Crystallization of PGN1459 and Sample Diffraction Pattern 81

Figure 43. Multiple Sequence Alignment of DUF4959/5126 Domains from Conserved Putative Galactosidases 82

Figure 44. Coordinated Ethylene Glycol Molecules in BT2857C Crystal Structure 83

Figure 45. Agarose Gel Analysis of BT2857N SOMA Mutagenesis Post-DpnI Digestion 84

Figure 46. Agarose Gel Analysis of BT2857C SOMA Mutagenesis Post-DpnI Digestion 84

Figure 47. Agarose Gel Analysis of BT2857N/BT3158-DUF4959 Amplification 86

Figure 48. Agarose Gel Analysis of pET21b(+) and BT2857N/BT3158-DUF4959 Digestions 86

Figure 49. Agarose Gel DNA-Shift Analysis of PGN1459 and Selected Plasmids 88

LIST OF ABBREVIATIONS

APS - ammonium persulfate

CAPS - N-cyclohexyl-3-aminopropanesulfonic acid

CHES - N-cyclohexyl-2-aminoethanesulfonic acid

EDTA - ethylenediaminetetraacetic acid

HEPES - 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HMBC – Heteronuclear Multiple Bond Correlation

IPTG - isopropyl β-D-1-thiogalactopyranoside

JCSG - Joint Center for Structural Genomics

MS - mass spectrometry

MSCG - Midwest Center for Structural Genomics

m/z – mass-to-charge ratio

NMR - nuclear magnetic resonance

PAGE - polyacrylamide gel electrophoresis

PCR - polymerase chain reaction

PIPES - piperazine-N,N′-bis(2-ethanesulfonic acid)

RPM - rotations per minute

S-DM - site-directed mutagenesis

SDM – sequence distance measure

SDS - sodium dodecyl sulfate

SOMA - single-oligonucleotide mutagenesis and cloning approach

TEMED - tetramethylethylenediamine

TLC - thin-layer chromatography

w/v - weight/volume

v/v - volume/volume

xg - times force of gravity

CHAPTER 1 - INTRODUCTION

1.1 Carbohydrates, Metabolic Chemistry, and Prokaryotes

Carbohydrates, also frequently referred to as either saccharides or, informally, sugars, are a type of biomolecule found in all cells in every kingdom of life (1). These molecules, ranging in size from single monosaccharide units to polysaccharides consisting of thousands of monosaccharide subunits, are involved in a multitude of biochemical processes and structural roles (2). This class of compounds can also be decorated with a variety of functional groups and serve as functional conjugates for other non-saccharide macro- or small-molecules(3, 4). Because of this enormous diversity, carbohydrates are thus ubiquitous in cells, and are essential to the metabolism and survival of any living organism (5). In these cellular locations carbohydrates can be found serving roles such as extra-/inter-cellular structure and communication (6), energy storage and transport (7), serving as the backbone for DNA (8), functionalization through glycosylation of proteins and lipids (9, 10), cell adhesion and biofilm formation (11, 12), and cellular protection (13, 14). The wide variety of functions played by carbohydrates in cellular processes makes these biomolecules, and the macromolecules that interact with them, a highly relevant area of study. Of particular interest in this study are how specific prokaryotes rely on carbohydrate biochemistry as the metabolism of these macromolecules is the preferred avenue of energy production for the vast majority of bacteria(15, 16). Polysaccharides produced and consumed by bacteria are far more varied than those of other organisms: complex carbohydrates produced by this class of organisms are approximately 10-fold and 9-fold more diverse than mammalian glycans at monosaccharide and disaccharide levels, respectively(17), and certain species have been seen to be able to live on the most complex polysaccharide yet known(18). The plethora of catalytic and non-catalytic

interactions prokaryotes have with carbohydrates include two groups of proteins: carbohydrate-active enzymes and carbohydrate-binding modules, respectively.

1.2 Carbohydrate-Active Enzymes and Carbohydrate-Binding Modules

Carbohydrate-active enzymes (CAZymes) are defined as any enzyme or enzymatic domain capable of acting on a carbohydrate substrate for either breakdown, synthesis, or modification(19). As of August 2019, the Carbohydrate-Active Enzyme Database (www.cazy.org) contains information on 165 glycoside hydrolases, 107 glycosyltransferases, 37 polysaccharide lyases, 16 carbohydrate esterases, and 16 auxiliary activity enzymes(20). Glycoside hydrolases breakdown glycosidic bonds between saccharide residues, or between a saccharide residue and a non-saccharide moiety, typically with a Koshland-style mechanism utilizing a nucleophilic amino acid general acid/base amino acid as the catalytic residues(21, 22). Glycosyltransferases function as the builders of oligo-/poly-saccharide, forming glycosidic bonds between an activated donor saccharide (containing a nucleotide- or lipid-phosphate leaving group) and a specific acceptor molecule(23, 24). Polysaccharide lyase enzymes cleave glycosidic bonds in polysaccharide chains that contain uronic acid saccharides (sugars with both carboxylic acid and carbonyl functional groups) via β-elimination, yielding a saturated and unsaturated hexenuronic acid product(25). Carbohydrate esterase enzymes catalyze to removal of O- and N-acetyl groups on substituted saccharides, and a number of different mechanisms may be utilized(26). The auxiliary activity group of CAZymes are unique in that members of this family contain lytic redox enzymes/enzymatic domains that break-down polysaccharides such as cellulose, and in some cases the complex organic polymer lignin(27). The database has also classified 85 distinct types of carbohydrate-binding module: a domain, typically appended to enzymatic domains within a larger protein sequence, that can bind saccharide substrates but on their own lack catalytic activity(28).

CAZymes are distinguished from one another and grouped into families on the basis of primary sequence and structural similarity, rather than by substrate specificity or mechanism of action(29). Due to the method of grouping utilized for CAZymes, substrate specificity, catalytic mechanisms, and kinetic profiles may vary quite a bit even between enzymes that belong to the same family. Thus, biochemical characterization is needed to truly determine the functional character of any CAZyme(30). Characterization, both structurally and functionally, of CAZymes is critical to understanding the litany of carbohydrate and carbohydrate-related processes and interactions within the biosphere, understanding which can be applied to fields such as agriculture, biofuel production, and medicine(31).

1.3 The Human Microbiome

The human microbiome is defined as the collection of all organisms, symbiotic, commensal, and pathogenic, that reside within and on the human body(32). Interest in the human microbiome has increased drastically since the publication of the initial draft of the human genome by the International Human Genome Sequencing Consortium in 2001(33) as the role this collection of organisms plays in human health has become more and more apparent(34). The initial estimates for the number of protein products the human genome encoded for were between 70,000 and 100,000 based on the number of biological functions necessary to maintain life as well as extrapolation from CpG island counts in previously sequenced gene fragments and the frequency with which they are associated with known genes(35, 36). After the completion of the Human Genome Project, a new estimate of 30,000 - 40,000 genes was made, and the current estimate has decreased that to only 21,306 protein-coding genes(37). Many of the microorganisms that constitute the human microbiome help to make up for this apparent functional metabolic deficit in the human genome(38). Only recently have biotechnological tools and techniques allowing for the

study of the complex microbial communities that make up the human microbiome been developed, studies which have revealed that gaps in the current understanding of host-microbial interactions are larger than previously thought(39). These complex microbial communities that make up the microbiome are not uniformly dispersed, but rather clustered in certain sections of the human body, such as the intestinal tract and oral cavity(40). Collectively, these microbial clusters house trillions of microorganisms representing thousands of different species and strains, with these organisms estimated to outnumber human cells by a factor of 1.3 - 2.3(41, 42). Many of the organisms have also been identified as abundant sources of CAZymes, enabling them to survive in the unique polymicrobial environments which they are found in(43, 44). Some of these organisms may dedicate up to 6% of their genome to coding for CAZymes, a significant increase over the typical 1 – 2% seen in other organisms(45). Being so numerous and playing such a prominent role in human physiology, characterizing the genomes of these organisms and protein products for which they encode is essential to further the understanding of human-microbiota relationships, as well as providing a source of novel CAZymes for potential application in biotechnology and industry(46).

1.4 *Bacteroides thetaiotaomicron* and the Gut Microbiome

The human gut microflora is a diverse group of microorganisms that live within the intestinal tract(47). The ecology of this particular microflora is affected by several factors such as age, diet, and host genotype(48). *Bacteroides thetaiotaomicron* is a major component of this group, accounting for a total of 6% of all microorganisms and 12% of all *Bacteroides* species in the distal intestinal tract, by population(49). *B. thetaiotaomicron* and other mutualistic organisms within the gut, play several roles in regulating the proper function of the intestine, including: formation and maintenance of the protective mucosal barrier, enhancement of nutrient absorption, host immune activation, and stimulation of the local circulatory system(50–52). Previous studies involving the

colonization of germ-free mice with *B. thetaiotaomicron* illustrated *in vivo* how critical the organism, and indeed the rest of gut microflora, are to proper physiology, with mice experiencing fold increases in host expression between 2.6 and 205 of multiple protein markers associated with colonic health following colonization(53). Mice with a healthy microflora also required 30% fewer calories to maintain body weight compared to their germ-free counterparts(54). Short-chain fatty acids produced by gut microbes through fermentation of undigested carbohydrates is estimated to contribute up to 10% of a person's daily calories, depending on the composition of an individual's diet and microbiome(55). In addition to being vital for the proper expression of host gene products, the intestinal (and human) microflora contributes a great deal of metabolic functionality, making up for the deficit in the human genome(38). *B. thetaiotaomicron* is one of the only organisms know to be capable of surviving on exclusively the most complicated carbohydrate known, rhamnogalacturonan-II, thanks to its large array of CAZyme and CAZyme-related genes(18). The degradation, modification, transportation, and formation of carbohydrates by *B. thetaiotaomicron* is strictly controlled, and CAZymes are organized into discreet clusters on the genome known as polysaccharide utilized loci(56).

1.4.1 Polysaccharide Utilization Loci and the *B. thetaiotaomicron* Genome

The tightly regulated linear clusters of genes known as Polysaccharide Utilization Loci (PULs) are found throughout the genome of *B. thetaiotaomicron*(57). A PUL is a single gene locus which contains all genes necessary for the association with a polysaccharide at the cell surface, the subsequent processing of this polysaccharide into large oligosaccharides, followed by the importation of these oligosaccharides into the periplasmic space and/or cytoplasm, and lastly, the degradation of the oligosaccharides into monosaccharide components. Each PUL also contains a regulatory gene necessary for modulation of its activity(58). The VPI-5482 *B. thetaiotaomicron*

strain contains a total of 88 of these utilization loci, encoding for a total of 866 genes and making up 18% of the genome(59). Of the organisms that comprise the human microflora, and in fact all organisms, the genome of *B. thetaiotaomicron* codes for the expression of the largest number of CAZymes; comprising 279 glycoside hydrolases, 87 glycoside transferases, 15 polysaccharide lyases, 27 carbohydrate-binding modules, and 19 carbohydrate esterases identified to date(45). Despite the significance of *B. thetaiotaomicron* in relation to both the human microflora and glycomics, it remains poorly studied, with 24% of its genes not being significantly similar to other characterized proteins in public databases, and 18% that are related to proteins or domains of unknown function(60). The *B. thetaiotaomicron* enzymes which are the focus of this study are all found in predictable positions within various PULs(61). Complete characterization, or at least functional annotation of these protein products is critical to understanding biological processes and the regulation thereof in the PULs of *B. thetaiotaomicron* and would contribute to knowledge of the relationship between this organism and the human body(62). The PULs of interest in this study can be seen below in Figure 1.

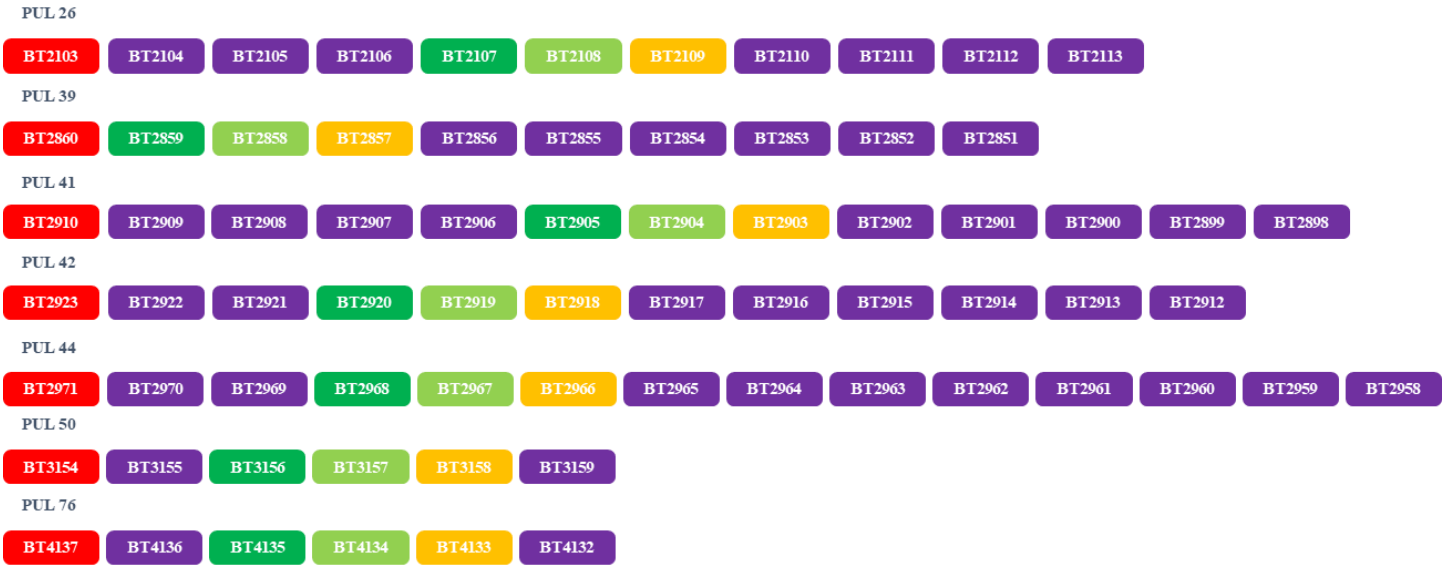


Figure 1. Illustration of PULs of Interest. Red cells indicate genes coding for two-component histidine kinase response regulator proteins, dark green cells for Starch Utilization System Type-C (SusC) proteins, light green for Starch Utilization System Type-D (SusD) proteins, yellow for the proteins of interest, and purple cells for the miscellaneous other members of the polysaccharide utilization loci. Notably, the targeted factors always are preceded by the SusC/D in these PULs.

1.4.2 Protein/Domains of Unknown Function

As DNA sequencing tools and technologies have become progressively less expensive over the past two decades the number of organisms with fully-sequenced, published genomes has increased exponentially; from only 35 in the year 2000(63) to nearly 60,000 by 2015(64). As of writing, the Joint Genome Institute (Walnut Creek, CA, USA) Genomes OnLine Database (GOLD) has accepted a total of 361,646 genome sequences(65). The UniProt/TrEMBL database contained slightly more than 230,000 protein sequences in 2000(66), with this increasing to 89 million protein sequence entries at the start of 2015(67). Functional annotation of these genomes and proteins lags far behind, resulting in proteins and domains of unknown function outnumbering characterized ones in nearly all categories, and with annotations often circulated through different genomes without any experimental or gene context validation, misannotation is a common problem(30). Domains of unknown function (DUFs) are a major contributing factor to the issue of

the deficit in functional attribution, structural characterization, and misannotation of genomes, comprising approximately 22% of all families in the Pfam database(68). As the functional and structural units of proteins are domains, and each domain tends to be distinct in these features, experimental validation is required for each individual domain's characterization(69). For these reasons genomic, structural, and biochemical analysis of gene products is required to truly determine the characteristics of almost any enzyme, particularly those containing DUFs(70), including the proteins of interest selected for this study which is predicted to encode for three DUFs. The problem of the prevalence of DUFs throughout genomes and gaps in our knowledge with respect to proteins essential for the survival of a plethora of organisms has received significant attention in the field of biochemistry as the number of sequenced genomes has ballooned in recent years(71). As such, in addition to identifying CAZymes potentially useful in biotechnological applications and adding context to the biological processes necessary for the survival of a vital microbial symbiont, characterization of these enzymes will add to the collection of knowledge on protein motifs and functional, as well as decrease the total number of domains of unknown function in genome databases.

1.5 The Oral Microbiome, Periodontal Disease, and *Porphyromonas gingivalis*

Like the gut, the human oral cavity also contains a high concentration of bacteria, being the second most diverse microbial community in the body with over 700 species of microorganisms colonizing the various tissues(72). Like the gut microbiome, the oral microbiome maintains a critical homeostasis, and dysbiosis in this community has a deleterious effect on the health of both the symbiotic microorganisms of the oral microbiota and the host organism(73). Periodontal diseases commonly arise as a result of oral microbiome dysbiosis(74) and are pervasive across nearly all demographics; estimated to have affected at least 3.58 billion

individuals globally in 2016(75), with severe periodontitis resulting in tooth loss occurring in 10 - 15% of the world population(76). The disease is not limited to just underdeveloped nations: over 50% of American adults are estimated to be affected by the disease(77). Economically, periodontitis is responsible for approximately $54 billion globally in lost productivity and a significant portion of the approximate $442 billion per annum cost of treating oral diseases(78). Periodontal diseases are polymicrobial in nature and the pathogenic organisms responsible for periodontal disease, such as those of the "Red Complex", are also a component of the human microbiome(79, 80). The Red Complex is a collection of three organisms: *Treponema denticola*, *Tannerella forsythia*, and *Porphyromonas gingivalis*, each of which are critical keystone organisms in the development of advanced periodontitis(81). The organism of interest in this study is *P. gingivalis*, an organism with a species name that shares lineage with the condition called gingivitis, for the tissue in which both are found. It is the organism most strongly associated of all three in progression of the condition, being found in upwards of 85% of all subgingival plaque sites(82, 83). The organism is a rod-shaped, non-motile, gram-negative, obligate anaerobe(84) highly capable of adapting to the dynamic environment of the oral cavity and colonization of plaque formations and host epithelial tissues(85). An important part of environmental adaptation for *P. gingivalis* is modulation of the expression of genes in response to growth in a polymicrobial biolfilm(81). Genes that are found to be upregulated in *P. gingivalis* during colonization of this biofilm are of biological significance as they can be implicated in the survival and pathogenesis of not just the organism itself, but also other organisms that exist within the biofilm as well(86). These genes are often referred to as "putative virulence factors"(87).

1.5.1 Putative Virulence Factors and Upregulation in Biofilm

Critical to the colonization and pathogenesis of *P. gingivalis* in the oral cavity is the formation of a carbohydrate-rich biofilm(88). Establishment and colonization of this biofilm, as well as invasion of host cells and tissues, requires a change in the expression of metabolic and regulatory genes in the organism(82, 88), as well as the genes necessary for mitigation of the immune response from the host organism(87, 89). Transcriptomic analysis was conducted on *in vitro* co-cultures of the Red Complex organisms during the establishment of biofilms, in order to identify genes that are upregulated relative to expression in planktonic cells(90). In the context of microbial-host colonization, genes that were found to be upregulated in the biofilm can be defined as biologically-significant factors, and characterizing the protein products may provide insight into the pathogenic processes of *P. gingivalis* during this important colonization stage(91). Several of these biologically-significant factors have been found to be putative CAZymes or CAZyme-related proteins, or putatively exported factors(90). While it has long been known the organism is not adept at utilizing carbohydrates as an energy source(92) (with some going so far as to classify it as asacchrolytic(93, 94)), the genome of *P. gingivalis* encodes the genes necessary to conduct glycolysis; as well as genes for the transport of glucose/galactose and degradation of complex amino sugars(95). Other genes, often found to be upregulated during biofilm colonization, include those related to glutamate metabolism, iron uptake, and cellular adhesion(84). The proteins selected for study herein either have putative roles in, or are related to proteins with roles in, these three biological functional. Elucidating the structure and function of these upregulated proteins will assist in not only building a better model of *P. gingivalis* biofilm formation, but also reveal potential therapeutic targets(96, 97).

1.6 Research Objectives

The main goal of the *B. thetaiotaomicron* project is to structurally and functionally characterize the conserved putative galactosidase enzymes in the selected polysaccharide utilization loci, and to assign function to the domains of unknown function present in these enzymes. The outcomes of protein characterization will then be submitted included in the CAZy database in order to add to the collective knowledge of carbohydrate-degrading enzymes and classify the as-yet unknown function of the domains within these proteins, as well as increase understanding of the metabolism and glycobiology of *B. thetaiotaomicron.* It was hypothesized that the protein would exist as two globular domains roughly split in half by sequence length, and that C-terminal domain would contain the catalytic region of the protein.

Aim one for the BT project was to achieve the preliminary crystallization of the conserved galactosidases BT2857, BT2966, and BT3158. This step was achieved through cloning, expression, and analysis of the proteins for solubility and crystal growth in sitting-drop additive screens covering a range of different buffered pH values, salts, and precipitating additives (including sugars, alcohols, and various polyether compounds). After preliminary crystallization was achieved using these sitting-drop screens, optimization of conditions were pursued using hanging-drop 24-well expansion plates. Crystals deemed of an appropriate size and morphology were cryoprotected, flash-frozen in liquid nitrogen, and sent to the Canadian Synchrotron Light Source for X-ray diffraction analysis in order to determine the 3D structure of the enzymes.

Aim two for the BT project was to evaluate the functional hypotheses generated through bioinformatics analysis of the proteins through kinetic assays utilizing substrate analogs and through screening of possible carbohydrate substrates through thin-layer chromatography. Kinetic assays were used to generate reaction rate constants for each of the enzymes. Once substrates that

the enzymes can degrade had been identified, isolated samples of the product were then analyzed via Nuclear Magnetic Resonance (NMR) and Mass Spectrometric (MS) analysis.

The third aim of the BT project was to determine the active domain of the enzymes through truncation of the genes and to determine the active-site residues within these domains utilizing site-directed mutagenesis and gene subcloning. Kinetic assays utilizing substrate-analogs were then employed to assess catalytic activity in the truncated and mutated enzymes, and crystallization trials were undertaken in an effort to attain enzyme structures complexed with substrate. The activity (or the absence thereof) in mutant enzymes/truncations can be used to infer which domain is most likely responsible for catalysis, as well as the active residues within those domains.

The goal of the *P. gingivalis* project was also to determine the structure and function of the putative virulence factors PGN1176 and PGN1461, as well as the PGN1461-related protein PGN1459. Classification of *P. gingivalis* proteins as putative virulence factors was determined previously by transcriptomic analysis, which showed upregulation of these genes in *P. gingivalis* during the co-colonization and localization of these organisms in the biofilm(98). PGN1459 was determined to be another protein of interest due to its close location to the gene encoding PGN1461 in the *P. gingivalis* genome. Proteins successfully characterized can then be used to form a more detailed model of pathogenicity for *P. gingivalis* as it relates to oral disease progression, and structural/functional information can then be used to both determine if the proteins are possible therapeutic targets and provide a starting for therapeutic development.

The first aim for this part of the project is to achieve the preliminary crystallization of the proteins PGN1176, PGN1461, and PGN1459 through screening with a range of buffered pH values, salts, and precipitating additives in 96-well sitting-drop plates. Optimization of these preliminary conditions were pursued using both 24-well hanging drop and 96-well sitting-drop

plates. Diffraction-quality crystals were then cryoprotected and sent for X-ray diffraction and eventual structural solution, and iterative rounds of model building, refinement and validation.

CHAPTER 2 - EXPERIMENTAL TECHNIQUES

2.1 Bioinformatics and Functional Hypothesis

The FASTA-format amino acid sequences of the selected proteins targets were taken from the *National Center for Biotechnology Information* (NCBI) Protein database, and analyzed through a series of bioinformatics resources in order to help predict the structure and function of the protein, as well as determine certain parameters such as size and theoretical isoelectric points that will assist in separating and identifying proteins throughout purification.

Gene codes for proteins of interest were searched for in the ELIXIR UniProt database in order to view any putative functional annotation the target may have, and to provide easy access the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) entry for the protein. The KEGG is a genome annotation database (https://www.genome.jp/kegg/) that uniquely allows for higher-level functional speculation through its unique classification architecture and automated workflow. Molecular functions of proteins are associated with ortholog groups so they can be associated across organisms, while separately these ortholog groups are defined in the context of cellular pathways and molecular relation networks(61). All members of the PUL from which the genes of interest belong can be studied to fit the proteins into the broader genomic context of the loci, and to try and ascertain the specific substrate being targeted(99). The amino acid sequences (found in Appendix 2) will then be submitted to the STRING database (https://string-db.org/), which generates a network of protein-protein interactions, physical and genomic, based on: experimental co-expression analysis; detecting shared selective signals across the genome of the organism; text-mining of scientific literature; and comparing interaction knowledge between organisms with gene orthology(100). Combined with the user interface of the STRING tool, it allows for more genome-wide detection of potential interactions, which will complement the information gathered from the

KEGG analysis of the PULs present in BT. Proteins that are present and ordered into PULs, as determined by KEGG, can be compared to the genome-wide results generated by STRING, confirming the makeup of the KEGG PULs and potentially connecting the PULs of interest to other pathways, providing more information that may help with determining the substrates and/or roles of these polysaccharide-active gene clusters.

The amino acid sequences of target genes are then submitted to the ExPASy ProtParam server to analyze the physical and chemical characteristics of the proteins. Output from this analysis tool provides information such as: molecular weight, amino acid composition, theoretical pI, 280 nm extinction coefficient, instability and aliphatic index scores based on the amino acid sequences(101). The information on physiochemical properties can be applied in the identification, solubilisation, purification, concentration, and storage of protein samples. Likely sites of subcellular localization for the proteins were determined using the online PSORTb server, a resource which analyzes the protein sequence for signal peptides, amino acid composition, transmembrane helices, structural/sequence motifs, and homology to proteins of known localization in the UniProt and Swiss-Prot databases(102). Information on putative subcellular location will provide insight into not only where the proteins will be located after expression, but also how these proteins may function and what their role is in their respective polysaccharide utilization pathways. The ExPASy online tool TMpred were used to scan each protein of interest for transmembrane domains. The tool searches for homology in submitted sequences against experimentally determined transmembrane topologies in the Swiss-Prot database(103). Identification, or lack thereof, of transmembrane domains will allow for an even more accurate picture of protein location and function to be derived. The SignalP4.1 server was used to search for the presence of signal peptide sequences in each protein as well(104), allowing for a more

accurate determination of what the amino acid sequence will be *in vivo* following translation and folding of the protein; information which can be used to obtain both a better preliminary structure and fit for residues in the structures determined by X-ray diffraction. Conserved domains for each protein will be investigated using the both the NCBI Basic Local Alignment Search Tool - protein (BLASTp) and the EMBL-EBI InterPro online resources. Both tools seek to perform the same task: accurately identify conserved domains in amino acid sequences, which may suggest a certain function and/or structural feature of that protein. The main difference is the method of searching, with BLAST utilizing multiple-sequence alignments of the query sequence with groups of previously grouped conserved domains and calculating a position-specific score matrix score(105), while InterPro utilizes an extra step where a Markov Matrix is generated and a node map based on domain similarity is constructed, with the "flow" through nodes being used to most similar domains(106). Domains determined to be similar to those of the target proteins may shed light on the functional/mechanistic/structural properties of these proteins.

Putative 3D structural models of each protein will be generated with the online tools Phyre^2 and RaptorX. These resources utilize initial multiple sequence alignments of the query protein with template proteins in the PDB possessing experimentally determined structures followed by Hidden-Markov Modelling to assess query-template fit, but differ in template selection parameters and the subsequent iterative processes used in structural refinement. Phyre^2 (Protein Homology/analogY Recognition Engine V2.0) refines possible structures by first modelling a crude backbone consisting of conserved regions, then adds loop and connective regions between domains, and finally adding side chains in the lowest energy configuration(107). Query models are constructed based on single templates, and each different query-template model is ranked based on total coverage of the query sequence. RaptorX on the other hand integrates local and global

structural context to better match distantly-related proteins or those with a low level of sequence conservation(108).

Multiple sequence alignments of the galactose-binding like-domain protein sequences from all homologous 4959/5126/5000 BT proteins, as well as previously characterized proteins from other organisms, will be conducted using the online EMBL-EBI tool Clustal Omega, which shows comparable accuracy to slower intensive alignments while being considerably faster and more consistent than previous rapid methods(109). These alignments can be useful in identifying conserved residues across the homologous BT proteins and orthologs found in other organisms, from which common sugar-binding/catalysis motifs such and aspartate-x-aspartate/glutamate (D-X-D/E)(110–112). These residues can then be visualized in either the homology- or experimentally-modelled 3D structures of the target proteins to determine whether these residue side chains are solvent-facing and capable of interacting with substrate.

2.2 Protein Expression and Solubility Screening

Initial attempts to isolate proteins from pelleted cell cultures were conducted through the solubilisation of cells in a lysis solution containing a high concentration of sucrose (to initiate osmotic shock), as well as sodium chloride and Tris:HCl buffer at pH 8.0. After homogenization in the lysis buffer a detergent solution containing a low concentration of deoxycholic acid and Triton X-100, along with sodium chloride and pH 8.0 Tris:HCl buffer, is added in order to stabilize recombinant proteins. Lysis suspensions were centrifuged to pellet the insoluble fraction and yield a supernatant that may contain the proteins of interest to be purified. If the protein of interest is found in the pellet, rather than the supernatant, protein solubility conditions were screened for alternative lysis conditions. Solution components that were varied include: buffering compound, pH, salt type and concentration, and detergent conditions. Stock solutions containing high

concentrations of these adjustable components were used to create a number of varied conditions against which solubility can then be tested. To perform this assay, small amounts of cell pellet are introduced into these custom conditions and sonicated for a short period at an ultrasonic frequency to homogenize the pellets and disrupt cell walls/membranes. Solutions were then centrifuged and analyzed for solubility of the recombinant protein.

If the protein was neither found in the pellet or the supernatant, the variation of expression conditions were then tested. Conditions that can be altered include: makeup of media (ie. salts, yeast extract amount, tryptone amount), media volume, induction temperature, usage of autoinduction, and type of cellular expression system (ie. Tuner(DE3), BL21-CodonPlus).

2.3 Purification of Proteins

The techniques used to isolate and purify proteins during the course of these studies were: immobilized-metal affinity chromatography (IMAC), anion-exchange chromatography, and cation-exchange chromatography.

The form of IMAC employed was nickel-nitrilotriacetic acid (Ni-NTA), which employed a resin designed to separate polyhistidine-tagged (6xHis) recombinant proteins from crude cell lysate. Crosslinked agarose beads make up this resin, to which alkyl linkers are covalently bound. Affixed to the end of these alkyl linkers are nitrilotriacetic acid molecules, which act as quadridentate chelators for nickel(II) cations, effectively immobilizing the metal ions on the resin. The two remaining coordination sites present on the immobilized nickel cation may then interact with free electrons on the nitrogen atoms of the histidine side chains of C- or N-terminally 6xHis-tagged recombinant proteins. Once cell lysate has been applied through the resin, buffer was run through the resin in order to wash residual crude extract proteins, and prevent the agarose beads from damage as a result of drying out. Buffer containing incrementally higher concentrations of

imidazole was then run through the resin in order to release the recombinant protein in a stepwise manner. Imidazole is a cyclic aromatic molecule containing two non-adjacent nitrogen atoms which competes with the 6xHis tags on recombinant proteins for nickel coordination sites. The higher affinity of imidazole for these coordination sites causes the bound protein to elute off the resin. Theoretically, Ni-NTA IMAC should be a "one-step" purification method, as non-recombinant proteins lack a flexible 6xHis tag, however in practice significant amounts of contaminant proteins may be present in fractionated samples due to protein-protein interactions, non-specific protein-bead interactions, and/or presence of contaminants with charged surface residues capable of coordinating to nickel(II). To remove these contaminants prior to fractionation, the resin is washed with a comparatively large amount of buffer containing a low concentration of imidazole.

Anion-exchange and cation-exchange chromatography are methods of purification that isolate proteins based on a positively or negatively charged resin, respectively, interacting with proteins that have a net complementary charge. In anion-exchange, the resin is positively charged across an operational pH range of 2 to 12 and attracts negatively-charged proteins due to the presence of a quaternary amine connected to crosslinked agarose beads via an alkyl linker. Positively-charged buffers, such as Tris, are used in anion exchange to prevent interactions with the resin and reduce the potential for inhibition of protein binding. The pH of the chosen buffer must be greater than the isoelectric point (pI) of the protein of interest, typically by at least 0.5 – 1 pH units, while remaining ~0.6 pH units from the pKa of the buffer in order to both effectively buffer the solution against pH change and negatively charge the protein. In cation-exchange, the crosslinked agarose resin contains alkyl linkers with terminal sulphonyl groups that are negative across a pH range of 2 to 14 and attracts positively-charged proteins. In the case of cation-

exchange, negatively-charged buffers, such as HEPES, are preferred. The pH of the buffer must be below that of the pI of the protein of interest, again by at least 0.5 – 1 pH units and within ~0.6 pH of buffer pKa. In both anion- and cation-exchange, proteins can be released from the resin through the gradient addition of buffer containing a high concentration sodium chloride, which contains both positive and negative ions that outcompete the proteins for the charged sites on the resin, causing the protein of interest to be eluted out.

2.4 Separation and Purity Analysis of Proteins through SDS-PAGE

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a technique that can be used to separate proteins based on molecular weight. SDS-PAGE gels generally consist of a stacking gel, typically containing 4 – 6% (bis)acrylamide, layered over a resolving gel that ranges between 8% and 12% (bis-)acrylamide for separation of mid-size proteins or 14 – 18% for small proteins. The typical pH of the stacking and resolving gel sections are 6.8 and 8.8, respectively. Crosslinking and polymerization of acrylamide and bis-acrylamide in solution is catalyzed by the addition of tetramethylethylenediamine (TEMED) and ammonium persulfate (APS). The stacking gel contains the wells in which protein samples are deposited and a short strip of gel beneath these wells that serves to align the entire protein sample in as compact a band for uniform introduction into the resolving gel. Before running protein samples on SDS-PAGE gels, a loading buffer solution containing glycerol, sodium dodecyl sulfate (SDS), bromophenol blue, and dithiothreitol (DTT) is added to the protein sample. Glycerol increases sample density easing loading of sample into wells; SDS assists in denaturation of, and coats, proteins in the samples with negative charge allowing for effective sample migration towards the anode; bromophenol blue allows for easy visualization of samples during application, and DTT reduces disulfide bonds in order to further denature proteins in samples. Protein/loading buffer mixtures are heated to

accelerate denaturation of the sample. Total denaturation of the proteins in samples is vital to the success of SDS-PAGE analysis, as any preserved secondary, tertiary, or quaternary structure will affect the migration rate of the protein through the pores of the resolving gel and may result in erroneous size or purity determinations being made. Gels are removed from glass casts after running and submerged in Coomassie blue stain (containing Brilliant Blue R250 dye) before being heated in order to visualize the protein(s) present. Stain is then decanted and gels are rinsed with water, then heated in a destaining solution to remove excess stain that has been absorbed into the gel.

2.5 Protein Concentration Determination through Spectral Analysis

Absorbance spectroscopy is an analytic technique that can be employed to determine the concentration of protein present in a sample, based on the intrinsic fluorescent properties of the particular protein being analyzed. The three amino acid residues that are considered when quantifying the intrinsic fluorescent property of a protein are tryptophan (W), tyrosine (Y), and phenylalanine (F), with maximal absorption wavelengths of 280, 274, and 257 nm, respectively. Cysteine (C) residues may in some cases oxidatively dimerize and form disulfide-linked cystine residues that somewhat contribute to the intrinsic fluorescence of a protein as well. The absorbance of a protein sample is typically measured at 280 nm due to tryptophan having the highest quantum yield (alternatively, having the lowest ratio of emitted electromagnetic radiation to absorbed radiation) of the considered residues, and its absorbance spectra overlapping with that of tyrosine and phenylalanine. The extinction coefficient for a given protein at 280 nm can be calculated using the online bioinformatics tool ProtParam (by ExPASy), requiring only the amino acid sequence of the protein. To practically apply this technique, a small amount of concentrated protein sample can be diluted into a measured amount of buffer or water and absorbance at 280 nm (A_{280}) can be

determined. The Beer's Law equation ($A = \varepsilon \cdot c \cdot l$) is then applied using the predicted extinction coefficient and known sample pathlength to calculate the concentration of the diluted protein. The original concentration of the undiluted protein sample can then be calculated, usually expressed in units of micromolarity (μM). This micromolar measurement can also be converted to mg/mL using the calculated molecular weight of the protein.

2.6 Kinetic Activity Analysis of Enzymes through Spectroscopic Measurement of Nitrophenol Release from Substrate Analogs

Optimal operating pH/temperature, kinetic rates, and Michaelis-Menten constants for certain proteins may be determined through the liberation of a nitrophenol indicator molecule from a substrate analog. This type of assay can be especially valuable for assessing the activity, and to some degree stereochemistry, of glycosidase enzymes where the mechanism involves the cleavage of O-glycosidic bonds. Examples of prototypical retaining and inverting hydrolase reactions can be seen in Figures 2 and 3.

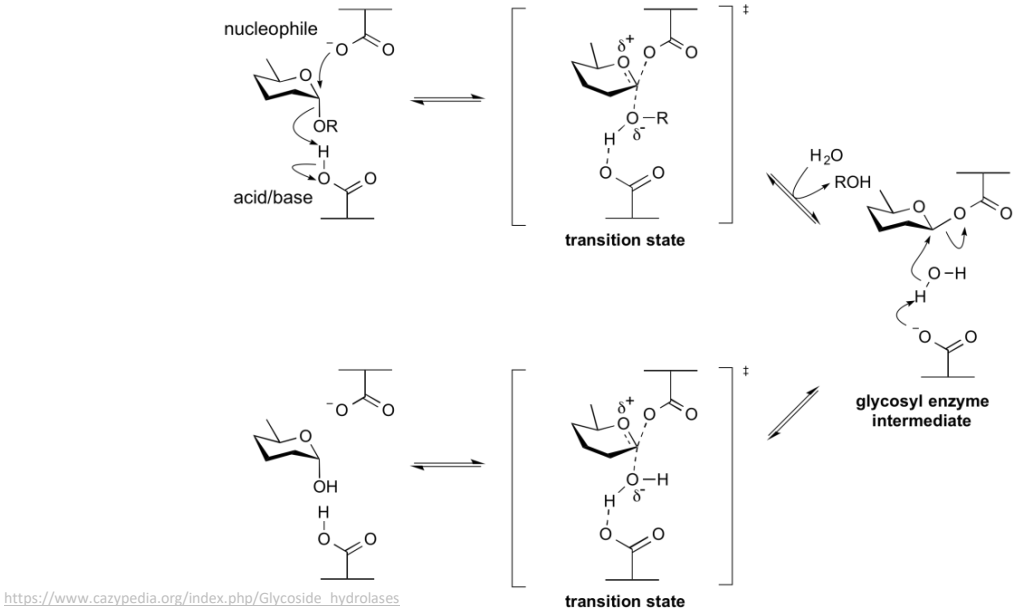


Figure 2. Example of the typical mechanism utilized by retaining glycoside hydrolases. Substrate shown in illustration is of "α" anomeric carbon conformation. The nucleophilic residue directly attacks the anomeric carbon of the substrate, while the acid/base residue provides a proton allowing for bond cleavage which is then followed by activation of a water molecule that allows for cleavage of the bond between the substrate and nucleophilic residue.

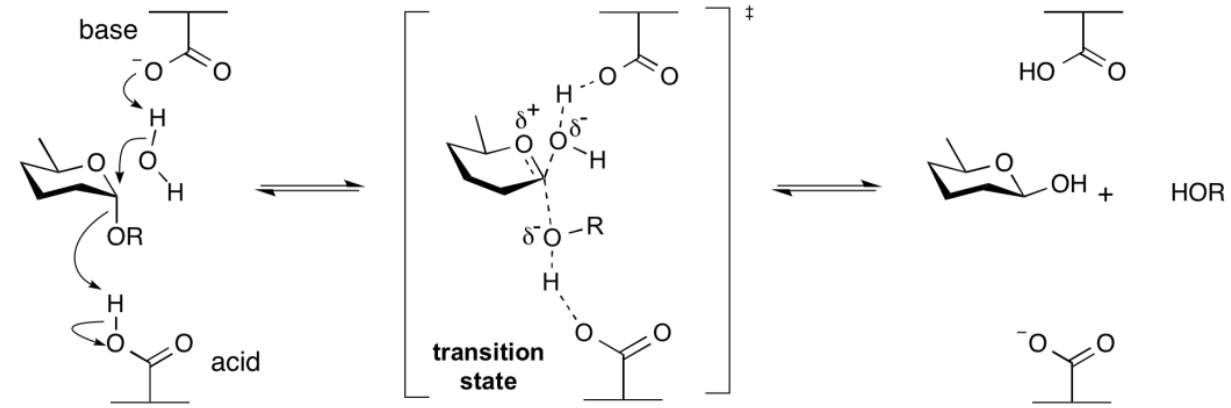


Figure 3. Example of the typical mechanism utilized by inverting glycoside hydrolases. Substrate shown in illustration is of "α" anomeric carbon conformation. The Lewis base residue directly activates a water molecule allowing for hydrolysis, while the leaving group is protonated by a Brønsted-Lowry acid residue.

In glycoside hydrolase assays utilizing nitrophenol-based colourimetry, a saccharide (ie. galactose, xylose, maltose, etc.) covalently bound via an ether linkage to a nitrophenyl moiety. As with natural carbohydrates, the nitrophenyl linkage may assume either "α" or "β" conformation, allowing for the selective testing of either α- and β-glycosidase activity. The most common type of nitrophenol moiety, and the one utilized throughout this study, is para-nitrophenol (pNP), though ortho- and meta- forms of the analog may also be used. These substrate-mimicking compounds are initially colourless in solution, but upon cleavage of the nitrophenyl group from the sugar substituent the solution will turn a yellow colour due to the resonance structure of the nitrophenolate ion at basic pH values, which can be easily measured at a wavelength of 405 nm. Illustration of pNP-based substrate analogs and the pNP liberated from the analogs via enzymatic cleavage can be seen in Figure 4.

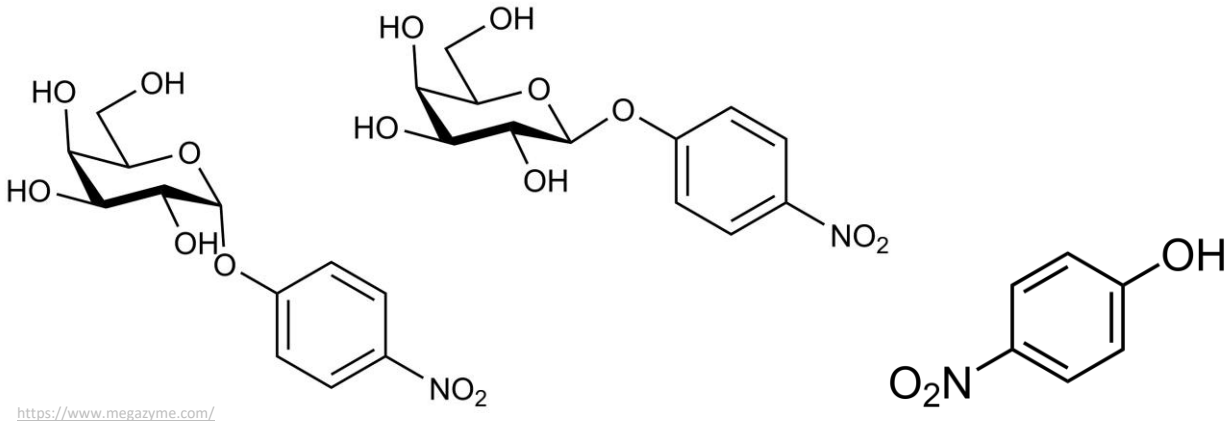


Figure 4. para-nitrophenyl-based substrate analogs used in the kinetic analysis of putative glycoside hydrolase enzymes. Left to right: 4-nitrophenyl-α-D-galactopyranoside, 4-nitrophenyl-β-D-galactopyranoside, para-nitrophenol catalysis product. The para-nitrophenol product is generated by enzyme-mediated hydrolysis of the ether bond in substrate analog and allows for colourimetric determination of enzyme activity.

The structure of pNP analogs allows for the compounds to be dissolved in water for pH optima determination, or buffer, for usage in Michaelis-Menten kinetic assay. Due to the pH of the given system affecting the amount of free nitrophenol that is present in the ionic form, a range of standards with a known concentration of pNP and solution pathlength will be needed in order to determine the extinction coefficient for the range of pHs utilized in a pH optima assay. Extinction coefficients can be calculated by rearranging Beer's Law to $\varepsilon = (A / [c \cdot l])$. This standard assay, along with pH optima and Michaelis-Menten assays, can be conducted in a microplate to easily allow for multiple readings with a variety of solution conditions. Absorbance readings from pH/Michaelis-Menten assays can be used to calculate the amount of pNP present in the solution, and from this calculate the rate of degradation of the substrate analog, which can be used to determine the kinetic parameters for each enzyme.

2.7 Site-Directed Mutagenesis of Potential Active-Site Residues

Site-directed mutagenesis can be used to identify potential catalytically-active residues in the enzymes of interest through the substitution of these residues to ones that contain inert side-chains. The method utilized in this study to produce such mutants was the Single-Oligonucleotide

Mutagenesis and Cloning Approach (SOMA). This process takes advantage of the ability of high-processivity DNA polymerases to replicate entire plasmids and certain stains of *E. coli* to re-circularize and complement single-stranded plasmid DNA. To conduct mutagenesis using SOMA, a single oligonucleotide (referred to subsequently as the "primer") is synthesized to complement an 18 - 25 nucleotide section of the plasmid that contains the codon for the amino acid designated to be mutated, with a codon on the primer containing the minimal number of "mismatched" nucleotides required to result in the section coding for an amino acid with an inert side-chain. In order to generate a mutant plasmid, a high-fidelity high-processivity polymerase, such as the *Pyrococcus furiosus*-derived *Pfu*, is utilized. A standard PCR protocol is followed, with the time in the elongation step suitably extended to allow for the polymerase to copy the full length of the plasmid, only terminating once it has reached the opposite end of the primer. Following generation of these single-stranded "plasmids", the restriction enzyme *Dpn*I is introduced into the reaction mixture. *Dpn*I recognizes the nucleotide sequence GATC/CTAG, where the adenine on both strands is methylated, and cuts DNA at these points, between the adenine and thymine nucleotides. Bacteria extensively methylate plasmid DNA to protect it from degradation, while DNA polymerase is capable of generating only unmethylated DNA. As such, *Dpn*I will breakdown the parent strands of plasmid DNA while leaving the single-stranded mutant DNA untouched. Reactions can then be run on an agarose gel to assess if mutant plasmid has been generated. Mixtures that contain plasmid may be purified and transformed into competent DH5α *E. coli* cells, where the linear strands will be recircularized and complementary strands formed with innate cellular machinery. Cells can be plated onto LB-agar media supplemented with a selective antibiotic and colonies selected for, which will be grown prior to plasmid DNA extraction. Extracted DNA can then be validated for the presence of the desired mutation.

2.8 Screening of Potential Carbohydrate Substrates through Thin-Layer Chromatography

Targets that showed activity against pNP analogs can be tested against potential carbohydrate substrates using thin-layer chromatography. In this assay, di-/oligo-saccharides containing (a) monosaccharide component(s) that are the same as or structurally similar to that of the carbohydrate substituent in the pNP analog from the kinetic assays would be tested for degradation by the target enzymes. Carbohydrates were first dissolved in the buffer conditions that were found to be suitable for protein stability. Buffer/carbohydrate mix were then divided into equal volumes into two tubes for incubation, with one being left without protein and the other having a set amount of protein introduced into it. The carbohydrate mixtures may then be incubated under agitation for an extended period of time in order to allow for the potential degradation to proceed. Following this incubation 95% ethanol was added to all tubes and incubated for one hour in order to arrest the reaction and denature the protein present in solution. Tubes can then be spun at high rpm to pellet denatured protein, leaving a mixture of buffer, ethanol, and carbohydrate. The remaining solution can be transferred to a new centrifuge tube and concentrated via rotary evaporation. Following the concentration of the mixtures to an appropriate volume, the solutions were then spotted onto a silica TLC plate and run using a polar solvent. After drying, carbohydrates were visualized by staining through the application to and heating of p-anisidine on the plate, which under heat will condense with any carbonyl/amine groups present on the saccharides, making it an excellent general-purpose stain(113).

2.9 Product Characterization Using Nuclear-Magnetic Resonance and Mass Spectrometry

Due to the high variability of carbohydrate substrates/products and the diversity of ways in which CAZymes may degrade these carbohydrates, multiple methods of analysis such as mass spectrometry (MS) and nuclear-magnetic resonance (NMR) spectroscopy is often required to determine the structure of even simple mono/di-saccharides. Digestion of carbohydrates by proteins can be conducted in the same or similar manner to that of those conducted for TLC. Purification of carbohydrates from protein mixture may require extra care as analyte conditions are more stringent for these methods of analysis, requiring not only the removal of protein, but often also all salts, buffering agents, and certain solvents. Reaction products may also require separation prior to analysis.

Nuclear-magnetic resonance spectroscopy is a useful analytical technique for the elucidation of carbohydrate structure, providing information on the number, identity, modifying groups, and anomeric configuration of monosaccharide residues, as well as linkages sequences(114). The process takes advantage the intrinsic nuclear-magnetic moment of atoms containing an odd number of protons/neutrons, most commonly single-proton hydrogen (^{1}H) and the six-proton/seven-neutron isotope carbon-13 (^{13}C). A strong magnetic field is applied to the samples to align the nuclear-magnetic spin of the atoms, after which a weak oscillating magnetic field is applied in order to disturb the magnetic fields in the aligned atoms. As the magnetic spin of the atoms relaxes back to equilibrium, a voltage is produced within a detection coil. The voltages generated are then used to generate a spectrum based on the frequency and intensity of these voltages as compared to standard.

Mass spectrometry is a technique that is commonly combined with NMR to determine the structure of carbohydrates. MS can provide information on the molecular mass, monosaccharide substituents and sequence, linkage types, modifying groups, and stereochemistry of carbohydrates(115). In this technique a purified sample is ionized, a process which results in the molecules being transferred into the gas phase, and the ions are transported to a mass analyzer under the power of a magnetic or electric field. A mass analyzer separates these ionized sample molecules based on their respective mass-to-charge ratios (*m/z*), which can measured in several different ways including: sector instruments that alter the linear trajectory of the sample using a static magnetic/electric field; time-of-flight analyzers which accelerate particles based on their potential and measures the time it takes for the particles to reach the detector; quadrupole analyzers that destabilize particles outside of a particular *m/z* range the device has been tuned to; and ion traps that capture and sequentially eject particles as the electric field within the trap is altered.

2.10 Crystal Screening and Diffraction Optimization

X-ray crystallography has been considered the gold-standard for protein structure determination for over 50 years(116), and thus will be critical to the determination of the structure of the target proteins. The first step in the determination of an X-ray crystal structure is the preliminary crystallization of the target proteins. To achieve preliminary crystallization, proteins that have been purified to near homogeneity and had all inclusions/aggregates removed from concentrated samples must be tested for stability, nucleation, and crystal growth across a wide variety of buffer, pH, salt, precipitant, and additive concentrations. For the purposes of preliminary crystallization commercially-available pre-made sparse-matrix screening solutions containing these variety of components are utilized. Protein samples were mixed with screening solution at certain ratios and situated above a reservoir of the same screening solution (known as the sitting-

drop vapour-diffusion method) and each well containing protein/precipitant droplets and the corresponding solution reservoir was individually sealed. For optimization of conditions found to produce crystals in sitting-drop conditions, the hanging-drop method is frequently utilized, where the protein/precipitant mixtures are suspended above solution reservoir on an inverted coverslip, sealed with vacuum grease, which isolates the droplet/reservoir system. In this method buffer, precipitating agent, salt, additive, reservoir volume, and protein concentration can be finely tailored in an attempt to increase crystal size and quality. In both methods vapour diffusion is the driving force behind crystallization. As the suspended protein droplet contains a higher percentage of water compared to the reservoir solution, over time this solution component will diffuse from the droplet into the reservoir solution, shrinking the droplet and gradually increasing the concentration of all other components in the system, including protein. As the protein and precipitating compounds are gradually concentrated in this dynamic liquid environment the chances of nucleation increase as eventually crystallization of protein and precipitant becomes the most entropically favourable state for the components(117). Addition of substrates or substrate-mimicking compounds may also be added to protein/precipitant mixtures before or after crystallization in an attempt to populate active sites, which can may help in the determination of catalytic function/mechanism.

CHAPTER 3 - MATERIALS AND METHODOLOGY

3.1 Materials

The cloned gene targets, BT2109, BT2918, BT3158 in pET15b vectors; BT2857 in the pET15TV-LIC vector, and BT2966 in pET28a were sent to us in collaboration with Dr. Alicia Lammerts Van Bueren (Faculty of Mathematics and Natural Sciences, Microbial Physiology) from the Gron Institute of Biomolecular Sciences & Biotechnology (Groningen, Netherlands). *fd* codon optimized synthetic genes of BT2857C, BT2857N, PGN1176, PGN1461, and PGN1459 in pET21b(+) plasmids were ordered from BioBasic (Markham, ON, Canada). Calcium-competent NEB 5-α (cat# C2987I) and BL21 (DE3) (cat# C2527I) *E. coli* cells were purchased from New England Biolabs Inc. (Ipswich, MA, USA). Competent BL21-CodonPlus (DE3)-RIPL *E. coli* cells (cat# 230280) from Agilent (Santa Clara, CA, USA) were borrowed from Dr. Weadge's lab. Competent Novagen Tuner™ (DE3)pLysS *E. coli* cells (cat# 70624-3) were purchased from MilliporeSigma (Burlington, MA, USA). Plasmid miniprep kits GeneJET and QIAprep® were purchased from Thermo Fisher Scientific (Waltham, MA, USA) and QIAGEN (Hilden, Germany), respectively. Restriction enzymes XhoI, NdeI, BamHI, DpnI; Cutsmart buffer, Phusion® High-Fidelity PCR Kit, Calf-Intestine Alkaline Phosphatase, T4 DNA ligase, and T4 DNA ligase buffer were also purchased from New England Biolabs Inc. The premixed DNA polymerase solution "2X PCR Bestaq™ MasterMix" was ordered through Applied Biological Materials (Richmond, BC, Canada). JumpStart™ Taq ReadyMix™ was ordered through Sigma-Aldrich (St. Louis, MO, USA). QIAquick PCR purification and QIAquick Gel Extraction kits were purchased through QIAGEN. Primers were ordered through Integrated DNA Technologies (Coralville, IA, USA).

Tryptone, yeast extract, agar A, monobasic sodium phosphate dihydrate ($NaH_2PO_4 \cdot 2H_2O$), sodium acetate trihydrate ($CH_3COO \cdot 3H_2O$), PIPES (free acid form), hydrochloric acid

(HCl), bicine, calcium chloride dihydrate ($CaCl_2 \cdot 2H_2O$), ammonium sulfate ($(NH_4)_2SO_4$), nickel(II) chloride hexahydrate ($NiCl_2 \cdot 6H_2O$), potassium bromide (KBr), iron(III) chloride hexahydrate ($FeCl_3 \cdot 6H_2O$), zinc(II) sulfate heptahydrate ($ZnSO_4 \cdot 7H_2O$), copper(II) chloride monohydrate ($CuCl_2 \cdot H_2O$), sodium sulfate anhydrous (Na_2SO_4), deoxycholic acid sodium salt, ethidium bromide, D-lactose monohydrate, isopropanol, were purchased through BioBasic Inc. (Markham, ON, Canada).

Agarose, tris (base), bis-tris, bis-tris propane, CAPS, CHES, HEPES, anhydrous dibasic sodium phosphate (Na_2HPO_4), glacial acetic acid, succinic acid, glycine, trisodium citrate dihydrate ($Na_3C_6H_5O_7 \cdot 2H_2O$), magnesium chloride hexahydrate ($MgCl_2 \cdot 6H_2O$), sodium chloride (NaCl), magnesium sulfate heptahydrate ($MgSO_4 \cdot 7H_2O$), imidazole, anhydrous monobasic potassium phosphate (KH_2PO_4), manganese chloride tetrahydrate ($MnCl_2 \cdot 4H_2O$), Triton X-100, sucrose, isopropyl β-D-1-thiogalactopyranoside, ampicillin, kanamycin, chloramphenicol, guanidine hydrochloride, 1-butanol, EDTA disodium dihydrate, 40% 19:1 acrylamide:bis-acrylamide solution, TEMED, glycerol, and ethylene glycol were purchased from Bioshop Canada Inc. (Burlington, ON, Canada).

Ammonium acetate, potassium tartrate, sodium potassium tartrate tetrahydrate, cobalt(II) chloride anhydrous ($CoCl_2$), glucose, TLC-tested carbohydrates, pNP-carbohydrate analogs, p-anisidine, D_2O, polyethylene glycol 1500, 3000, 4000, 6000, and polyethylene glycol monomethyl ether 2000 were purchased from Sigma-Aldrich Inc. (St. Louis, MO, USA). Sodium hydroxide and anhydrous citric acid were purchased through Anachemia (Mississauga, ON, Canada). Ammonium persulfate ($(NH_4)_2S_2O_8$) and Coomassie Brilliant Blue R-250 was purchased from Fisher Scientific Inc. (Hampton, NH, USA). 100% methanol and 95% ethanol were obtained from CFS Chemicals (Montréal, QC, Canada). Potassium chloride (KCl) and lithium sulfate

monohydrate were purchased from Alfa Aesar (Ward Hill, MA, USA). Lithium nitrate and aluminum-backed F60 silica TLC plates were purchased from EMD Millipore Inc. (Burlington, MA, USA). 96-well flat-bottom Nunc™ MicroWell microplates used in functional assays were purchased through ThermoFisher Scientic (Waltham, MA, USA). 96-well sitting-drop crystal trays were purchased from Art Robbins Instruments LLC (Sunnyvale, CA, USA). 24-well cell culture plates used in crystallization trials, 1.5mL and 0.6mL microcentrifuge tubes, and pipette tips were purchased from Sarstedt Ag & Co. (Nümbrecht, Germany). High-vacuum grease was purchased from Dow-Corning Inc. (Midland, MI, USA). Silica coverslips, Crystal Screen, and Crystal Screen 2 were purchased from Hampton Research Corp. (Aliso Viejo, CA, USA). MCSG crystal screens were purchased from Anatrace Inc. (Maumee, OH, Canada).

3.2 Methodology and Experimental Design

3.2.1 Introduction into and Extraction of Recombinant Plasmids in *E. coli* Cells

BT2109, BT2857, BT2918, BT2966, BT3158 plasmids and BT2857N, BT2857C, PGN1176, PGN1459, PGN1461 custom genes were transformed into calcium-competent NEB 5-α, BL21, BL21-CodonPlus, and Tuner *E. coli* cells via heat shock methodology. 50µL aliquots of competent cells in 1.5mL microcentrifuge tubes were thawed on ice for 5 minutes, after which 1 - 5µL of vector DNA solution was added to cells and incubated on ice for 30 minutes. Following incubation cells were subjected to heat shock at 42°C for 60 seconds and placed back on ice for 5 minutes. 200µL of high-NaCl LB broth medium was then added to tubes and transformations were incubated for one hour at 37°C and 240rpm before being spread on ampicillin-, kanamycin-, or ampicillin/kanamycin+chloramphenicol-containing LB-agar plates as appropriate. Plates were incubated overnight at 37°C to allow for growth, and single colonies were picked from plates and grown overnight in either 5mL or 10mL cultures of LB broth containing antibiotics at 37°C and

240rpm. Standard antibiotic working concentrations used for LB broth cultures was 100µg/mL for ampicillin, 50µg/mL for kanamycin, and 25µg/mL for chloramphenicol.

Glycerol stocks of transformed cells were generated by pipetting 650µL of overnight cultures into 1.5mL microcentrifuge tubes after which 650µL of sterile 50% glycerol was then added and mixed via pipetting. Tubes were then submerged upright in liquid nitrogen until completely frozen, then stored in -80°C ultra-low freezer (Thermo Scientific™ Forma 88000 Series).

Plasmid vectors were harvested from NEB 5-α overnight cultures using either GeneJet or QIAprep plasmid miniprep kits. 1mL sample of overnight culture was pelleted down at 17,900xg for 5 minutes in tabletop centrifuge, supernatant was decanted, and pellet was then resuspended in 250µL of a resuspension solution. 250µL of a lysis solution containing RNase was then added to the resuspension and the reaction was allowed to proceed for up to 5 minutes. 350µL of neutralization solution was then added to the solution and tubes were mixed by inverting, before being centrifuged for 10 minutes at 17,900xg to remove precipitated cellular components. 800µL of supernatant was then pipetted into a tube containing a silica column, which was then spun for one minute to absorb DNA into the silica. Flow through was decanted from the lower collection tube and 750µL of an ethanol-containing wash buffer was pipetted onto the silica column before being spun for one minute, decanted, and spun again for one minute. 50µL of an elution buffer of 10mM Tris:HCl at pH8.5 was then pipetted directly onto the silica and incubated for one minute after which the column was placed in a sterilized 1.5mL microcentrifuge tube and centrifuged for one minute to elute DNA. The concentration of the isolated DNA was determined via nanodrop analysis and tubes were then stored at -20°C for subsequent use.

3.2.2 Protein Expression and Initial Isolation

Overnight cultures of BL21 (DE3), BL21-CodonPlus (DE3)-RIPL, or Tuner™ (DE3)pLysS *E. coli* cells containing desired vector were prepared by introducing a small amount of ultra-frozen glycerol cell stock to 10mL of LB media containing antibiotics at standard working concentrations. Cultures were incubated overnight at 37°C and 240rpm before being inoculated into 1L cultures of high-NaCl LB broth containing appropriate antibiotic(s) at standard concentrations (100μg/mL ampicillin; 50μg/mL kanamycin; 25μg/mL chloramphenicol). 1L cultures were incubated under the same conditions as overnight cultures until an OD_{600} of 0.6 – 0.8 was reached, at which point isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added to the cultures to a final concentration of 1mM and incubation conditions were changed to 16°C and 160rpm to facilitate overexpression of the desired protein. Other expression media used included 2-YT and Terrific broth (both containing tryptone, yeast extract, and sodium chloride) at a volume of 500mL per flask, and autoinduction-LB at 1L/flask. For autoinduction-LB, standard LB was dissolved in 925mL of water and autoclaved, after which 50mL of 20x P salt solution, 20mL of 50x 5052 solution, 1mL of 1M $MgSO_4$, 1mL 1000x trace-metal mix, and 3.5mL 11M NaOH to achieve a media pH of ~7.5. Autoinduction cultures were left at 37°C, 240rpm, for 8 hours after inoculation before being moved to 16°C 160rpm conditions for overnight expression. After overnight expression, cells were harvested via centrifugation at 5000rpm for 15 minutes at 4°C. If not immediately lysed, cell pellets were stored at -20°C in 50mL falcon tubes.

Overnight cultures for cells expressed in SelenoMethionine minimal media were grown at a volume of 100mL per litre of culture. Cultures were pelleted by centrifugation at 5000rpm, 4°C, before being resuspended in a volume of M9 minimal media equal to that of the culture volume and pelleted again. Pellets were again resuspended in M9 minimal media, at a volume of

25mL/100mL of culture. 25mL of resuspended cells were added to each 1L culture contained in baffled 2L flasks, and cultures were grown and induced under the same conditions as described above for LB and 2YT cultures.

For chemical lysis cell pellets were initially dissolved in a beaker on ice with a volume of 730mM sucrose solution containing 300mM NaCl; buffered with Tris:HCl, HEPES:NaOH, or dibasic sodium phosphate (Na_2HPO_4), at pH7.0 - 8.0. Once the pellet had completely homogenized with sucrose solution, lysozyme was added to concentration of 1mg/mL and the solution was left under agitation for 10 minutes before the addition of 2 volumes of detergent solution (1%(w/v) deoxycholate, 2%(w/v) Triton X-100, 300mM NaCl, buffered identically to sucrose solution used). 1M $MgCl_2$ and 2mg/mL DNase I solutions were then added and the solution was left for up to 30 minutes, or until the lysis solution was no longer highly viscous. The soluble and insoluble fractions were separated by centrifugation at 25,000xg for 45 minutes at 4°C, supernatant was kept as crude extract for isolation of protein target.

Prior to immobilized metal affinity chromatography (IMAC), Ni-NTA resin was suspended in crude protein extract in 50mL falcon tubes and incubated under inversion for 30 - 90 minutes at 4°C in a Fisher Scientific 346 mixer. After incubation the resin/extract mixture was decanted into a Bio-Rad Econo-Column® and allowed to drain back into falcon tubes as isolate "flow-through". Wash and elution buffers used during IMAC contained the same buffering/salt conditions as lysis buffer. Wash buffer containing buffer, salt, and 5mM imidazole was poured over resin with care given to not disturb the settled resin. A flow adaptor, connected to a Bio-Rad Econo™ Gradient Pump, was placed in the column ~5mm above the resin bed and 50mL of wash buffer was run through the column. After washing, proteins were eluted off the resin through gradient application of elution buffer containing increasing concentrations of imidazole (25, 50,

75, 100, 125, 150, 175, 200, 225, 250mM) and collected in 10mL fractions. Column and resin was then cleaned by adding one column volume 500mM imidazole, one column volume water, one volume 0.5M NaOH, one volume water, one volume 10%(v/v) acetic acid, one volume water, before being stored in 20%(v/v) ethanol.

3.2.3 SDS-PAGE Analysis

Insoluble pellet, IMAC flow-through, column wash, and eluted fractions were kept for assessment of expression, solubility, and purity after initial isolation. 20µL of each sample was mixed with 5µL of 5x SDS-PAGE sample buffer in 0.6mL microcentrifuge tubes, and then heated at 85°C for 5 minutes. Samples were loaded into 12%(w/v) SDS-PAGE gel along with Bio-Rad Precision Plus Protein™ Dual Color Standards ladder. Gels were run at 180V for 50 minutes. Gel recipe can be seen in Table 1.

Table 1. Solution compositions and volumes used in the making of SDS-PAGE gels. Both recipes makes a total of 10mL of gel solution. These volumes are suitable for the casting of 2 separating gels and 10 separating gels.

Solution	Separating Gel (12%)	Stacking Gel (4%)
40% 19:1 acrylamide/bis-acrylamide	3.1mL	1.008mL
milliQ H_2O	4.235mL	6.332mL
1.5M Tris:HCl pH8.8	2.5mL	-
0.5M Tris:HCl pH6.8	-	2.5mL
10% w/v SDS	100µL	100µL
TEMED	15µL	10µL
10% w/v APS	50µL	50µL

After running, gels were placed in Tupperware containers with Coomassie stain and heated in a 700W microwave for one minute. Stain was emptied from container and gels were rinsed and submerged in clean water, then heated in microwave for 3 minutes to de-stain and visualize protein

bands. De-stained gels were imaged using Bio-Rad VersaDoc™ MP 4000 Molecular Imager and adjusted using Quantity One® 1-dimensional analysis software.

3.2.4 Dialysis of Protein

SDS-PAGE analysis of IMAC samples was used to determine which fractions contained an acceptable amount of contamination bands as compared to the amount of purified protein target. Fractions with a high degree of purity were dialyzed to decrease the concentrations imidazole and salt, and in some cases change buffering conditions, in preparation for further purification via fast protein liquid chromatography. Selected fractions were pooled in Fisherbrand™ 50mm Regenerated Cellulose dialysis tubing with a molecular weight cut-off (MWCO) of 6,000 - 8,000Da. Tubing was clipped at both ends with 50mm polyamide SnakeSkin™ or Sigma closures. Dialysis bags containing pooled fractions were then submerged in 2L of milliQ water agitated with a magnetic stir bar and left at 4°C for one hour. After one hour, dialysis bags were transferred to 2L beakers of desired buffer (ie. Tris:HCl, HEPES:NaOH, Na_2HPO_4) and left overnight at 4°C under stir bar agitation. If aggregates/inclusion bodies were seen in protein solution after overnight dialysis, the solution was poured into 50mL falcon tubes and spun at 4300xg for 20 minutes at 4°C in Beckman Coulter Allegra X-14R swinging-bucket centrifuge to pellet the insoluble protein. Soluble protein solution, free of aggregates, was stored at 4°C until further purification could be conducted.

3.2.5 Purification Through Anion and Cation Exchange Chromatography

Protein samples that had undergone dialysis after IMAC purification were run through anion- and or cation-exchange fast protein liquid chromatography (FPLC) to further purify the protein prior to crystallization and functional assay. Purification via these techniques was performed using the GE Healthcare ÄKTA pure protein purification system, controlled by

protocols in the UNICORN 6.4 Workstation software. Before running ion-exchange protocols, lines were purged with buffers being used in the protocol, with protein sample line purge bypassing the ion exchange column, 1M NaCl-containing buffer being run over the column until the UV signal at 280nm reached 0 milli-Absorbance Units (mAU) to clear any protein left in the column, and non-salt buffer being run over the column until conductivity flatlined to prime column for sample application. Flow rate for these steps was 5mL/min, and column pressure was ensured to be below 0.5 megapascals (MPa). For anion exchange with a GE 5mL HiTrap Q FF column the positively charged buffer Tris:HCl at pH8.0 was used, while negatively charged HEPES:NaOH at pH7.5 was used for cation exchange with a GE 5mL SP Sepharose FF column. In the standard protocols utilized for both anion and cation exchange all buffer and sample lines were primed before the application of sample. After priming, sample was run over the column twice to ensure maximum interaction between resin and protein. The column was then washed with 15mL of the respective wash buffer before elution. Elution buffer, comprised of 1M NaCl in the same buffer condition used in the wash, was then applied to the columns in gradient from 1 - 100%, causing the gradual release of proteins from the column as chloride ions (in anion exchange) or sodium ions (in cation exchange) outcompete the proteins for resin interaction. Elution was collected in 3mL fractions, and fractions that corresponded to the greatest peak in the UV_{280} signal were run on SDS-PAGE gels to verify the presence and purity of the target protein.

3.2.6 Ultrafiltration and Determination of Protein Concentration

FPLC-purified protein was concentrated using Amicron® Ultracel-15 regenerated cellulose centrifugal filters purchased through Millipore. Centrifugal filters with MWCO of 10 or 30kDa were used depending on the size of the protein being concentrated, with 30kDa being used for all full-length BT proteins and PGN1461, 10kDa being used for BT2857C, BT2857N, PGN1176 and

PGN1459. Before concentration filters were filled with milliQ water, mixed briefly, and left to soak for 10 minutes in order to hydrate the cellulose membrane. Water was emptied from filter unit, protein sample was placed in the upper compartment of the filter, and filters were spun in Allegra X-14R swinging bucket centrifuge at 4000xg, 4°C, for 15 minutes. Flow-through was emptied from the bottom compartment of the filter and the upper compartment was refilled with unconcentrated protein sample, and the centrifugation process was repeated until protein had reached an acceptable level of concentration, typically between 10mg/mL and 40mg/mL. If the protein sample at the bottom of the upper filter compartment began to appear yellow or brown, a P1000 pipette would be used to gently agitate and homogenize the sample. After ultrafiltration, samples were pipetted into 1.5mL microcentrifuge tubes and spun for 30 minutes in Beckman Coulter Microfuge® 20R at 20000xg, 4°C, to pellet any inclusion bodies that may have formed during concentration. Concentrated protein sample was moved into fresh 1.5mL microcentrifuge tubes after spinning.

Concentration of protein samples was determined by measuring the absorbance of tryptophan residues at UV_{280} using Thermo Scientific™ Genesys 10S UV-Vis Spectrophotometer. Spectrophotometer was blanked with 1000µL of milliQ water in Sarstedt UV transparent LCG 8.5mm cuvettes, after which 5µL of protein sample was added to the water and homogenized in solution by pipetting up and down with a P1000 pipette. Absorbance spectra from 260 - 300nm was collected and the reading at 280nm was taken to calculate the protein concentration in the cuvette using Beer's Law. The equation is rearranged from "$A = \varepsilon \cdot c \cdot l$" to "$c = A / (\varepsilon \cdot l)$", where $A = UV_{280}$ value, ε = proteins extinction coefficient as determined by EXPasy Protparam, and l = 1cm. Extinction coefficients for each protein can be found in Appendix 3. The dilution factor is then calculated and applied to the sample, where "sample concentration = cuvette concentration ·

(1005µL / 5µL)". Protein concentration in units of mg/mL can also be calculated using the equation "concentration(mg/mL) = concentration(mol/L) · MW(g/mol) · (1L/1000mL) · (1000mg/1g)". These calculations were done automatically using Microsoft Excel software.

3.2.7 Crystal Screening and Expansion Plating

Initial screening for stable, meta-stable, and crystal-forming conditions was conducted in Art Robbins Instruments (ARI) 96-well sitting-drop crystal trays (102-0001-03 type) using the ARI Crystal Gryphon automated plating system and pre-made first-pass crystallization condition screens from Anatrace Inc., Hampton Research Corp., and Molecular Dimensions Ltd. Protein sample to reservoir solution ratios used in sitting-drop screening were 1:1, 2:3, or 2:1. After setup, sitting-drop trays were labelled and sealed with Duck® HD Clear tape before being stored at 18°C in Molecular Dimensions 100L Benchtop Incubator. Trays were observed under a Zeiss SteREO Discovery.V8 microscope after 3 days, then at the end of every multiple of 7 days after incubation. Successful hits for crystallization conditions were recorded on scoring sheets for crystal growth optimization.

The primary technique used for optimization was hanging-drop vapour-diffusion in 24-well Sarstedt TC cell culture plates. Each expansion plate focused on one or more conditions that yielded successful crystallization in 96-well plates, with aspects such buffer, pH, stabilizing additive, salt/protein concentration, and amount of precipitant being altered slightly to test for ideal crystal growth. Reservoir solutions were prepared in each well and vacuum grease was applied around the rim of each well. Six siliconized glass coverslips were laid on the benchtop and droplets of reservoir solution from wells A1 - A6 were applied to/around the centre of the slip using a P10 pipette. Protein was then added to A1 coverslip droplet(s) and the coverslip was inverted and placed overtop of well A1, with light pressure applied to seal the well and prevent the ingress of

air. Process was repeated with wells A2 - A6. The procedure outlined for row A was then repeated for rows B, C, and D, individually. Plates were labelled, dated and placed in 18°C incubator or 4°C cold room. Diagrams describing conditions for coverslips with multiple droplets were drawn on expansion plate formulation sheet. Volumes of both reservoir solution and protein sample were increased for hanging-drop expansion plate droplets, typically by a factor of 10 (ie. 5/6µL hanging droplet vs. 0.5/0.6µL sitting-drop).

Another technique utilized for optimization was formulation of custom 96-well sitting-drop plates. Reservoir solutions were prepared in empty sterile 96-well deep well blocks from BioBasic (BR581-96N). Plating was again performed using the ARI Gryphon automated system with 102-0001-03 sitting-drop trays, with final drop volumes being doubled to 1-1.2µL. Trays were sealed, stored, and observed in same manner as initial-screen plates.

Crystals generated by both hanging-drop and sitting-drop vapour diffusion were cryoprotected in a variety of solutions (usually liquid reservoir conditions supplemented with 30% glycerol or ethylene glycol) and frozen in liquid nitrogen prior to shipping. Crystals were analyzed at the Canadian Light Source Synchrotron (CLS/CCRS) in Saskatoon, SK, Canada. Datasets for native crystals were collected on both CMCF-ID and CMCF-BM beamlines, while anomalous diffraction data were collected on the CMCF-BM beamline. Examples of preliminary statistics for the diffraction of targets can be seen in Appendix 7.

3.2.8 Enzyme Activity Assays

Kinetic assaying of the *B. thetaiotaomicron* enzymes was performed using para-nitrophenyl substrate analogs for simple colourimetric detection and quantification of reaction rates. All para-nitrophenyl assays were conducted in Nunc™ Microwell 96-well flat-bottom microplates, measured using a BioTek® (Winooski, VT, USA) Cytation5 cell imaging multi-mode

reader, both graciously provided by Dr. Weadge's lab. pH profiles for each enzyme were determined by incubating 2µL of 625µM enzyme with 1mM para-nitrophenyl-β or α-galactopyranoside (pNP-Gal) in a pH range of 100mM McIlvaine buffers at a final volume of 250µL. Reactions were conducted in triplicate at room temperature (~25°C), for 60 minutes with reads taken at 52 second intervals. To produce Michaelis-Menten curves, 2µL of 625µM enzyme was incubated with pNP-Gal substrate analog at increasing concentrations. Reactions were again run in triplicate, 25°C, at 250µL, for 60 minutes with 52 second read intervals. Typical plate setup for Michaelis-Menten kinetic assays can be seen in Table 2.

Table 2. Typical Microplate Setup for pNP Michaelis-Menten Kinetic Assays. Volume of 10mM pNP-carbohydrate solution in 100mM buffer recorded in top third of each cell; volume of 100mM buffer recorded in middle of each cell; volume of protein solution listed in bottom third of each cell. Row A served as control for auto-hydrolysis of pNP substrates.

	1	2	3	4	5	6	7	8	9	10	11	12
A	1.25µL 248.75µL	2.5µL 247.5µL	3.75µL 246.25µL	5µL 245µL	6.25µL 243.75µL	12.5µL 237.5µL	25µL 225µL	50µL 200µL	75µL 175µL	100µL 150µL	125µL 125µL	150µL 100µL
B	1.25µL 246.75µL 2µL	2.5µL 245.5µL 2µL	3.75µL 244.25µL 2µL	5µL 243µL 2µL	6.25µL 241.75µL 2µL	12.5µL 235.5µL 2µL	25µL 223µL 2µL	50µL 198µL 2µL	75µL 173µL 2µL	100µL 148µL 2µL	125µL 123µL 2µL	150µL 98µL 2µL
C	1.25µL 246.75µL 2µL	2.5µL 245.5µL 2µL	3.75µL 244.25µL 2µL	5µL 243µL 2µL	6.25µL 241.75µL 2µL	12.5µL 235.5µL 2µL	25µL 223µL 2µL	50µL 198µL 2µL	75µL 173µL 2µL	100µL 148µL 2µL	125µL 123µL 2µL	150µL 98µL 2µL
D	1.25µL 246.75µL 2µL	2.5µL 245.5µL 2µL	3.75µL 244.25µL 2µL	5µL 243µL 2µL	6.25µL 241.75µL 2µL	12.5µL 235.5µL 2µL	25µL 223µL 2µL	50µL 198µL 2µL	75µL 173µL 2µL	100µL 148µL 2µL	125µL 123µL 2µL	150µL 98µL 2µL

The rate for each reaction (in Abs/min) was calculated for each pH/substrate concentration, and converted to M/min using Beer's Law. Michaelis-Menten values of V_{max} and K_M were determined by the equation of the curve for each construct. The k_{cat} value was calculated using the formula $k_{cat} = V_{max}/E_t$, where "E_t" is the defined the total enzyme concentration. Enzymes that showed activity against pNP-β-Gal were expressed in *E. coli* Tuner (DE3)pLyss cells to confirm galactosidase activity, a cell line that does not contain the lacZ gene, which codes for an

endogenous β-galactosidase that may cause above-background hydrolysis of the pNP-β-Gal substrate.

Enzymes determined to have potential activity against carbohydrates were screened against variety of available (poly)saccharides. 5mg of each polysaccharide was dissolved in 495μL of phosphate buffer at ideal pH as determined by pH optima assay, in 1.5mL microcentrifuge tubes. 5μL of enzyme at 25mg/mL was added to the solution and left overnight at 240rpm, 37°C. Reactions were quenched with 500μL of 95% reagent ethanol and left for 60 minutes at room temperature. Microcentrifuge tubes were spun at 14,800rpm in tabletop centrifuge for 30 minutes to pellet protein in reaction mixture, and supernatant was transferred into sterile 1.5mL microcentrifuge tubes via pipetting. Supernatants were vacuum concentrated at 40°C to ~50-100μL using the tabletop CentriVap concentrator from the Labconco Corporation (Kansas City, MO, USA). Samples were run on Silica gel 60 aluminum TLC plates (MilliporeSigma) using 1-butanol:acetic acid:water solvent at a 2:1:1 volume ratio until the solvent front was 1cm from the top of the plate. Plates were left to dry completely before 1% (w/v) p-anisidine stain (dissolved in 10% (v/v) methanol, 90% (v/v) 1-butanol) was applied, plates were then baked at 130°C for 10 minutes to visualize carbohydrates. TLC separation was also utilized to separate products for NMR, with a larger amount of product applied to plates (5 - 20mg). 1cm sections of plate were then removed from the edges and stained to check if angular migration on the plate had occurred. After verification of migration uniformity, a line was drawn across the remaining unstained plate separating the areas where digested products were assumed to be. Silica gel was then scraped from each plate section and placed in a beaker. Each sectional scrape was dissolved in 35mL of methanol and agitated with a magnetic stir bar at 200rpm, room temperature, for 30 minutes. Samples were vacuum concentrated until dry and redissolved in D_2O in preparation for NMR analysis.

3.2.9 Site-Directed Mutagenesis and Subcloning

Site-directed mutagenesis (SDM) of BT2857C and BT2857N was undertaken in an attempt to identify active-site residues through subsequent pNP kinetic analyses. Primers were created based on guidelines in the Agilent QuickChange Site-Directed Mutagenesis Kit. Primer pellets were resuspended in QIAGEN "EB buffer" (10mM Tris:HCl pH8.0) to a final concentration of 1mM. Working primer stocks at 100µM were prepared via dilution with EB buffer. Mutants were generated using the single-oligonucleotide mutagenesis and cloning approach (SOMA) with Phusion® High-Fidelity DNA polymerase. Conditions such as DMSO/$MgCl_2$ concentration, buffer composition, and annealing temperatures were altered if standard procedure did not achieve mutagenesis. Primer sequences can be found in Appendix 4. Protocols can be seen in Table 3 and Table 4.

Table 3. Reagent volumes typically used in Phusion DNA Polymerase SOMA protocol for SDM of BT2857 constructs. Reactions were conducted in Bio Basic flat-cap 0.2mL thin-wall PCR tubes.

Reagent	Volume
5X Phusion buffer (HF or GC)	10µL
10mM dNTP mix	1 - 3µL
100ng/µL Primer	5µL
Template DNA	µL for 100 - 250ng
100% DMSO (optional)	0.5 - 4.5µL
50mM $MgCl_2$ (optional)	0.5 - 4.5µL
2,000U/µL Phusion polymerase	0.5µL
Nuclease-free H_2O	to 50µL total volume

Table 4. Typical thermal cycler temperature profile for SOMA mutagenesis of BT2857 constructs. Reactions were conducted in a Bio-Rad Laboratories Inc. T100™ Thermal Cycler utilizing a heated lid to prevent evaporation of reagents during procedure.

Step	Temperature (°C)	Time (s)	# of Cycles
Initial Denaturation	98	60	1
Denaturation Annealing Elongation	98 55 - 61 72	30 30 270	35
Final Elongation	72	360	1
Hold	4	∞	1

Following PCR, 2µL of 20000U/mL DpnI restriction enzyme was added directly to the reaction mixture, and tubes were incubated at 37°C, 240rpm, for one hour in order to degrade any methylated parental DNA, and then inactivated at 80°C for 20 minutes. Following DpnI digestion, PCR reactions were run on a 0.7% (w/v) agarose gel alongside an NEB Quick-Load® 1kb DNA Ladder in order to visualize banding and determine if the generation of mutagenic plasmids was successful. Digested reactions that showed a single band at the expected size were cleaned using the QIAquick PCR purification kit and transformed into competent NEB 5-α *E. coli.* cells. Colonies were picked and grown in 5mL overnight for plasmid isolation. Concentration of isolated plasmids was analyzed, and positive sequencing confirmed at the Robarts Research Institute (University of Western Ontario, London, ON, Canada).

The subcloning of the DUF4959 domains from BT2857N (pET21b(+)) and BT3158 (pET15b) into the pET21b plasmid first involved PCR amplification of the genes. PCR was conducted using both JumpStart™ Taq ReadyMix™ and 2X PCR Bestaq™ MasterMix, and the standard protocol can be seen in Table 5 and Table 6.

Table 5. Reagent volumes used in PCR amplifications of DUF4959 in BT2857/3158.
Reactions were conducted in Bio Basic flat-cap 0.2mL thin-wall PCR tubes.

Reagent	Volume
100ng/µL Forward Primer	5µL
100ng/µL Reverse Primer	5µL
Template DNA	µL for 100 - 250ng
Jumpstart ReadyMix Taq	25µL
Nuclease-free H_2O	to 50µL

Table 6. Typical thermal cycler temperature profile for DUF4959 PCR amplifications. Reactions were conducted in a Bio-Rad Laboratories Inc. T100™ Thermal Cycler utilizing a heated lid to prevent evaporation of reagents during procedure.

Step	Temperature (°C)	Time (s)	# of Cycles
Initial Denaturation	94	60	1
Denaturation Annealing Elongation	94 61 72	30 45 90	35
Final Elongation	72	300	1
Hold	4	∞	1

Successful amplification was identified by agarose gel electrophoresis, where reaction mixtures were run on 0.7% (w/v) gels along with NEB Quick-Load® 100bp DNA Ladder. Reactions that showed singular banding at the expected length were purified and analyzed for concentration. Purified mixtures were then digested using the restriction enzymes NdeI and XhoI. Standard protocol can be seen in Table 7. Plasmid vectors were also digested using this procedure, with alkaline phosphatase buffer and alkaline phosphatase enzyme added to prevent re-ligation.

Table 7. Reagent volumes for DUF4959/pET21b(+) digestion reactions. Reactions were conducted in Bio Basic flat-cap 0.2mL thin-wall PCR tubes heated at 37°C for 1 hour in a shaking incubator at 200rpm. Digestions were arrested via heating tubes to 80°C for 20 minutes in a Bio-Rad T100™ thermocycler.

Reagent	Volume
XhoI	1µL
NdeI	1µL
10X CutSmart buffer	5µL
5,000U/µL calf-intestine alkaline phosphatase	2.5µL
plasmid/fragment DNA	to 1µg
nuclease-free H_2O	to 50µL

Digested products were cleaned using QIAquick PCR purification kit and analyzed for concentration. Purified digestion products were ligated with T4 DNA ligase into plasmid vector at either a 1:3 or 1:5 plasmid:fragment concentration. Standard protocol can be seen in Table 8.

Table 8. Reagent volumes for DUF4959/pET21b(+) ligation reactions. Reactions were conducted in Bio Basic flat-cap 0.2mL thin-wall PCR tubes cooled to 16°C in a shaking incubator set to 200rpm for 18 hours. Reactions were inactivated by heating tubes to 65°C for 10 minutes.

Reagent	Volume
T4 DNA ligase	1μL
10X T4 DNA ligase buffer	2μL
Vector DNA	to
Insert DNA	to
nuclease-free H_2O	to 20μL

4.1.1 Structural Prediction and Functional Hypothesis for *B. thetaiotaomicron* Proteins

The amino acid sequences for BT2109, BT2857, BT2918, BT2966, and BT3158, were analyzed via a series bioinformatics tools in order to form a hypothesis on the structure and function of the target proteins, as well as to elucidate how these proteins may fit into the broader function of the polysaccharide utilization loci to which they belong. Outputs from InterPro and BLASTp both confirmed the conservation of three domains of interest/unknown function across the five target proteins. The BLASTp illustrations of DUF positions in the proteins of interest can be seen in Figure 5. The N-terminal DUF 4959 is classified under the Pfam family PF16323, and may contain a fibronectin type-III structure. DUF 5126 was found to be a member of Pfam family

Figure 5. Illustration of BLASTp Outputs for Target DUFs. While BLASTp did not register BT2966 as having a DUF5000 domain despite containing the F5/8-type C superfamily, InterPro analysis did confirm the domain as being present.

PF17166 and had no structural motif associated with it. The C-terminal DUF 5000, a member of PF16391, likely contains a short helical F5/F8-type C (discoidin) fold; a fold associated with amphipathic membrane binding (118). The DUF5000 sequences were also found to overlap significantly with galactose-binding-like domain superfamily IPR008979. The IPR008979 superfamily does not have a prototypical structure assigned to it, though a number of 3D XRD/NMR structures from both prokaryotic (ie. PDB IDs: 1CX1, 1EUT, 1DP0) and eukaryotic (ie. PDB IDs: 1BHG, 1GOF, 1I5P) CAZymes have been generated. IPR008979 domains present in these example proteins vary in length (140 – 370 amino acids) and have low sequence similarity but contain similar β-sandwich tertiary structure which forms a jelly-roll fold.

Output from the TMpred online resource showed short sections of a potential transmembrane region covering the first 21 to 24 residues in each target protein, however outputs from the SignalP4.1 server showed signal peptide regions overlapping with these potential transmembrane regions, suggesting the proteins do not possess a transmembrane region following subcellular localization. Results from the PSORTb online resource were inconsistent between the target proteins, with only BT3158 having a result of relatively high confidence, with a score of 8.96/10.00 for cytoplasmic localization, with scores of 0.51 for the cytoplasmic membrane, 0.26 for the periplasm, 0.01 for the outer membrane, and 0.26 for extracellular localization. Three target proteins: BT2109, BT2918, and BT2966 showed scores of 2.00/10.00 for all five potential localizations. BT2857 showed scores of 0.00, 4.90, 2.50, 0.10, and 2.50 for cytoplasmic, cytoplasmic membrane, periplasm, outer membrane, and extracellular localization, respectively, suggesting the protein may have multiple sites of subcellular localization.

The overall best homology models generated by intensive Phyre^2 modelling also showed a diversity in possible structure for each of the five target proteins, despite the conservation in DUFs.

A number of these structures lack a modelled fibronectin type-III domain predicted to be at the N-terminal of the protein. Homology models were superimposed using the MatchMaker function in UCSF Chimera with alignment conducted using the Needleman-Wunsch algorithm and BLOSUM62 matrix with a gap-extension penalty of 1 with distant residue pairs iteratively pruned to a level of 2.0Å or lower. These models can be seen in Figure 6., notably with the C-terminal

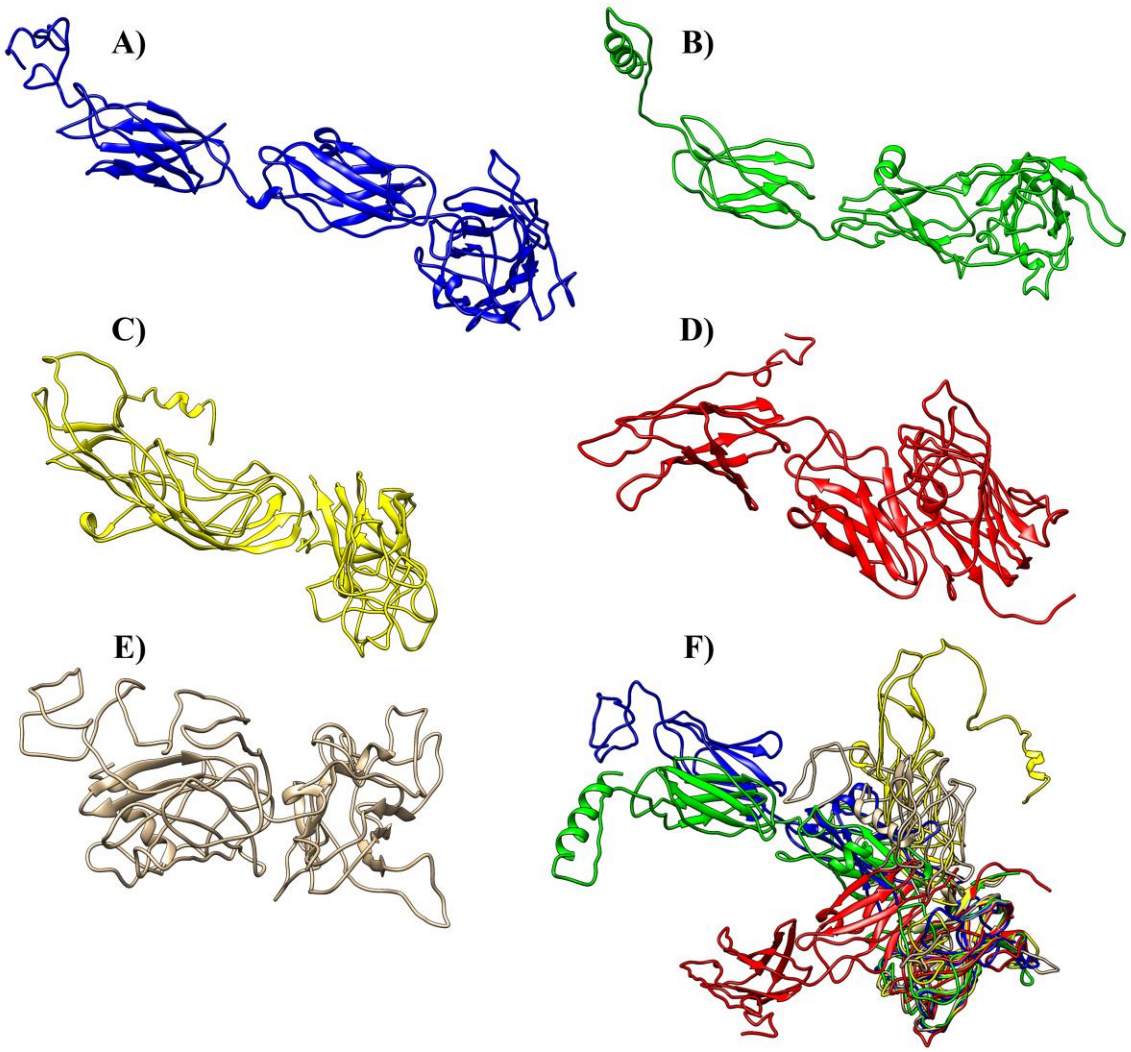


Figure 6. Phyre^2 Intensive Homology Models of *B. thetaiotaomicron* Target Proteins. A), B), C), D), and E) representing Phyre^2 models generated from the amino acid sequences of BT2109, BT2857, BT2918, BT2966, and BT3158, respectively. F) depicts all homology models overlaid based on structural similarity as determined by the UCSF Chimera MatchMaker function, with colours of each target model being identical to those of the A – E.

DUF5000 domain being the center of the alignment for all five of the models. The Match→Align function was used to analyze the similarity of the models after superimposition, with a residue-residue cutoff of 10.0Å used and alignment iterated until convergence across the whole alignment. This analysis generated values of 1.937Å for the RMSD, 47.9 for Sequence Distance-Measure, and 0.004 for Overall Quality (out of a possible 1.000). Notably, of the 5 targets that were modelled only the output for BT2109 had a region that overlapped with DUF4959 and showed significant resemblance to a fibronectin type-III domain. Protein target sequences were submitted to the RaptorX homology modelling server to validate the structures generated by Phyre[2]. Models produced by RaptorX can be seen in Figure 7.

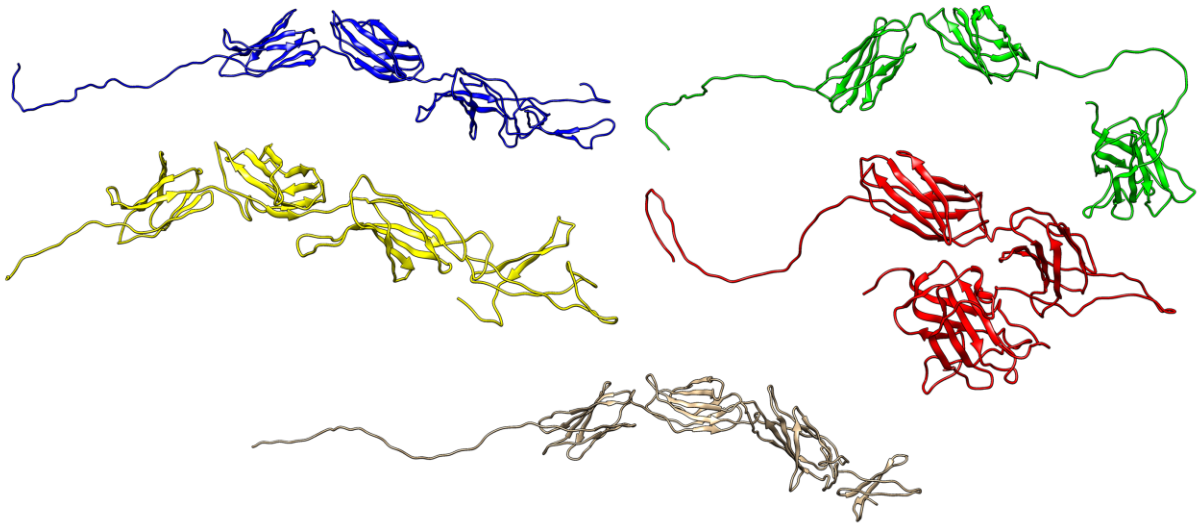


Figure 7. RaptorX Homology Models of *B. thetaiotaomicron* Target Proteins. A), B), C), D), and E) representing BT2109, BT2857, BT2918, BT2966, and BT3158, respectively. Notable differences from Phyre[2] includes: DUF4959 being modelled similarly to a fibronectin type-III domain in BT2109, BT2857, and BT3158; more extensive secondary-structure character across the sequences; distension of sequence segments that do not contain secondary structure character.

4.1.2 Model Prediction and Functional Hypothesis for *P. gingivalis* Proteins

The amino acid sequences of PGN1176, PGN1459, and PGN1461 were analyzed via the same bioinformatics tools as those used on the *B. thetaiotaomicron* proteins, also for the purpose of forming a hypothesis regarding the structure and function of the proteins, as well as hypothesize how they may contribute to the pathogen virulence for the purpose of building a better model of periodontal disease. Outputs from InterPro and BLASTp were used to either verify the putative functional annotation in the UniProt entry for the protein or form preliminary hypotheses for entries with no proposed function. The BLASTp results can be seen in Figure 8.

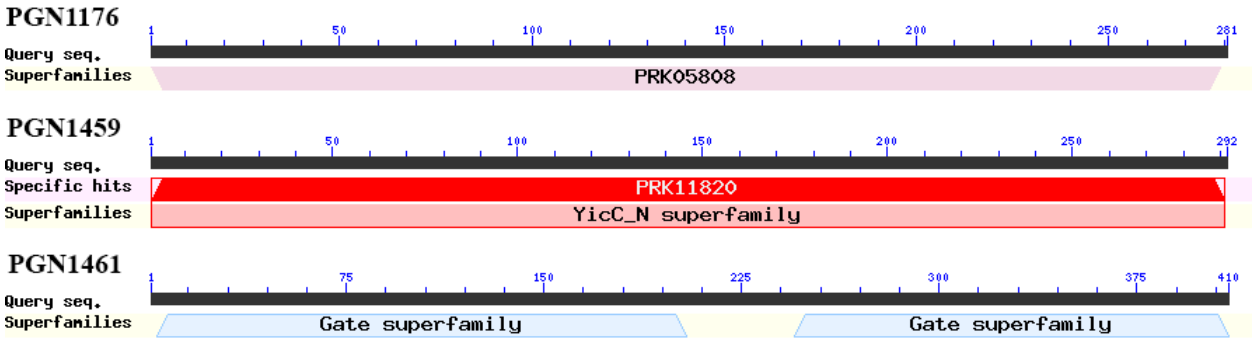


Figure 8. Illustrated BLASTp Domain Results for *Porphyromonas gingivalis* Target Proteins. Detailed results regarding all (non-)specific hits and some superfamily membership omitted for clarity. Based on sequence similarity Pgn1176 is suggested to belong to the PRK05808 (3-hydroxybutyryl-CoA dehydrogenase) superfamily, Pgn1459 is suggested to belong to the YicC uncharacterized conserved protein superfamily, and Pgn1461 is suggested to have two Gate superfamily domains found in nucleoside transporter FeoB GTPases.

In the case of PGN1176, the PRK05808 superfamily identified by BLASTp agreed with the UniProt proposed function of being a 3-hydroxybutyryl CoA-dehydrogenase, while a Specific Hit in BLASTp suggested the protein contains a FadB domain. The InterPro output matched with these results, identifying the IPR022694 family to cover the full length of the protein. An NAD(P)+ binding domain (a member of the adenosylmethionine-dependent methyltransferase superfamily) was found to be located between amino acids 2 and 181 by both BLASTp and InterPro. InterPro and BLASTp outputs for PGN1459 both identified the protein to contain a YicC-like domain

stretching across the full length of the sequence (InterPro family IPR005229, pfam PF03755), a domain that has been associated with maintaining proper morphology of bacterium in which it is expressed in stationary-phase growth, and required for growth under high temperatures (37°C - 45°C). InterPro output for PGN1459 also showed the presence of Domain of Unknown Function 1732 between amino acids of 208 and 291 (pfam: PF08340), a domain that may bind nucleic acids. PGN1461 was found to be annotated as a putative fused type A/B spore-maturation protein. BLASTp output for PGN1461 agreed with UniProt annotation, with the spore maturation domain SpmA covering amino acid residues 3 - 204, and SpmB covering residues 245 - 410. A suspected nucleoside gate domain is present within both the SpmA and SpmB domains according to the BLASTp analysis, at residues 52 - 155 and 278 - 379, respectively. InterPro outputs further validated the BLASTp and UniProt results, however the server only identified a single gate domain at residues 52 - 154. Gate domains in PGN1461 were also noted to have homology to those found in FeoB iron transport membrane proteins responsible for Fe^{2+} uptake(119). The SignalP4.1 server identified no signal peptide in the amino acid sequences for PGN1176, PGN1459, or PGN1461. TMPred predicted two possible transmembrane helices in PGN1176, one in outside→inside orientation at residues 1 – 19, and one in inside→outside orientation at residues 226 – 246, results which contradicted PSORTb results, which predicted cytosolic localization with a 99.7% degree of confidence (cytosolic membrane and periplasm localization were predicted with 0.1% and 0.1% confidence, respectively). For PGN1459 TMPred predicted no transmembrane domains, and subcellular localization as predicted by PSORTb showed the same results as PGN1176. TMpred predicted PGN1461 to contain a total of 10 transmembrane helices in alternating inside→outside/outside→inside orientation throughout the length of the protein, agreeing with

results from PSORTb analysis showing a 100% chance of the protein being localized to the cytosolic membrane. PGN1461 TMPred results can be seen in Figure 9.

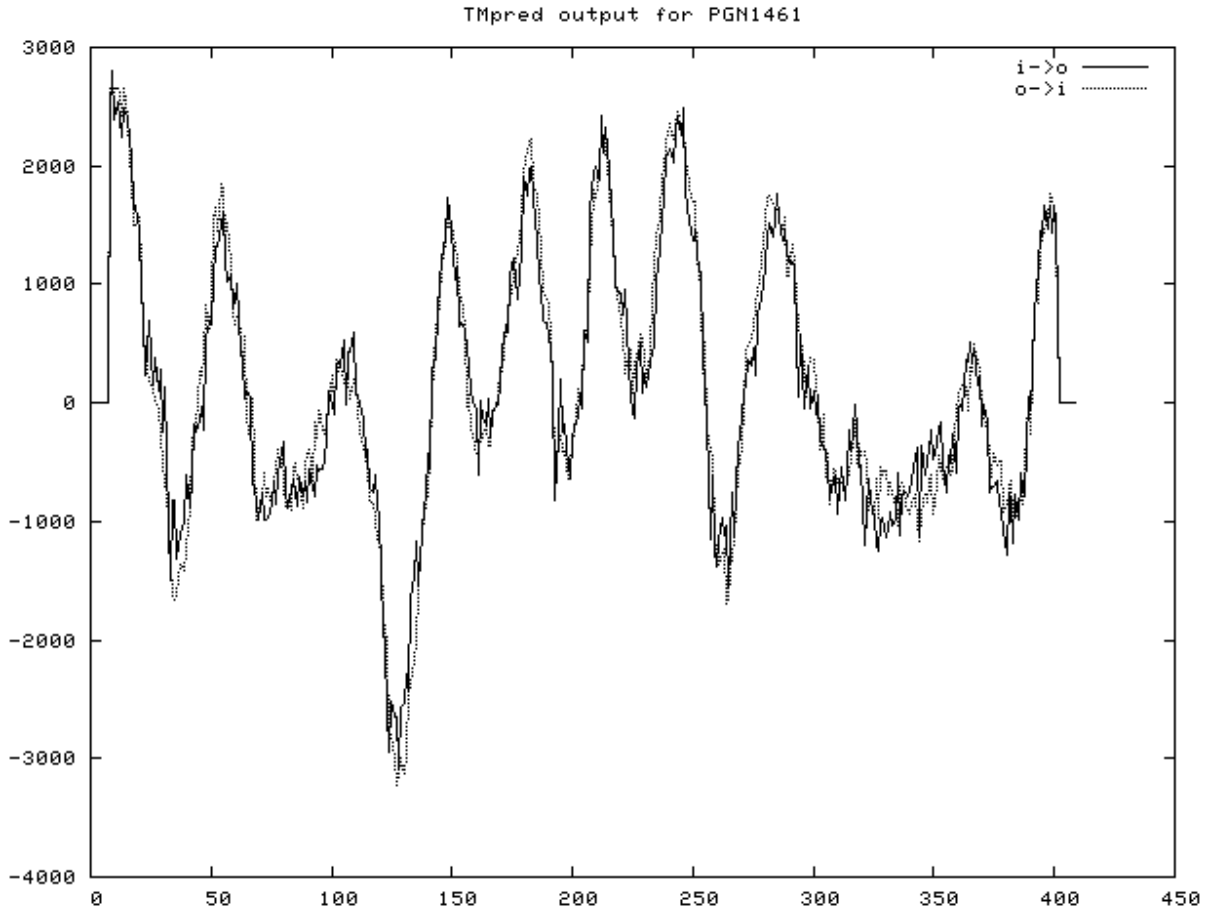


Figure 9. TMPred Output for *Porphyromonas gingivalis* Target Protein PGN1461. Values above 0 indicate regions of probable transmembrane character in the protein of interest. Legend can be seen in the top right corner, with i→o and o→i indicating inside→outside and outside→inside orientation predictions, with "inside" being cytosol and "outside" being periplasm. The preferred model determined the first transmembrane helix to be inside→outside, with subsequent helices alternating orientation.

Amino acid sequences for each target protein were submitted to the Phyre[2] and RaptorX servers for homology modelling to assess possible conformations of the proteins and find previously characterized proteins that may be used as templates for molecular replacement. PGN1176 was modelled with a high degree of confidence in Phyre[2], with 100% of residues being modelled at above 90% confidence, with the best template being another probable 3-

hydroxybutyryl-CoA dehydrogenase (PDB ID: 3MOG) from strain K12 *E. coli*, with which it shares 41% amino acid sequence identity. The model generated by RaptorX was based on the template 6ACQ, another 3-hydroxybutyryl-CoA dehydrogenase, from *Clostridium acetobutylicum*, with 100% of residues modelled having a Global Distance Test (GDT) score of

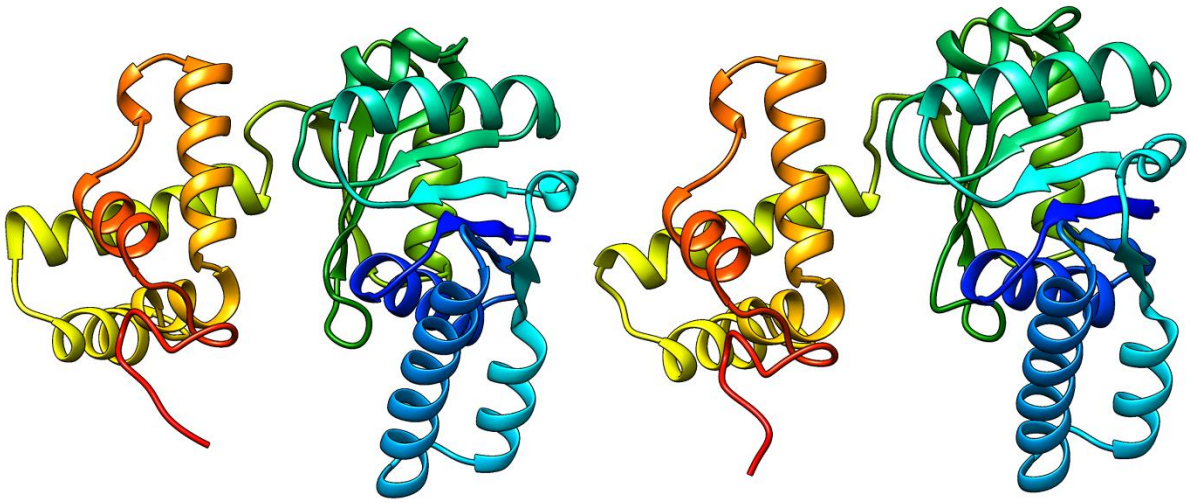


Figure 10. Phyre^2 (left) and RaptorX (right) Homology Models of PGN1176. Figures were generated and models compared using UCSF Chimera. Match→Align analysis of superimposed structures with parameters of 5.0Å cutoff resulted in an overall RMSD of 0.892Å; sequence length of 281; SDM of 8.912; and Quality score of 0.919, showing that the generated structures possess a high degree of similarity.

84, greatly exceeding the threshold of acceptability of 50. The homology models for PGN1176 can be seen Figure 10.

PGN1459 was modelled with a low degree of confidence in both Phyre^2 and RaptorX; with Phyre^2 modelling no residues at >90% confidence and the top template having only 62.8% confidence despite having 36% amino acid sequence identity, while the RaptorX model had a GDT significantly below the acceptable threshold of 50, at only 29. The RaptorX model had an unnormalized GDT of 84; indicating sections of the structure may be accurate to the structure. The Phyre^2 and RaptorX servers both modelled potential structures for PGN1459 on template proteins dominated by α-helical secondary structures and often involved in cellular signalling (6QAJ and

2D1L in Phyre[2]; 3I9W in RaptorX). Homology models generated by these resources can be seen in Figure 11.

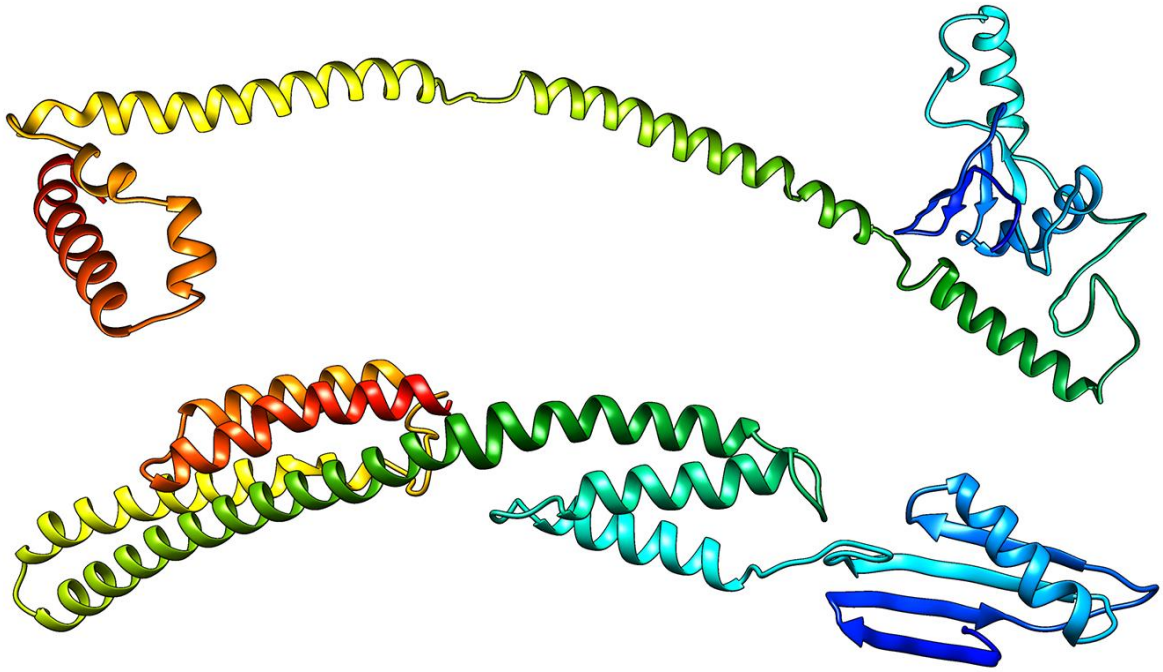


Figure 11. Phyre[2] (top) and RaptorX (bottom) homology models of PGN1459. Phyre[2] model based primarily on NMR solution structure of human calcium-binding non-muscle myosin IIA (PDB ID: 2LNK) and X-ray diffracted solution of human nuclear protein KAP1/TRIM28 (PDB ID: 6QAJ). RaptorX model based on X-ray diffraction structure of strain 12 *E. coli* histidine kinase sensor domain (PDB ID: 3I9W).

PGN1461 was modelled with moderate levels of confidence using Phyre[2], mainly across the sections of the primary sequence predicted to be nucleoside gate domains. Primary templates for Phyre[2] homology model were two nucleoside permease proteins, 5L2B from strain 12 *E. coli* and 3TIJ from *Vibrio cholera*, with 87.8%/83.2% homology confidence and 21%/18% sequence identity to PGN1461, respectively. Quality indicators for the RaptorX also predicted moderate confidence in the overall homology model, with an unnormalized GDT of 85 and GDT of 20, indicating sections of the model are likely accurate, though not the global structure. Top ranked

templates from both servers used to build PGN1461 homology models consisted of transmembrane transporters (including concentrative nucleoside transporters and aminobenzoyl-glutamate (AbgT) family transporters), the same class of proteins that PGN1461 was predicted to be a member of according to previous bioinformatics results. Also of note was that all top templates exist as trimers in crystal form according to the associated literature, suggesting the proteins may form a trimeric channel-type transporter with a variety of potential substrates in an intracellular environment(120–123). Model predictions for single polypeptide strands of PGN1461 can be seen in Figure 12.

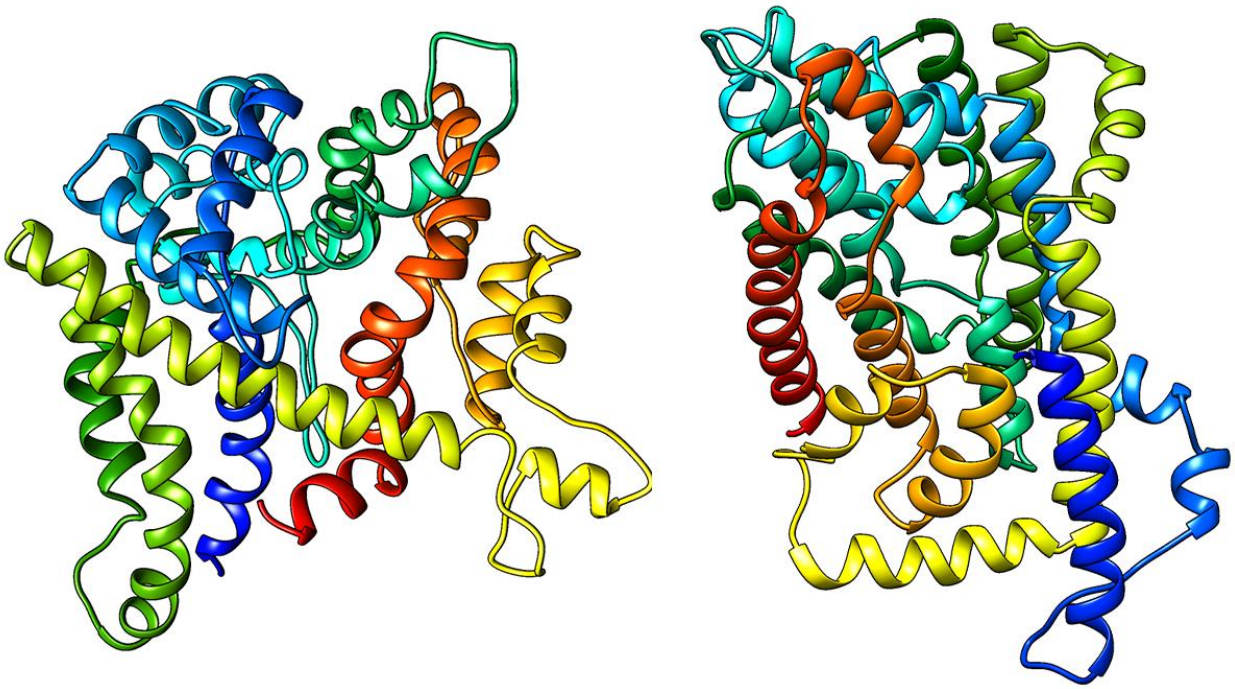


Figure 12. Phyre^2 (left) and RaptorX (right) homology models of PGN1461. Sequences were determined to be highly dissimilar according to a Match→Align analysis using a 5.0Å cutoff in UCSF Chimera, with results of 2.621Å for RMSD, 410 for sequence length, 113.386 for SDM, and 0.032 for Overall Quality.

4.2.1 Expression and Purification of *B. thetaiotaomicron* Proteins

The first step in the characterization of the *B. thetaiotaomicron* enzymes was optimizing the expression and purification of the proteins. Expression of BT2109, BT2857, and BT3158 was conducted in 1L high-salt (10mg/mL) LB cultures, supplemented with 100mg/L ampicillin and induced with IPTG at a final concentration of 1mM, using BL21 (DE3) *E. coli* cells. BT2918 and BT2966 were expressed in Rosetta™(DE3) *E. coli* cells in 1L high-salt LB cultures, supplemented with chloramphenicol and ampicillin (BT2918) or kanamycin (BT2966), and induced with 1mM IPTG. Cell lysis using solutions buffered with 50mM Tris:HCl pH8.0 and containing 300mM NaCl in addition to standard sucrose and detergent concentrations (730mM and 1% deoxycholate/2& Triton X-100, respectively) was found to be appropriate for all BT proteins. After homogenized cell lysate was centrifuged supernatant was removed and subjected to IMAC isolation. The standard IMAC resin incubation period used was 60 minutes, but could range from immediate flow through to 2 hours depending on what length of time was found to be ideal for the individual construct being isolated: 0 minutes for BT2109; 60 minutes for BT2857, BT2918, and BT3158; 2 hours for BT2966. SDS-PAGE analysis of collected IMAC fractions can be seen in Figure 13 – 14.

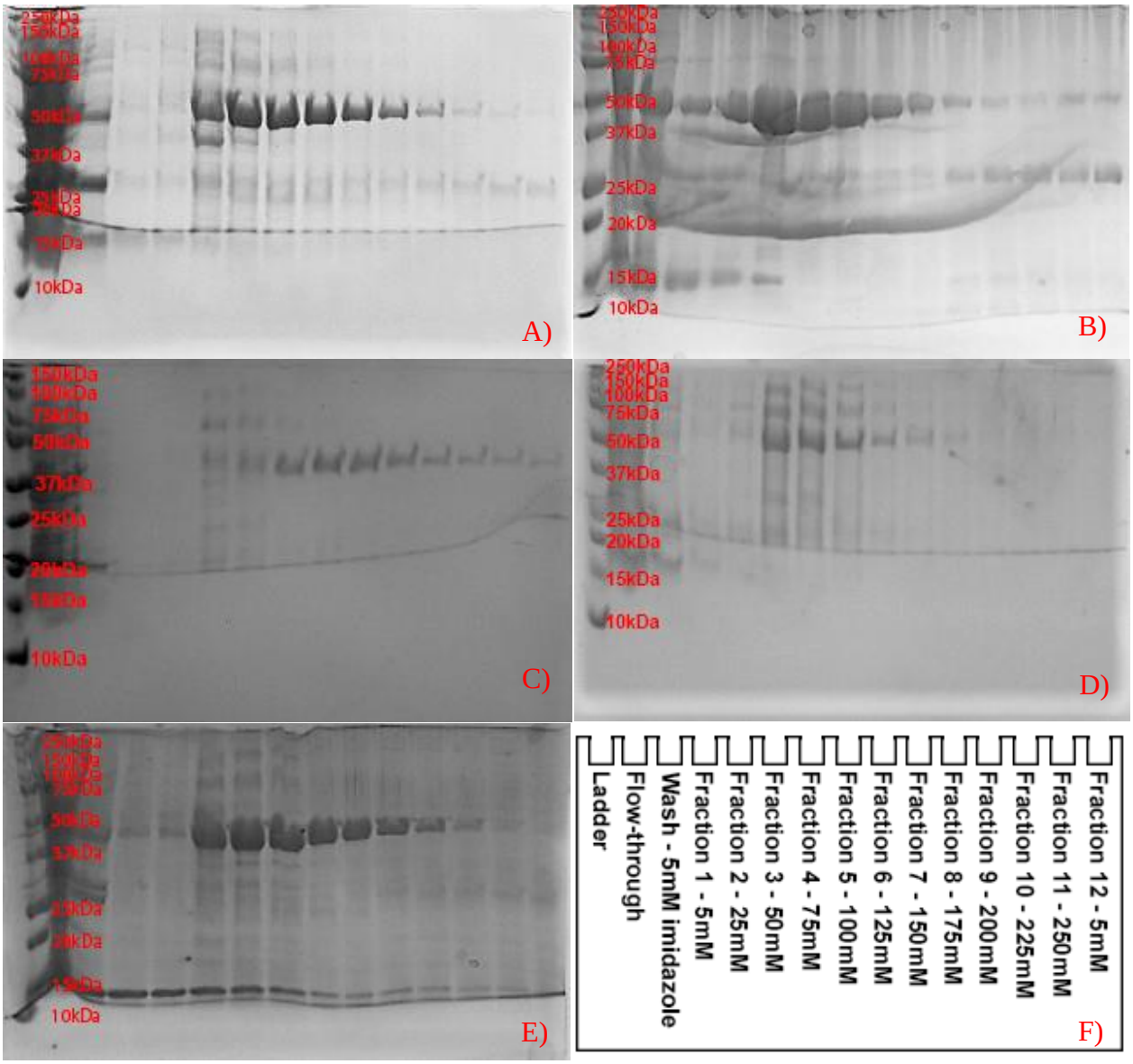


Figure 13. SDS-PAGE Gels for the IMAC Isolation of Full-Length *B. thetaiotaomicron* Proteins. 12% (w/v) SDS-PAGE gels were loaded with prepared elution samples of increasing imidazole concentration. Gels A), B), C), D), and E) correspond to isolations of BT2109, BT2857, BT2918, BT2966, and BT3158, respectively. Legend for lanes in each gel is depicted in F). Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1. All constructs expected to be found at ~45kDa.

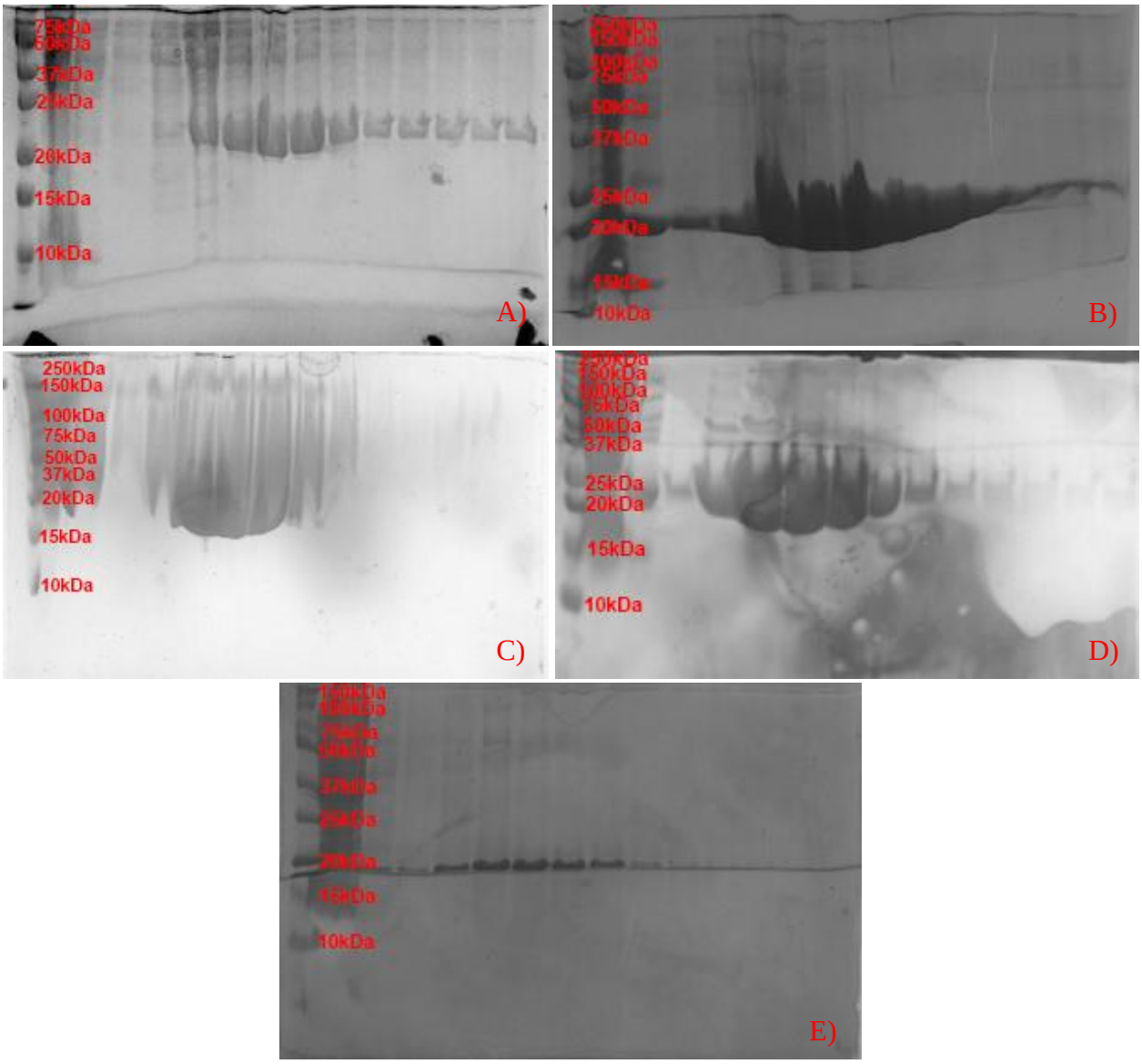


Figure 14. SDS-PAGE Gels for the IMAC Isolation of Truncated and Truncated Mutant BT2857 Proteins. 12% (w/v) SDS-PAGE gels were loaded with prepared elution samples of increasing imidazole concentration. Gels A), B), C), D), and E) correspond to isolations of BT2857N, BT2857C, BT2857C D256A mutant, BT2857C D239A-D242A mutant, and BT2857C H300A-D304A mutant, respectively. Order of elution/imidazole concentration is identical to that of the full-length *B. thetaiotaomicron* isolations. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1. BT2857N expected to be found at 22kDa and BT2857C constructs at 19kDa.

Following IMAC isolation fractions 3 - 8 were pooled for each full-length construct except BT2918, where fractions 5 - 12 were pooled, and dialyzed into salt-free Tris:HCl buffer prior to anion-exchange chromatography. Notably, BT2918 consistently formed a higher amount of visible

inclusion bodies than other constructs over the course of dialysis. Any protein samples that had visible inclusion bodies following dialysis were centrifuged at 4,300xg for 20 minutes to remove inclusions. IMAC fractions of BT2857C and its associated mutants were determined to be at apparent homogeneity after only this initial isolation, and did not undergo anion-exchange FPLC. Examples of anion-exchange FPLC chromatograms and associated SDS-PAGE analysis of protein purification can be seen in Figures 15 – 19.

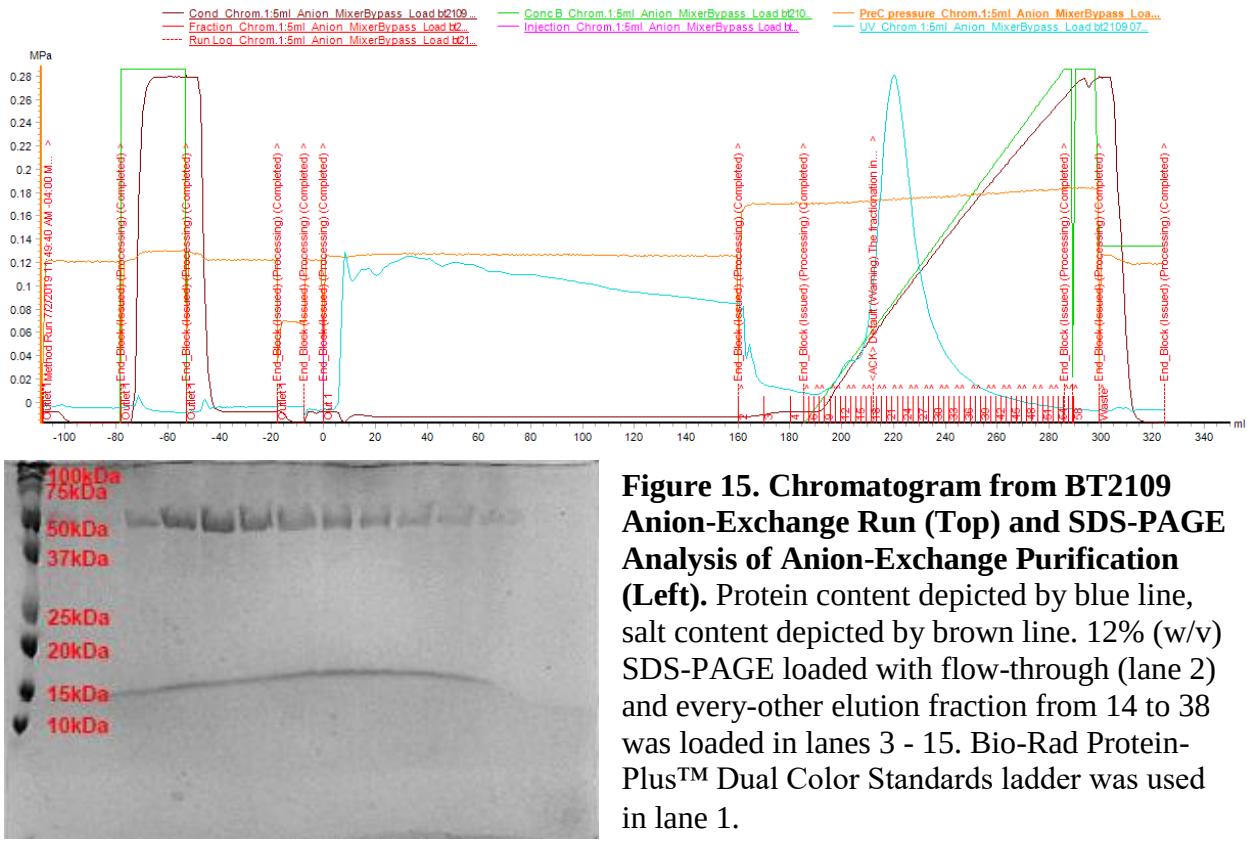


Figure 15. Chromatogram from BT2109 Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). Protein content depicted by blue line, salt content depicted by brown line. 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.

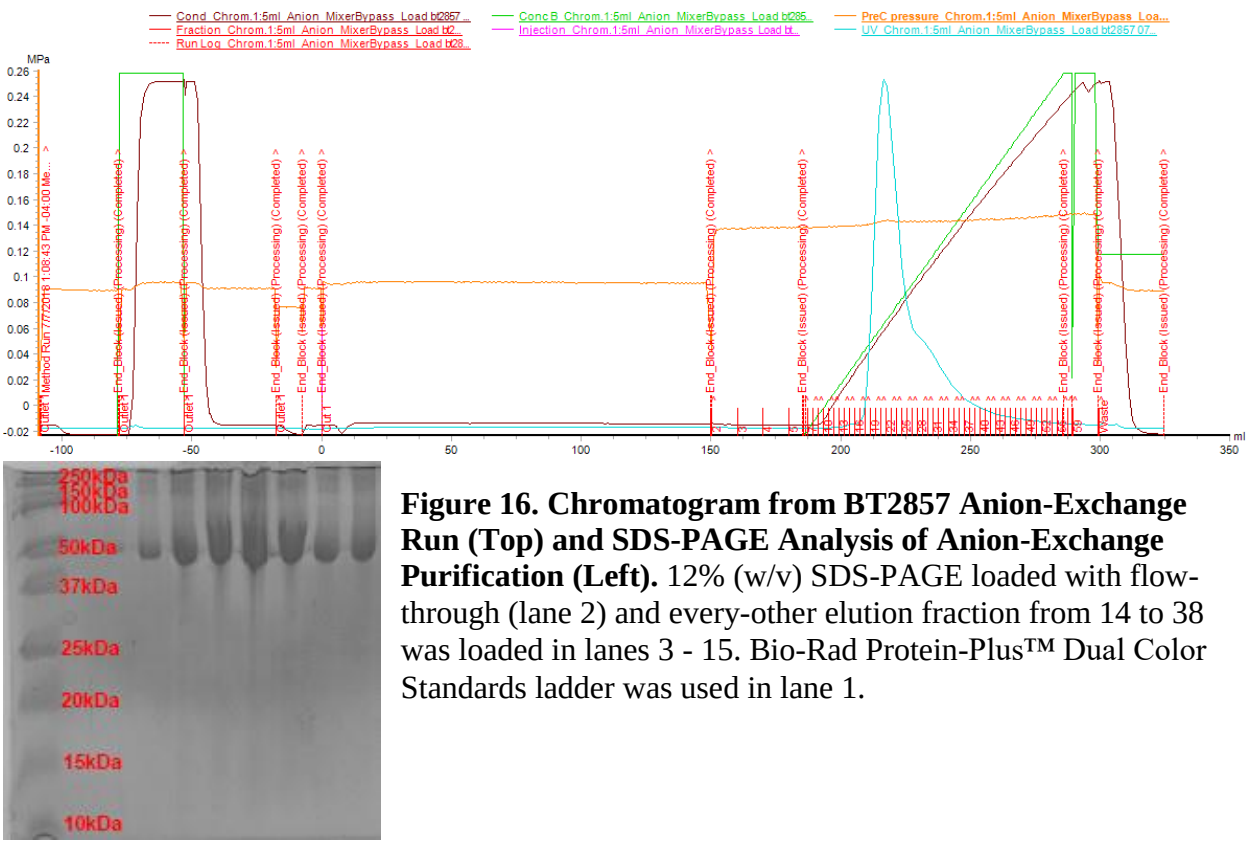


Figure 16. Chromatogram from BT2857 Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.

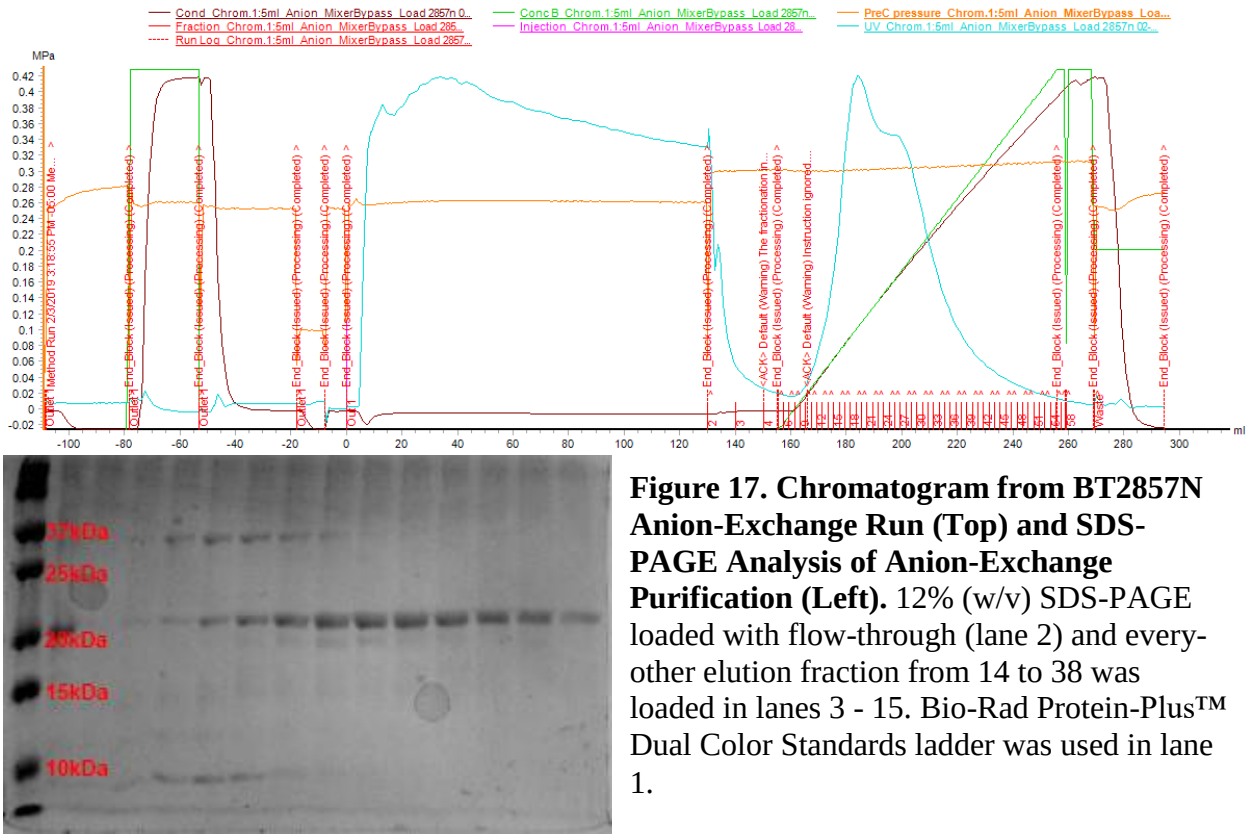
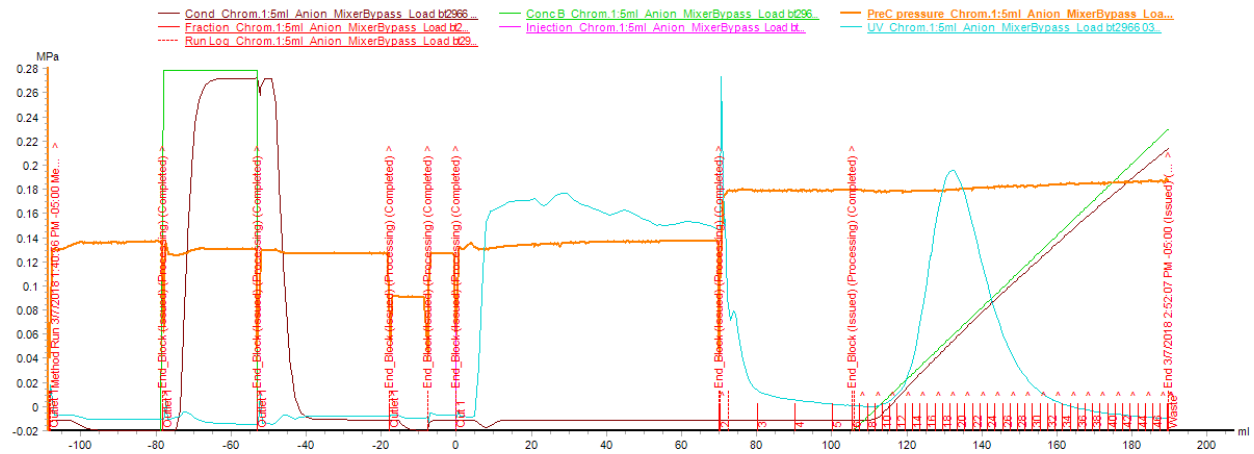


Figure 17. Chromatogram from BT2857N Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.



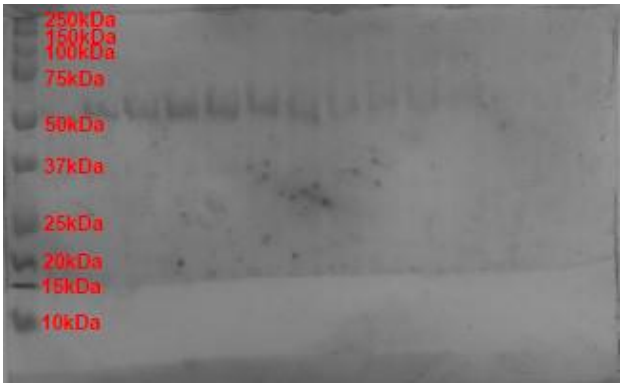
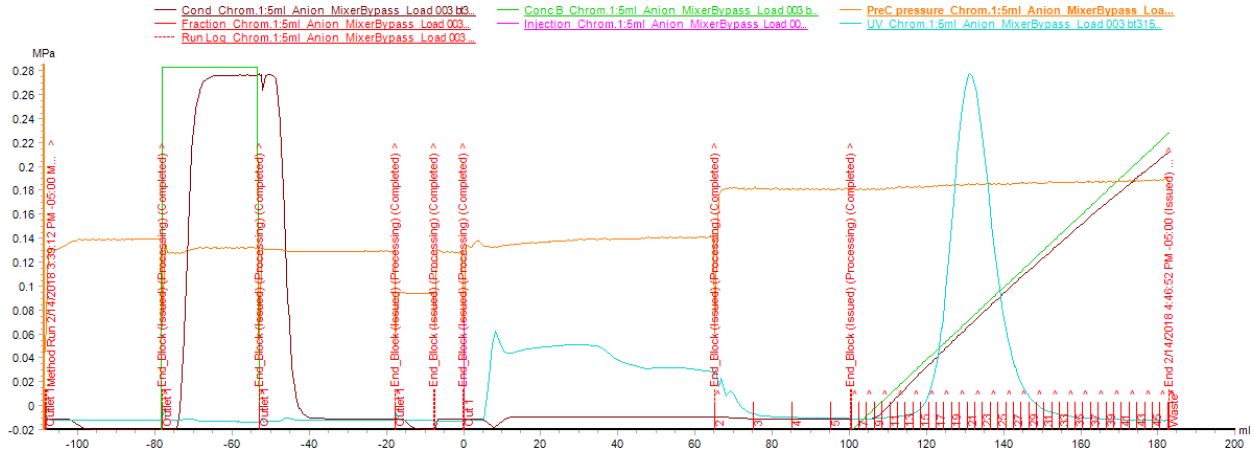


Figure 18. Chromatogram from BT2966 Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.



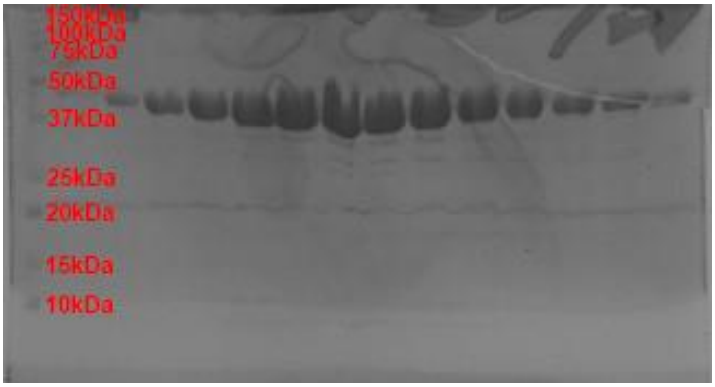


Figure 19. Chromatogram from BT3158 Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.

4.2.2 Expression and Purification of *P. gingivalis* Proteins

As with *B. thetaiotaomicron* proteins, expression of *P. gingivalis* targets was the first step in characterization. PGN1176 was expressed in high-salt 1L LB cultures, PGN1461 in 500mL 2YT cultures, and PGN1459 in 1L autoinduction-LB cultures with BL21 (DE3) *E. coli* cells. All cultures were supplemented with 100mg/L ampicillin and induced with IPTG at 1mM final concentration. 50mM Tris:HCl pH8.0 buffered lysis solutions containing 300mM NaCl were determined to be sufficient for solubilisation of PGN1176, while 100mM HEPES:NaOH pH7.5 with 750mM NaCl and 300mM NaCl was found to optimally solubilize PGN1459 and PGN1461, respectively. Cell lysate was initially subjected to IMAC isolation with resin incubations varied from 30 to 60 minutes depending on the construct being purified. SDS-PAGE gels can be seen in Figure 20.

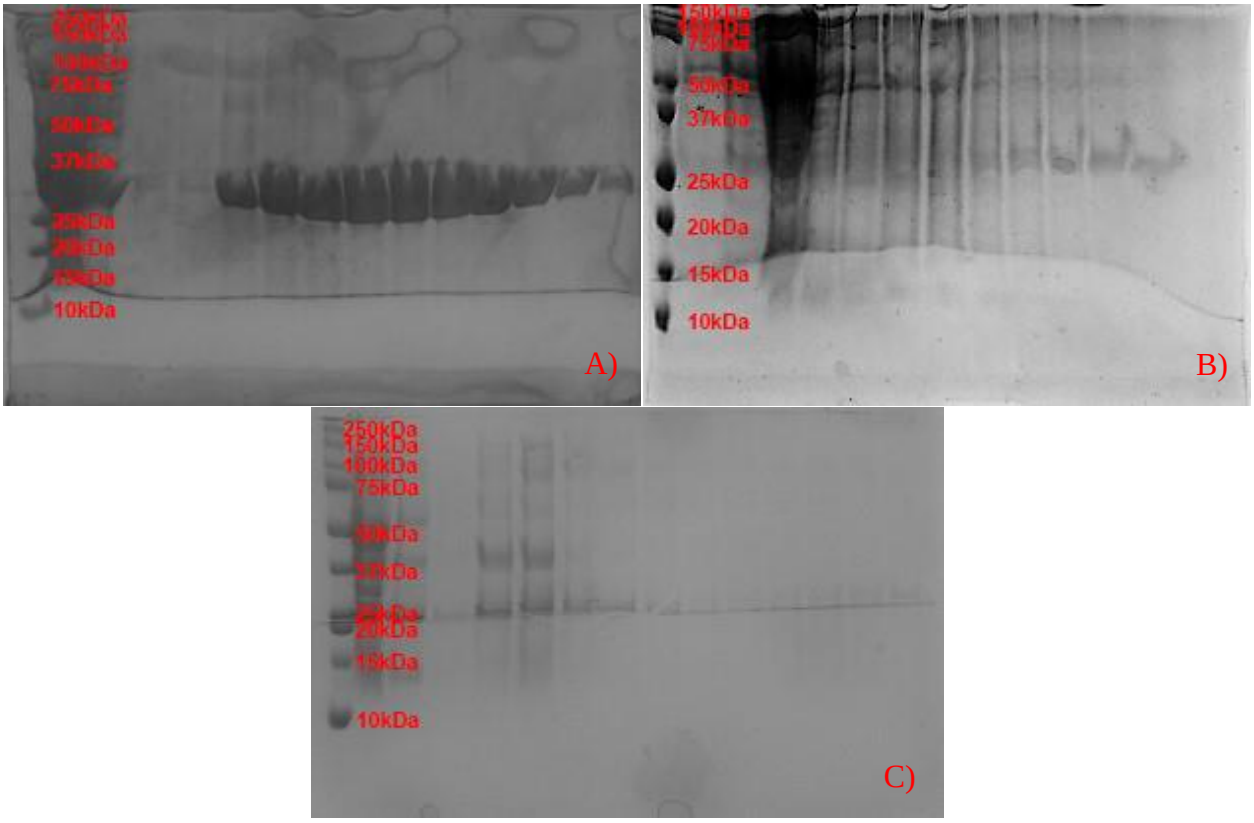


Figure 20. SDS-PAGE Gels for the IMAC Isolation of *P. gingivalis* Proteins. 12% (w/v) SDS-PAGE gels were loaded with prepared elution samples of increasing imidazole concentration. Gels A), B), C) correspond to isolations of PGN1176, PGN1459 and PGN1461, respectively. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1. Banding for PGN1176 expected at 30kDa, PGN1459 at 34kDa, and PGN1461 at 44kDa.

For constructs PGN1176 and PGN1459 fractions 6 – 12 (125mM – 250mM imidazole) and 8 – 12 (175mM – 250mM imidazole), respectively, were determined to be of high enough purity and yield to proceed to functional/crystallization trials without requiring further purification via FPLC. PGN1461 was dialyzed into salt-free 100mM HEPES:NaOH pH7.5 prior to FPLC purification. Cation-exchange chromatography was initially pursued as a purification method for PGN1461 as the theoretical pI of the protein was found to be 8.97 in ProtParam analysis; however it was found that the protein failed to bind to the column and anion-exchange was used in subsequent

purifications. Chromatograms and associated SDS-PAGE gels for these purifications can be seen in Figures 21 - 22.

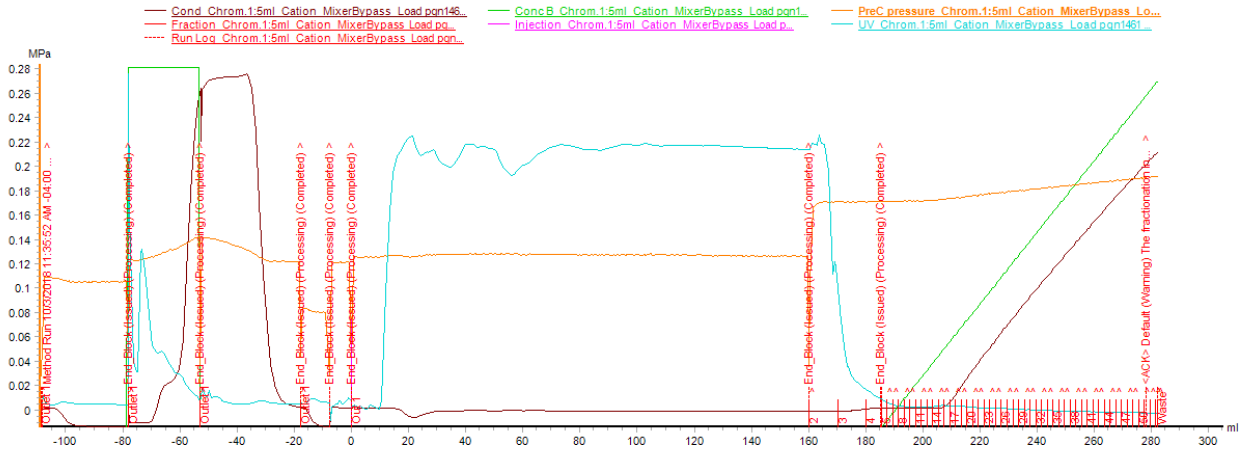
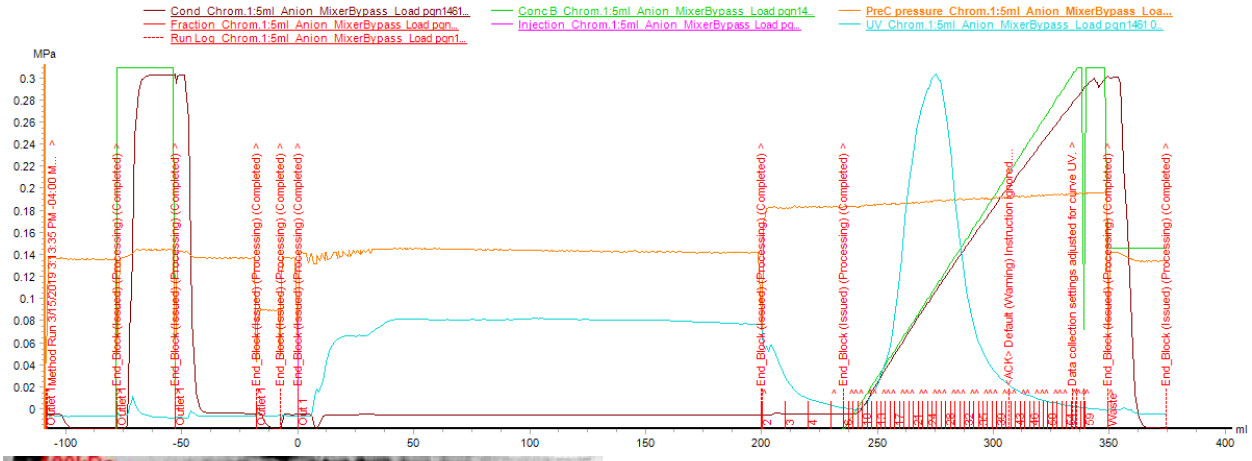


Figure 21. Chromatogram of PGN1461 Cation-Exchange FPLC Purification Attempt. Run was stopped and column was cleaned after it became apparent the protein had not bound to the cation exchange column. Protein content via UV 280nm absorbance depicted by blue line, salt content depicted by brown line. No PGN1461 adhered to the column.



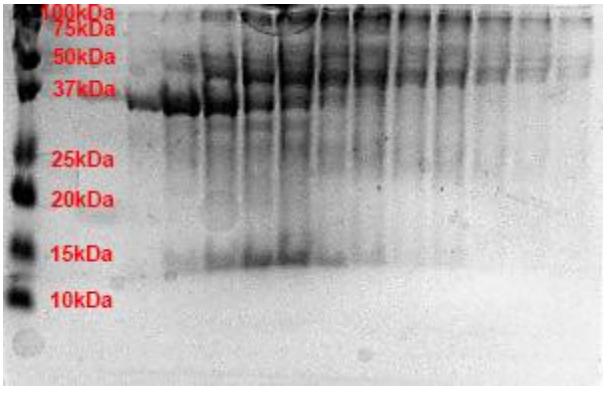


Figure 22. Chromatogram from PGN1461 Anion-Exchange Run (Top) and SDS-PAGE Analysis of Anion-Exchange Purification (Left). 12% (w/v) SDS-PAGE loaded with flow-through (lane 2) and every-other elution fraction from 14 to 38 was loaded in lanes 3 - 15. Bio-Rad Protein-Plus™ Dual Color Standards ladder was used in lane 1.

4.3 Kinetic and Mechanistic Characterization of *B. thetaiotaomicron* Proteins

Preceding the pNP-carbohydrate functional assay of BT proteins extinction coefficients for para-nitrophenol at each of the pH values being tested were generated. A series of kinetic and functional assays were run on the *B. thetaiotaomicron* enzymes BT2857, BT3158, and truncated/mutant truncations of BT2857. The first parameter tested was the pH at which BT2857, BT2857 truncations, and BT3158 achieved optimal activity against the substrate 4-nitrophenyl-β-D-galactopyranoside. It was found that BT2857 and both BT2857 truncations had maximal activity pH8.4, while BT3158 was found to be most active at pH8 (Figures 23 - 24). Activity of the enzymes against the substrate was confirmed using protein that was expressed in lacZ-deficient *E. coli* Tuner(DE3) cells, as this cell line lacks an endogenous β-galactosidase that may contribute to background pNP-β-D-galactosidase hydrolysis. Enzymes were also tested against the substrate 4-nitrophenyl-α-D-galactopyranoside at pH8.4 for BT2857 constructs and pH8.0 for BT3158, which did not yield absorbance increases above background. pH ranges displayed were limited to 5.8 – 9.2 as at pH values higher or lower than resulted in visible precipitation of protein out od solution.

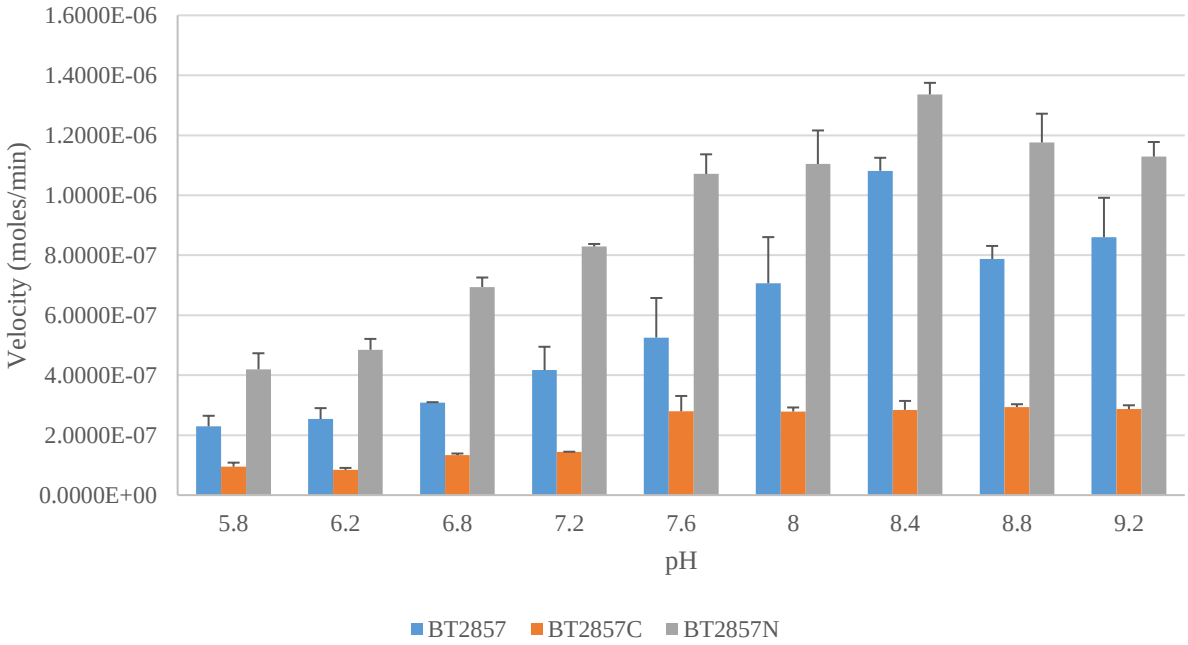


Figure 23. pH Optima Kinetic Assay of Full-Length and N/C-Terminal Truncations of BT2857. Triplicate reactions were conducted in a McIlvaine buffering system up to pH8.0 and 100mM phosphate buffer at pH 8.4 - 9.2 with 1mM pNP-β-Gal substrate and a final concentration of 25µM for each protein. Final volume of each reaction was 250µL at a temperature of 22°C.

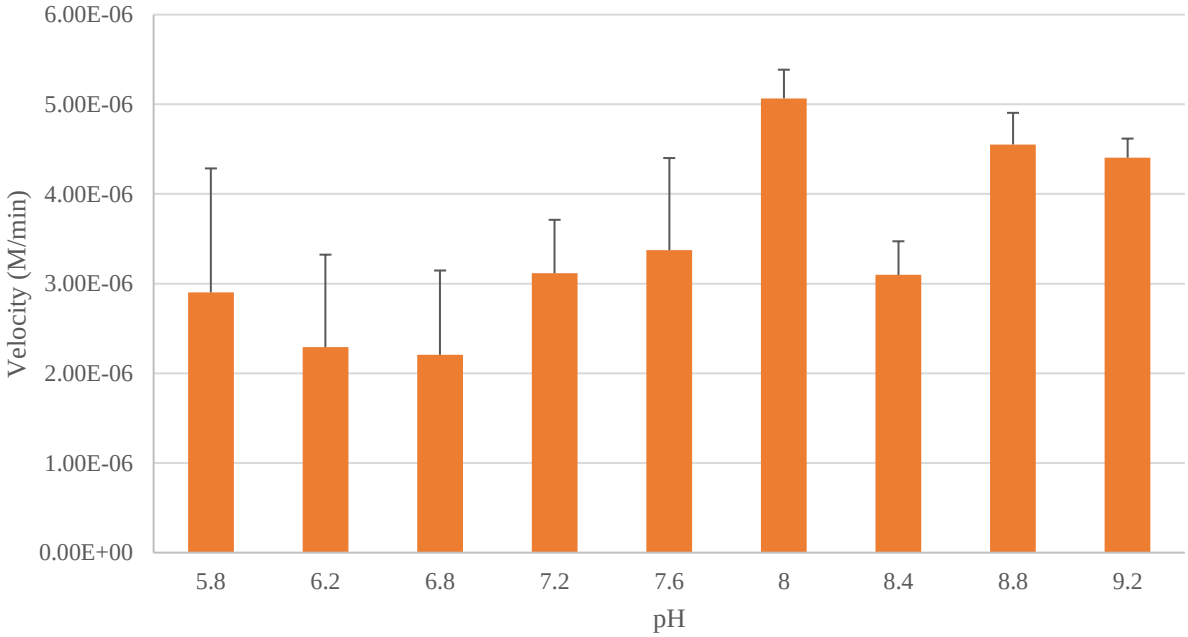


Figure 24. pH Optima Kinetic Assay of BT3158. Triplicate reactions were conducted in a McIlvaine buffering system up to pH8.0 and 100mM phosphate buffer at pH 8.4 - 9.2 with 1mM pNP-β-Gal substrate and a final protein concentration of 5µM. Final volume of each reaction was 250µL at a temperature of 22°C.

Following pH optima analysis of BT2857, BT2857 truncations, and BT3158, Michaelis-Menten kinetic assay of both full-length constructs and BT2857 truncations was done, as well as for three mutants of BT2857C: D239A-D242A, D256A, and H300A-D304A. The Michaelis-Menten curves for the tested proteins can be seen in Figures 25 - 26.

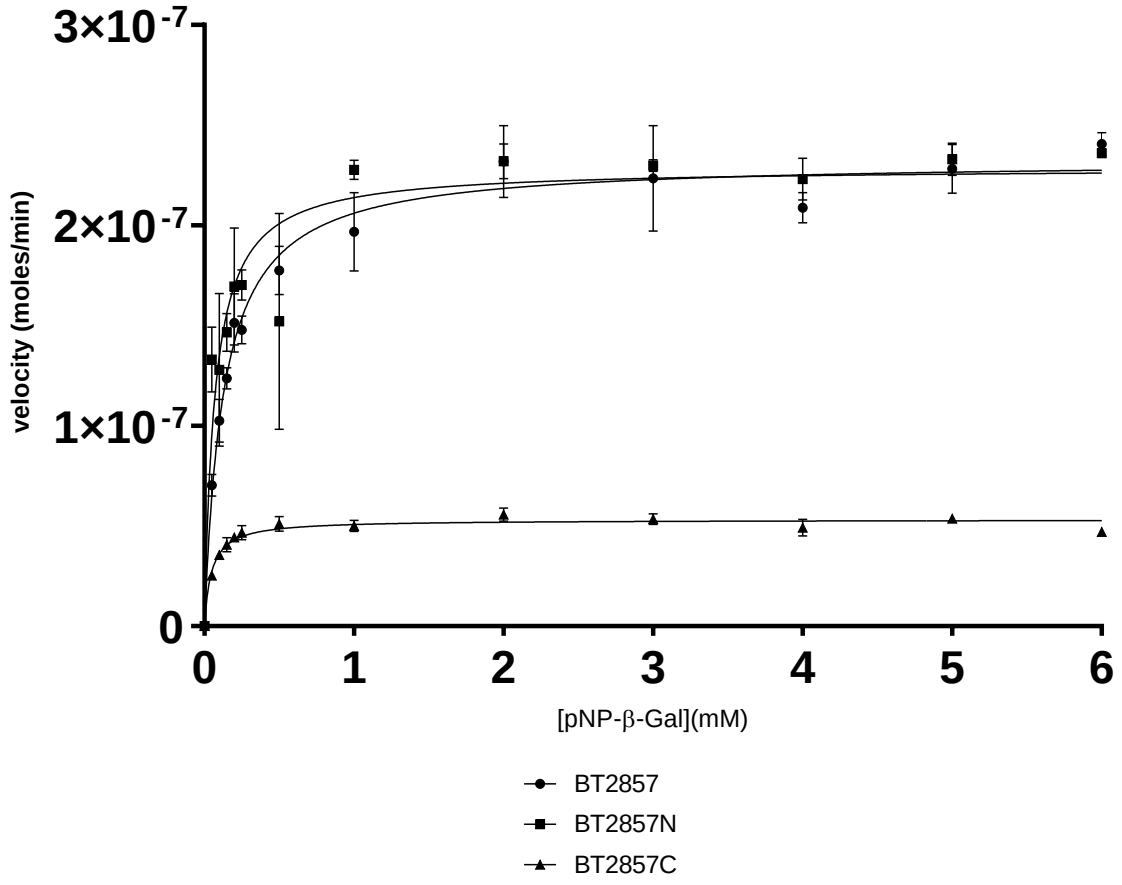


Figure 25. Michaelis-Menten Analysis of BT2857 and BT2857 Truncations. Reactions totalled 250µL with each protein being present at a final concentration of 5µM and were conducted in triplicate. Solutions were buffered with 100mM phosphate at pH 8.4 and a temperature of 22°C. BT2857 was determined to have a V_{max}, K_M, and k_{cat} of 2.32×10^{-7}moles/min ($\pm 3.83 \times 10^{-9}$), 0.13mM (±0.010), and $0.046 s^{-1}$, respectively. BT2857N was determined to have values of 2.29×10^{-7}moles/min ($\pm 6.59 \times 10^{-9}$), 0.07mM (±0.012), and $0.046 s^{-1}$; while BT2857C generated values of 5.31×10^{-8}moles/min ($\pm 8.09 \times 10^{-10}$), 0.047mM (±0.005), and $0.011 s^{-1}$.

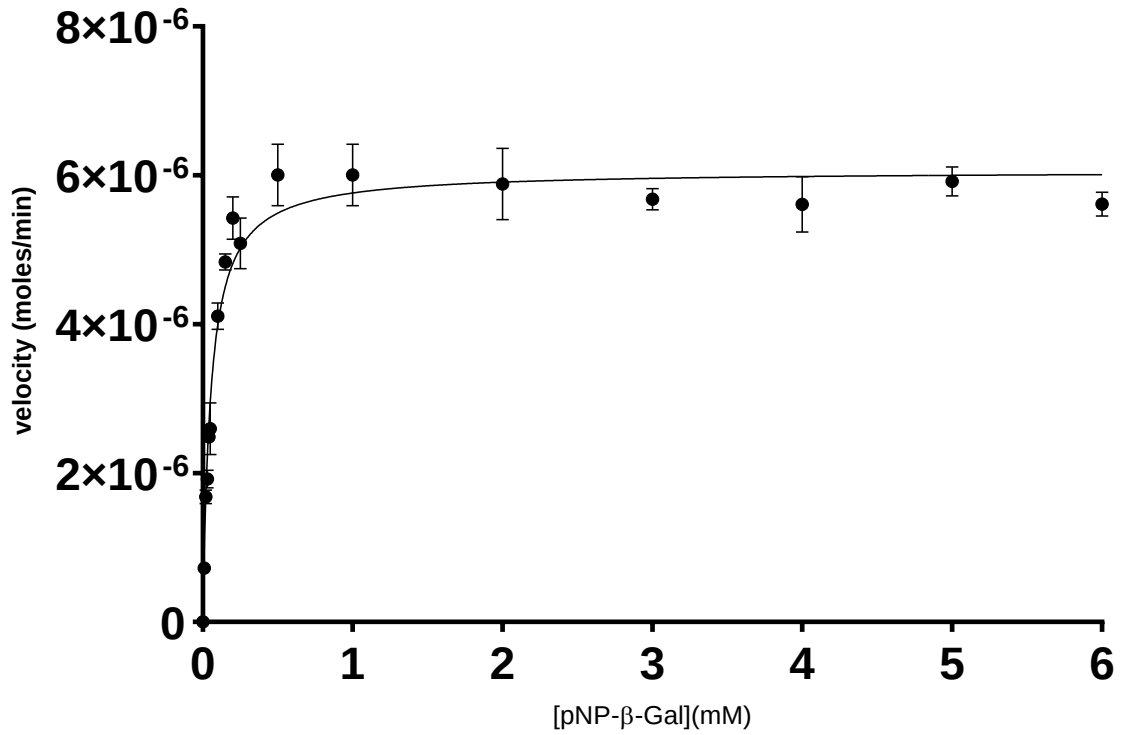


Figure 26. Michaelis-Menten Analysis of BT3158. Reactions totalled 250µL with protein being present at a final concentration of 5µM and conducted in triplicate. Solutions were buffered with 100mM phosphate at pH 8.0 and a temperature of 22°C. BT3158 was determined to have a V_{max}, K_M, and k_{cat} of 6.06×10^{-6}moles/min ($\pm 1.7 \times 10^{-7}$), 0.052mM (±0.007), and $1.21 s^{-1}$, respectively.

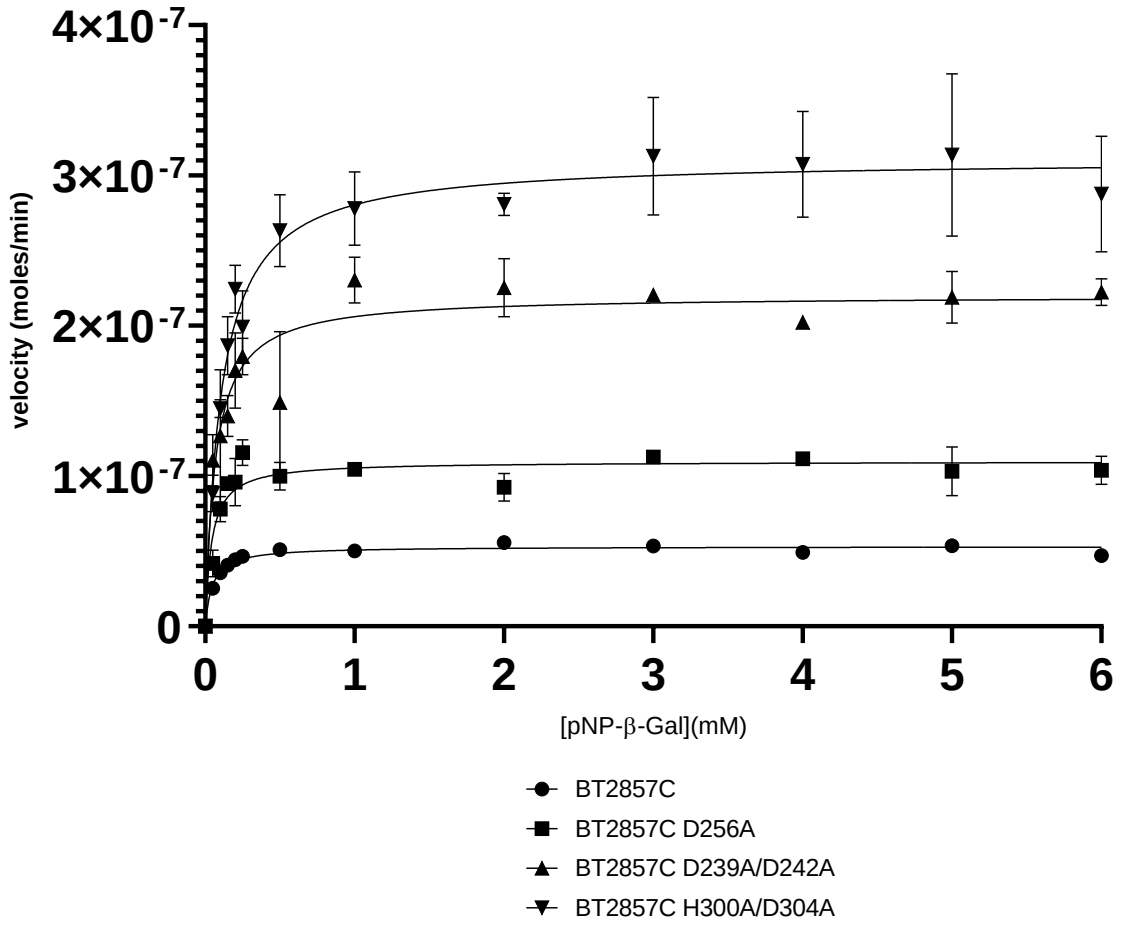


Figure 27. Michaelis-Menten Analysis of BT2857C Mutants. Reactions totalled 250µL with each protein being present at a final concentration of 5µM and were conducted in triplicate. Solutions were buffered with 100mM phosphate at pH 8.4 and a temperature of 22°C. All proteins had a final concentration of 5µM. Un-mutated BT2857C included to illustrate increased catalytic rates of mutants. BT2857C^{D256A} mutant showed values of 1.1x10^{-7}moles/min (±2.94x10^{-9}), 0.041mM (±0.008), and 0.022s^{-1} for V_{max}, K_M, and k_{cat}, respectively. BT2857C$^{D239A/D242A}$ mutant showed values of 2.2x10^{-7}moles/min (±6.33x10^{-9}), 0.068mM (±0.012), and 0.044s^{-1}; while the BT2857C$^{H300A/D304A}$ mutants showed values of 3.11x10^{-7}moles/min (±6.96x10^{-9}), 0.108mM (±0.012), and 0.062s^{-1} for V_{max}, K_M, and k_{cat}.

BT2109 was also tested against the substrate analogue 2-nitrophenyl-β-D-galactopyranoside, though Tuner-based expression of the protein failed to yield sufficiently pure protein to conduct confirmation assays. Preliminary analysis however showed activity against this substrate analog however, seen in Figure 28.

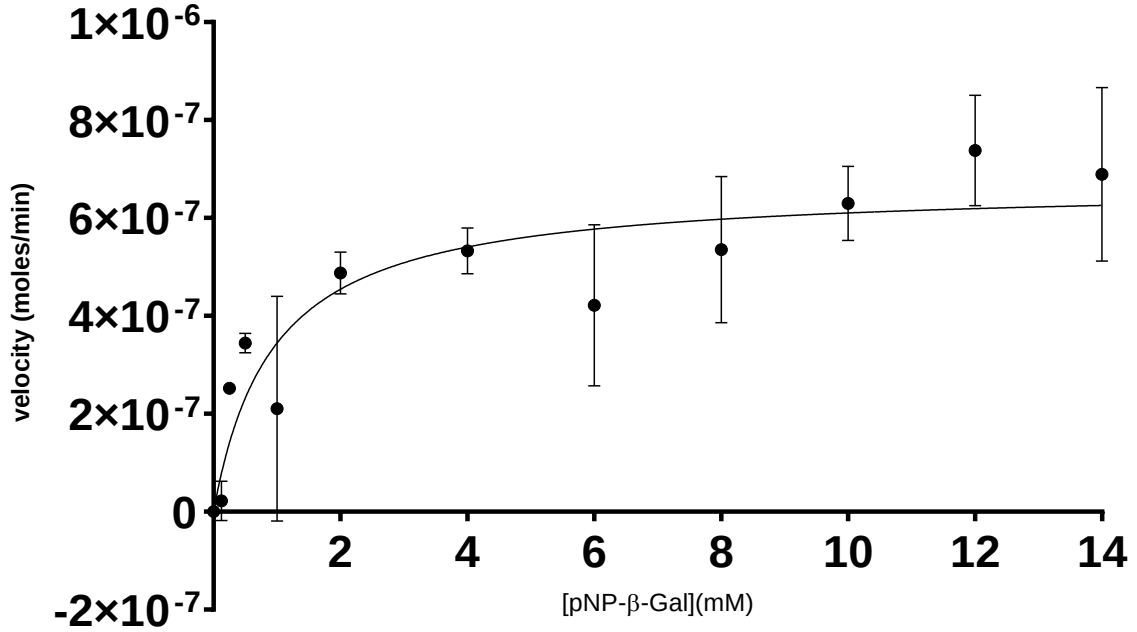


Figure 28. Michaelis-Menten Analysis of BT2109. Reactions totalled 250μL with each protein being present at a final concentration of 1.3μM and conducted in triplicate. Solutions were buffered with 100mM phosphate at pH 8.0 and a temperature of 22°C. BT2109 was determined to have a V_{max}, K_M, and k_{cat} of $6.67x10^{-7}$moles/min ($\pm 1.29x10^{-7}$), 0.94mM (±0.51), and $0.51s^{-1}$, respectively.

Following testing of the enzymes against analog nitrophenyl-carbohydrate substrates the full-length enzymes BT2857 and BT3158 were tested against a variety of natural substrates containing galactose via incubation at optimal buffer/pH conditions, with degradation/modification assessed via TLC. Of the various carbohydrates tested, BT2857 showed degradative capability against α-lactose, with the migration of the upper spot corresponding to galactose, while the second spot migrated a shorter distance suggesting the other product is less polar than either glucose or galactose. Analysis of digestions conducted with C- and N-terminal truncations showed that the N-terminal DUF4959/DUF5126-containing domain was also capable of degrading α-lactose, while the C-terminal DUF5000-containing domain was not. Digestions conducted with BT3158 against α-lactose showed similar results to those of BT2857, suggesting the enzymes work through

the same catalytic mechanism. TLC results for the various digested products can be seen in Figures 29 - 30.

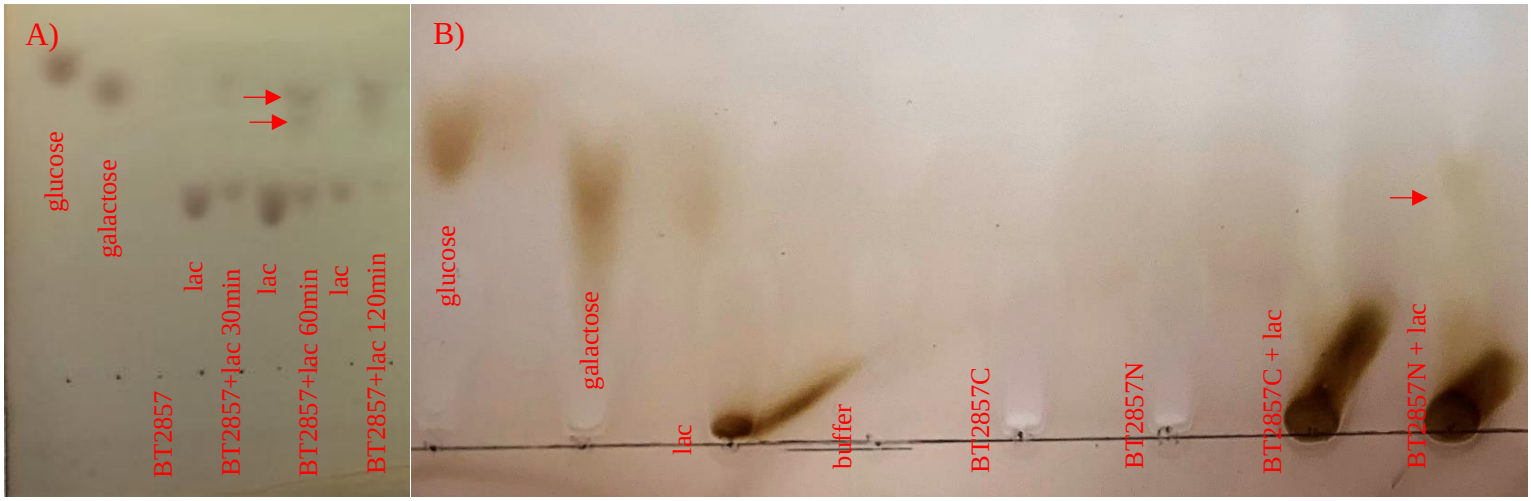


Figure 29. Thin-Layer Chromatography Analysis of BT2857, BT2857N, BT2857C Incubated with α-lactose. Incubations were conducted in phosphate buffer at pH 8.4, 37°C, under agitation. A) shows BT2857 incubated with lactose with reactions arrested at 30, 60, and 120min. B) shows BT2857 truncations incubated overnight.

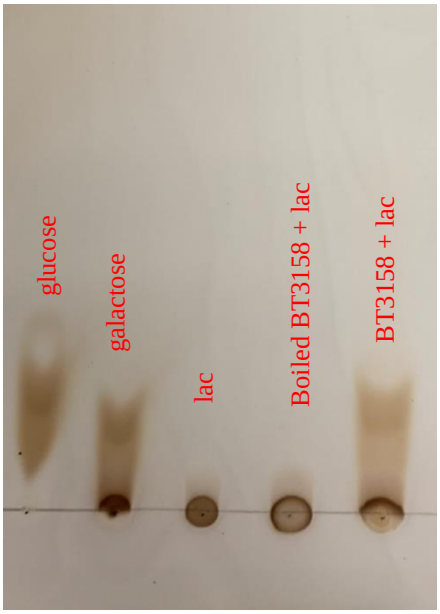


Figure 30. Thin-Layer Chromatography Analysis of BT3158 Incubated with α-lactose. Incubations were conducted in phosphate buffer at pH 8.0, 37°C, under agitation for 18 hours.

Further attempts at characterizing the products of BT2857/carbohydrate digestions were pursued in order to attempt to elucidate the identities of the spots seen on the TLC plates. Products were extracted from the TLC plates through scraping of the silica and immersion in methanol to desorb the carbohydrates from the silica. Silica was removed via centrifugation and methanol was evaporated from the samples before dilution of the samples in D_2O. NMR was then conducted on samples. The results of NMR analysis can be seen in Figure 31 - 35.

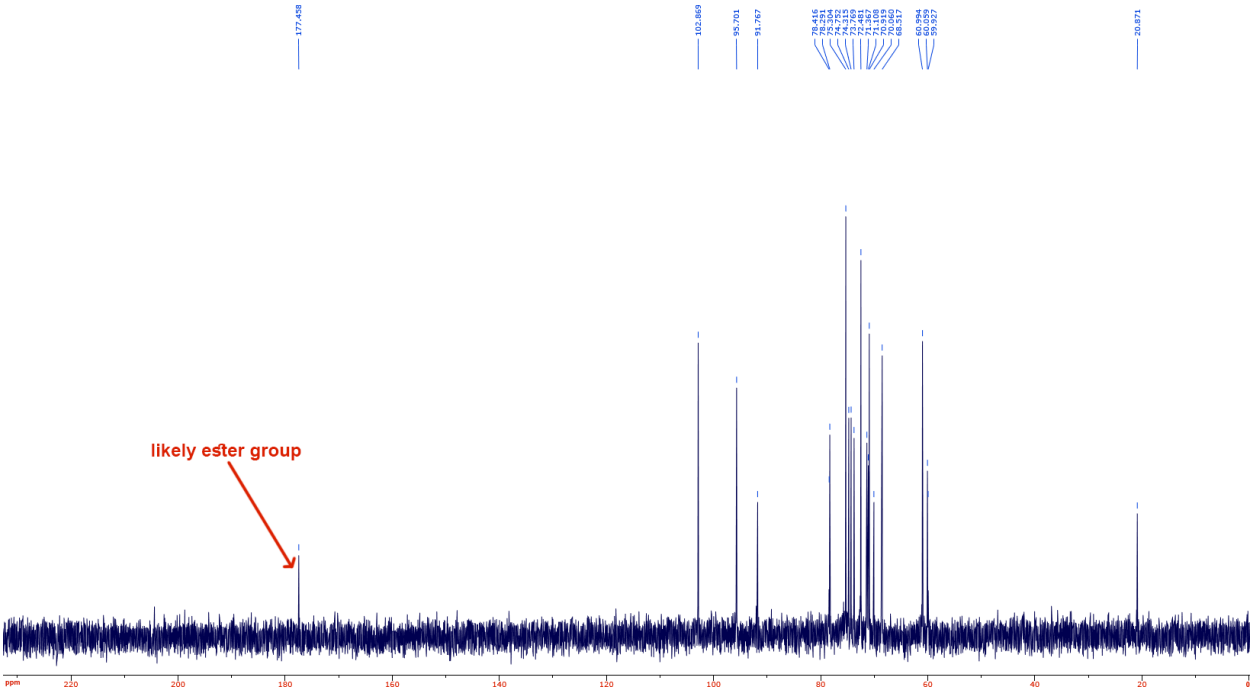


Figure 31. ^{13}CNMR of Lower Scrape from Lactose Digestion TLC. Spectra showed a total of 20 signals, too great a number to come from a single monosaccharide. The presence of a signal within the ester range of 160 – 185 ppm suggested the sample may be contaminated with undigested lactose, or that the protein modifies lactose without breaking the ester bond.

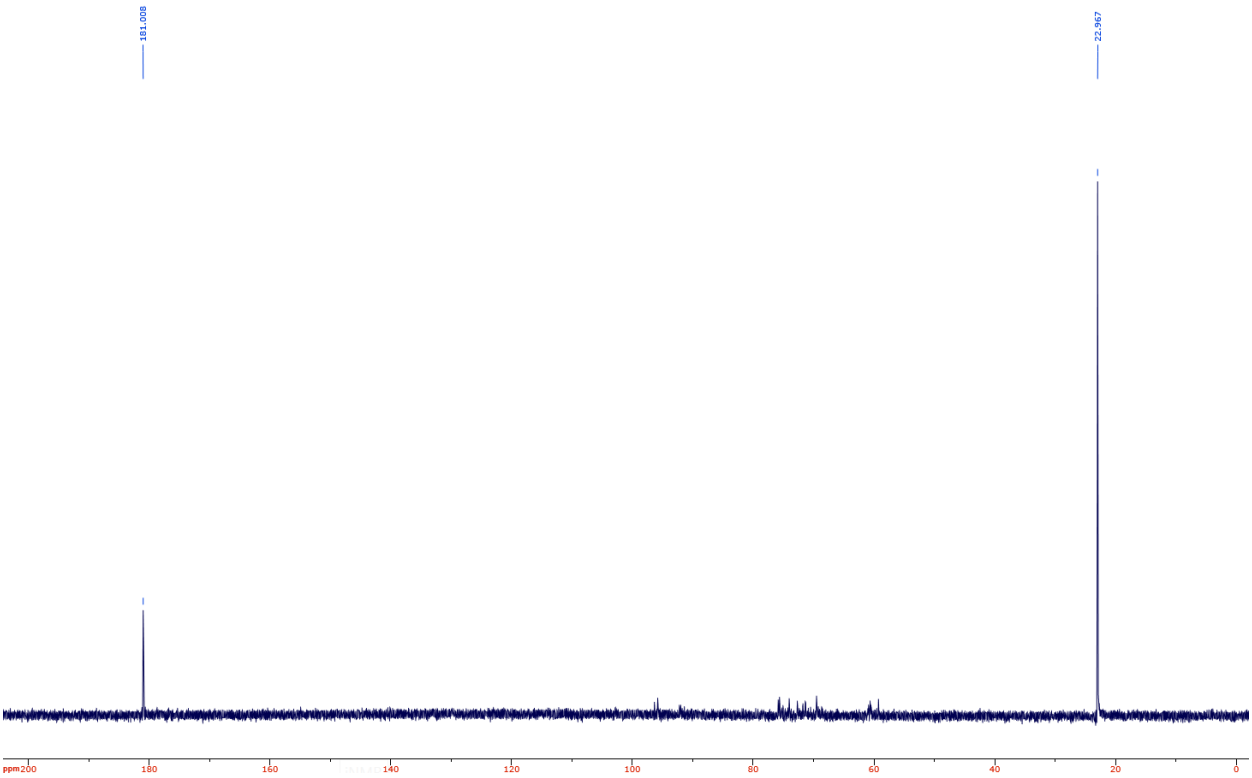


Figure 32. ^{13}C NMR of Upper Scrape from Lactose Digestion TLC. Spectra showed far smaller signals between 55 and 100ppm. Noteworthy is that certain signals appeared upshifted relative to the other isolated product in Figure 27, including the possible ester signal.

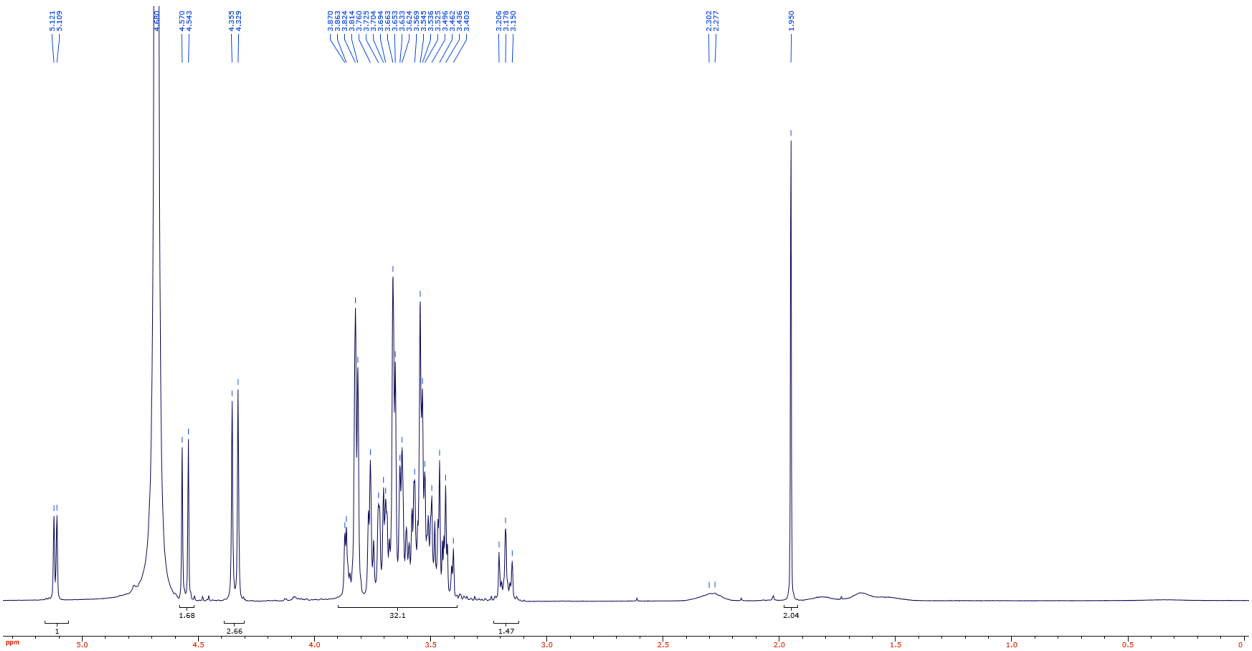


Figure 33. 1D ^{1}H NMR of Lower Scrape from Lactose Digestion TLC. Mass of signals integrated as 32.1 were not able to be correlated to structural features and may require higher-field NMR to resolve. The most down-field signal appears to indicate a CH_2 group in the product.

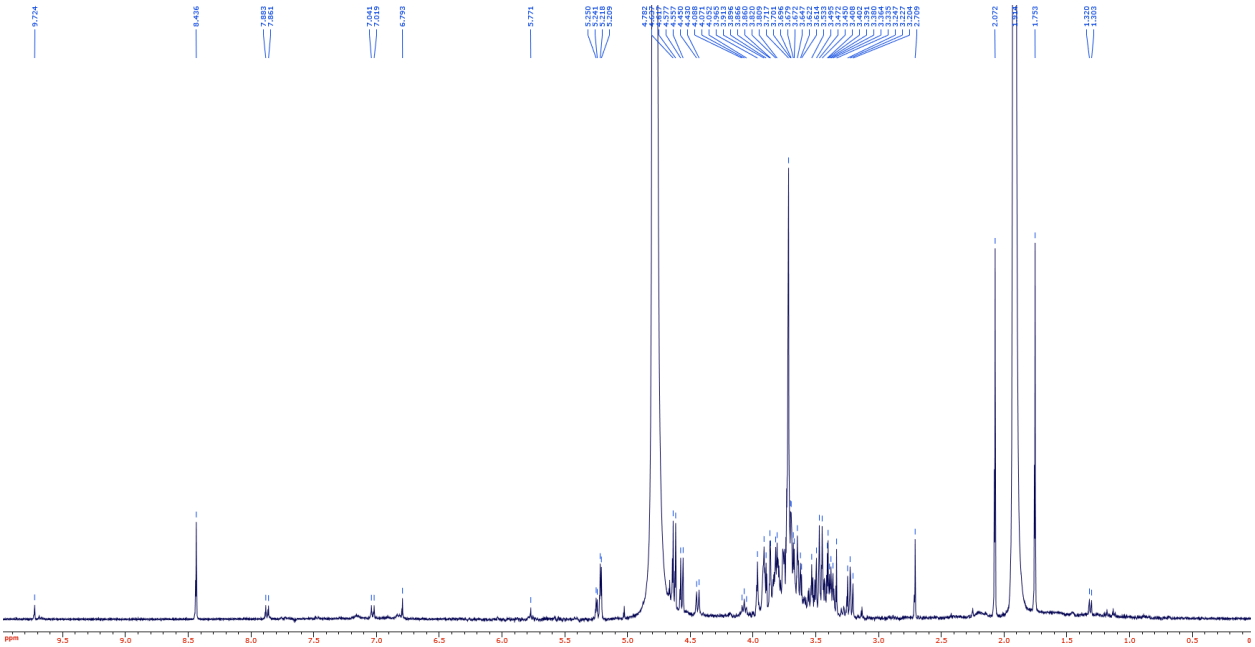


Figure 34. ^{1}H NMR of Upper Scrape from Lactose Digestion TLC. Of note are the signals much further up-field not seen in the product isolated in the lower scrape, and the variation in splitting in the 3 – 4ppm signals.

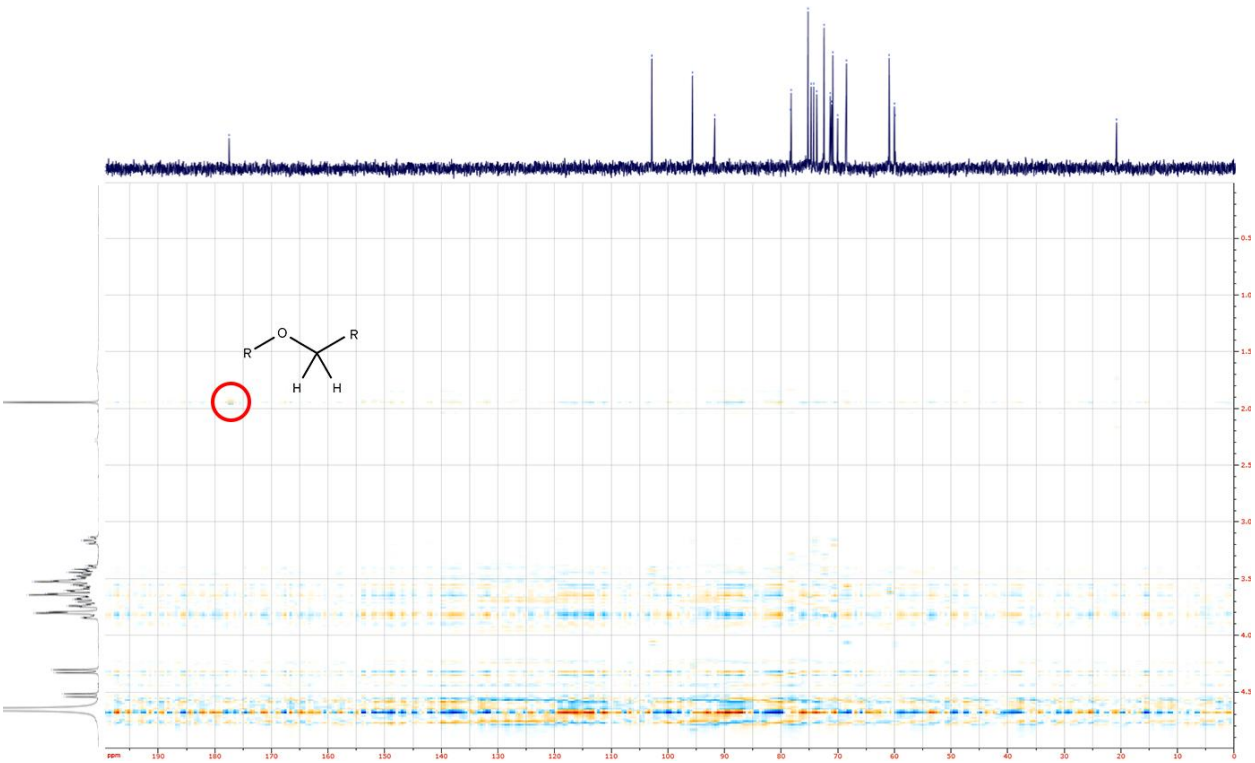


Figure 35. HMBC NMR Spectra of Lower Scrape from Lactose Digestion TLC. Analysis most notably seems to suggest the presence of a methanediyl group next to an ester bridge in the product. This

4.4 Crystallization and Diffraction of Selected Targets

Preliminary crystallization of the BT2857C construct was achieved by a previous MSc student (Maria Ngo) and conditions were expanded on in hanging drop expansions plates. Crystal morphology was varied between thin platelets, platelet clusters, and thick plates, with the thick plate crystals having enough 3-dimensional character to be appropriate for diffraction. Protein purified from cells expressed in selenomethionine-containing minimal media were crystallized in identical conditions to those of native crystals yielding 1.4Å resolution (Figure 36).

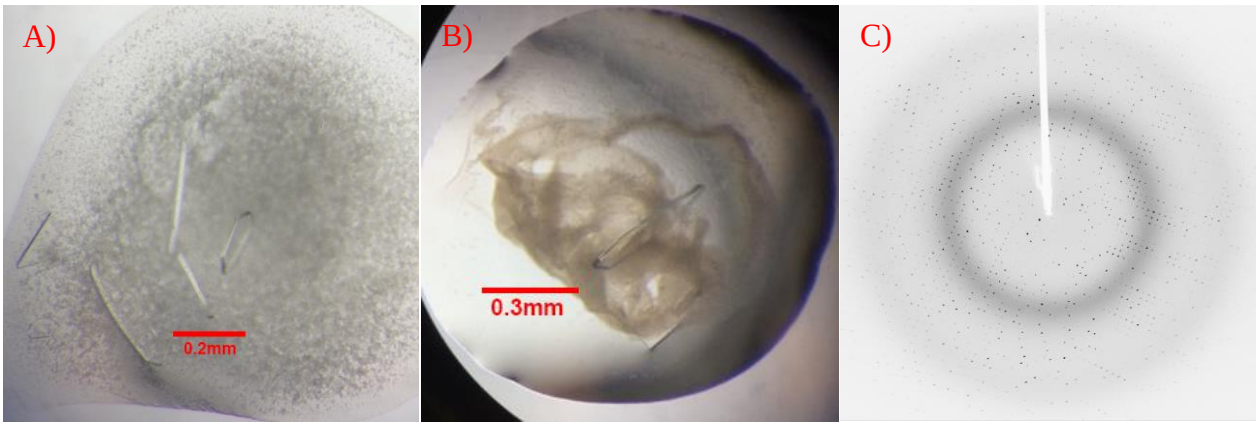


Figure 36. Optimized Crystals and Example Diffraction Pattern for BT2857C. Optimal reservoir conditions included: 0.1M sodium acetate buffer at pH 5.0; 6% (w/v) PEG 3.35K; 2% (w/v) glycerol; 0.2M ammonium acetate. Drops contained 1.5µL 25mg/mL protein and 1.5µL reservoir solution. A) shows crystal generated with protein derived from cells grown in LB culture, while B) shows selenomethionine-containing crystals. C) shows an example pattern taken from native-crystal diffraction to approximately 1.4Å resolution.

Preliminary crystallization was achieved for BT2857, BT2966, BT3158, BT2857C^{D256A}, and BT2857C$^{H300A/D304A}$ double mutant in sitting-drop vapour-diffusion plates. The BT2857 crystal morphology was entirely thin needle-like or thin plate-like. Attempts at expanding these conditions in hanging-drop expansion plates only led to thin needles and platelet clusters, with no diffraction-quality crystals formed (Figure 37).

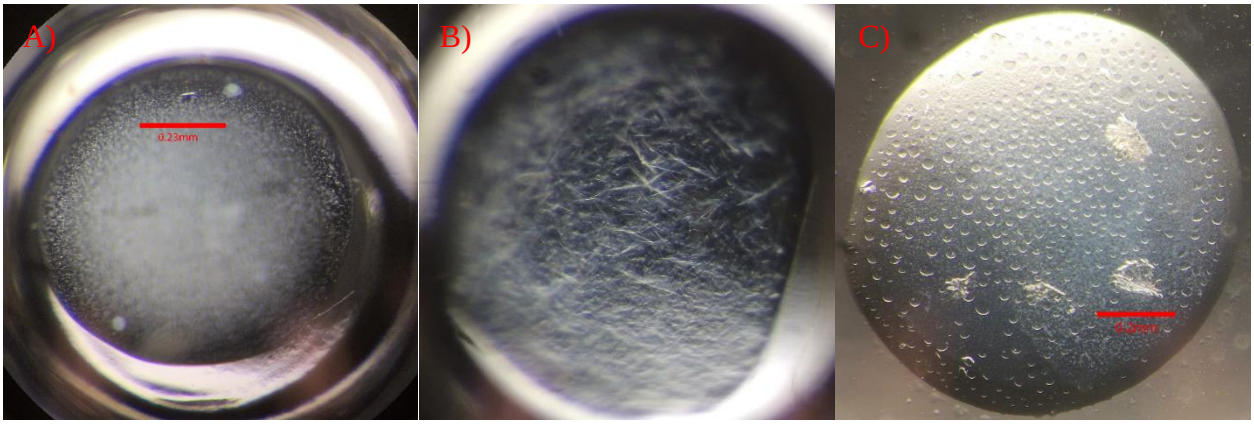


Figure 37. Initial Crystallization Hits and Hanging-Drop Optimization Attempt BT2857. Initial hit A) was generated in MCSG-II well A1 with a reservoir mixture of: 0.2M Potassium tartrate and 20% (w/v) PEG3350; with drop containing 0.5µL 30mg/mL protein + 5mM lactose mixed with 0.25µL reservoir solution. Initial hit B) was generated in MCSG-II well A12 with precipitant mixture of: 0.1M sodium citrate:citric acid pH5.5 and 40% (v/v) PEG600. Hanging-drop crystals in C) were generated based on conditions in initial hit A) and were not of diffraction data assessment quality.

BT2966 formed both small single plate and rod crystals in one sitting-well drop. These crystals failed to be reliably replicated in hanging-drop expansion plates, and crystals taken from the sitting-drop plates showed low resolution during diffraction and as such could not produce an acceptable electron density map to build a model (Figure 38).

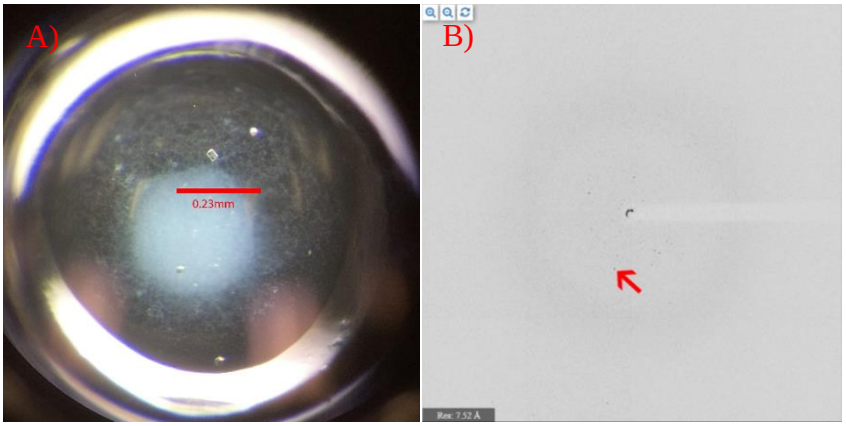


Figure 38. Initial Crystallization Hit for BT2966 and Corresponding Diffraction Pattern. A) Crystals formed in a single sitting drop condition (Molecular Dimensions PACT premier HT-96 well A6) with precipitant mixture containing: 0.1M SPG buffer at pH9.0; 25% w/v PEG1500. Drop consisted of 0.3µL protein at 20mg/mL and 0.2µL reservoir solution. Red arrow in B) indicates the diffraction spot that corresponds to the 7.52Å measurement in the lower left-hand corner of the image.

BT3158 formed crystals in several sitting-drop conditions, with low precipitant and good 3-dimensional character. These conditions were expanded successfully in hanging-drop expansion plates. Co-crystallization with carbohydrates that may bind the enzymes active site was also attempted in hanging-drop plates (Figure 39). Preliminary statistics for the diffraction maps generated from these crystals can be found in Appendix 7.

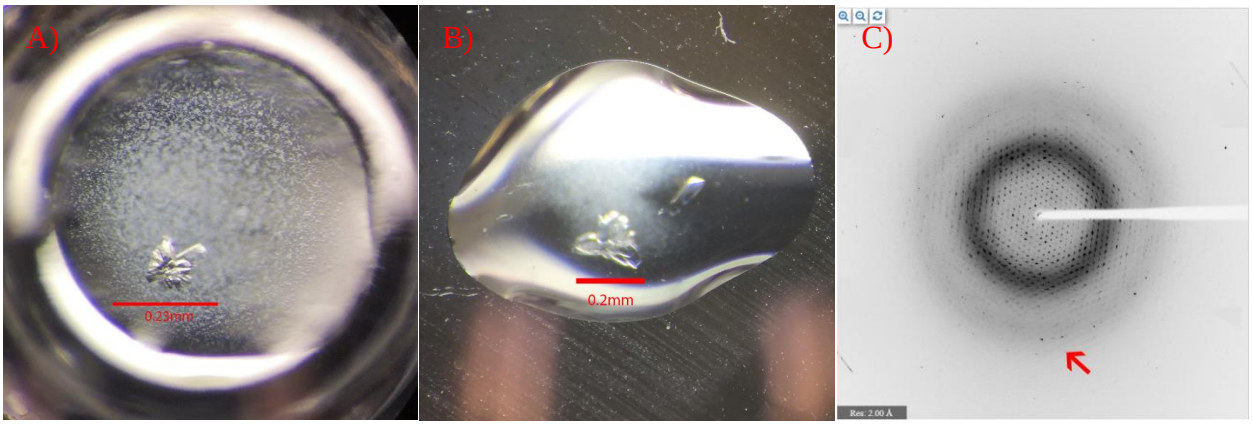


Figure 39. Initial Crystallization and Examples of Optimized Crystallization and Diffraction of BT3158. An initial crystal hit of favourable morphology was found in MCSG-I well C7 with a reservoir condition of: 0.2M $CaCl_2$; 0.1M Tris:HCl buffer pH 8.5; 25% w/v PEG4000; drops contained 0.5µL 20mg/mL protein + 5mM DTT, 0.25µL reservoir solution. A) depicts the initial sitting-drop; B) an expanded hanging-drop crystal of BT3158; C) an example diffraction pattern, with red arrow indicating dot corresponding to 2.00Å.

$BT2857C^{H300A/D304A}$ double mutant formed a single crystal in preliminary sitting-drop screening, alongside a significant amount of precipitate (Figure 40). The protein failed to be expanded on in hanging-drop plates, producing significant amounts of precipitate.

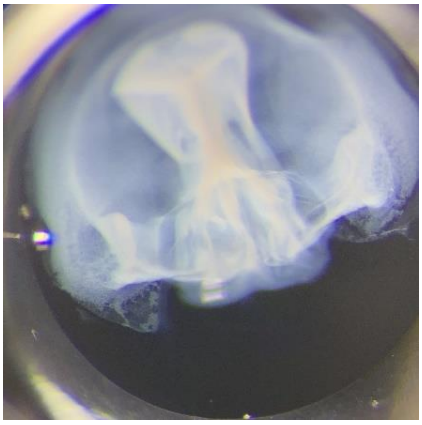


Figure 40. Crystallization of BT2857C H300A/D304A Mutant in Sitting Drop. Crystal was formed in MCSG-I sitting-drop condition A12 with reservoir condition of: 0.2M $CaCl_2$; 0.1M Tris:HCl pH 8.5; 20% w/v PEG4000; drops contained 0.3µL protein at 15mg/mL and 0.3µL reservoir solution.

Preliminary crystallization and hanging-drop expansion of PGN1176 was achieved in number of different sitting-drop conditions, and these conditions were successfully expanded in hanging-drop plates. Examples of the best crystal morphologies and preliminary diffraction patterns can be seen in Figure 41. Further crystallization attempts under conditions that generated favourable crystal morphology were pursued, however maximum resolution yielded by these crystals was limited to approximately 7.11Å.

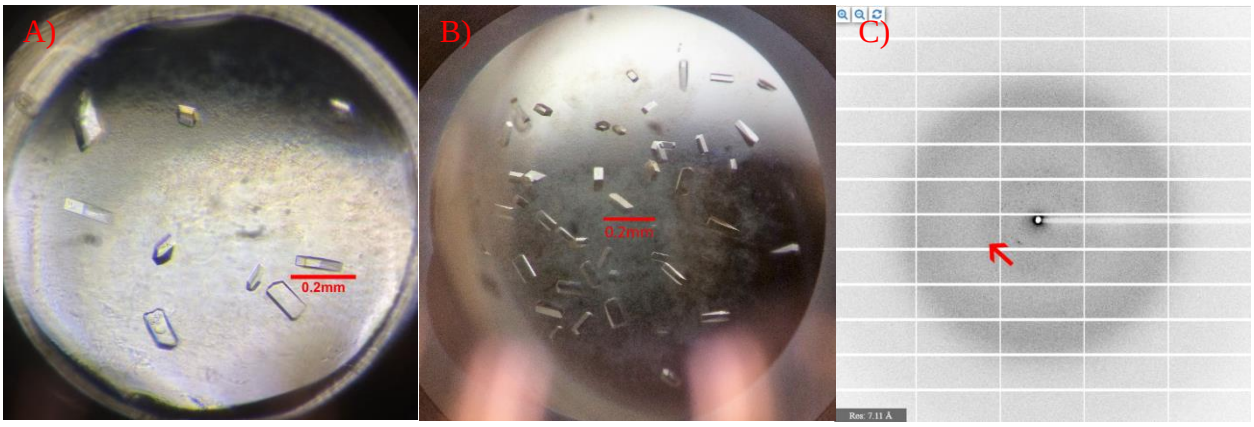


Figure 41. Crystallization of PGN1176 and Sample Diffraction Pattern. Crystals in A) were formed in MCSG-I sitting-drop condition B10 which contained: 0.16M $MgCl_2$; 0.08M Tris:HCl pH 8.5; 24% (w/v) PEG4000; 20% (v/v) glycerol; drops contained 0.5µL protein solution at 20mg/mL and 0.25µL reservoir solution. Crystals in B) were generated in hanging drops based on conditions in A). C) depicts a sample diffraction pattern for a crystal from drop in B), with red arrow pointed to spot corresponding to 7.11Å resolution.

Preliminary attempts at crystallization of PGN1459 in sitting-drop plates yielded several crystals of good morphology in conditions containing high concentrations of sodium citrate, however the diffraction patterns for them made it apparent the crystals were salt rather than protein. One other crystallization condition containing 20% (w/v) PEG3350 as the main precipitating agent and a low concentration of sodium tartrate was subsequently optimized upon in expansion plates. Crystals and a preliminary diffraction pattern can be seen in Figure 42.

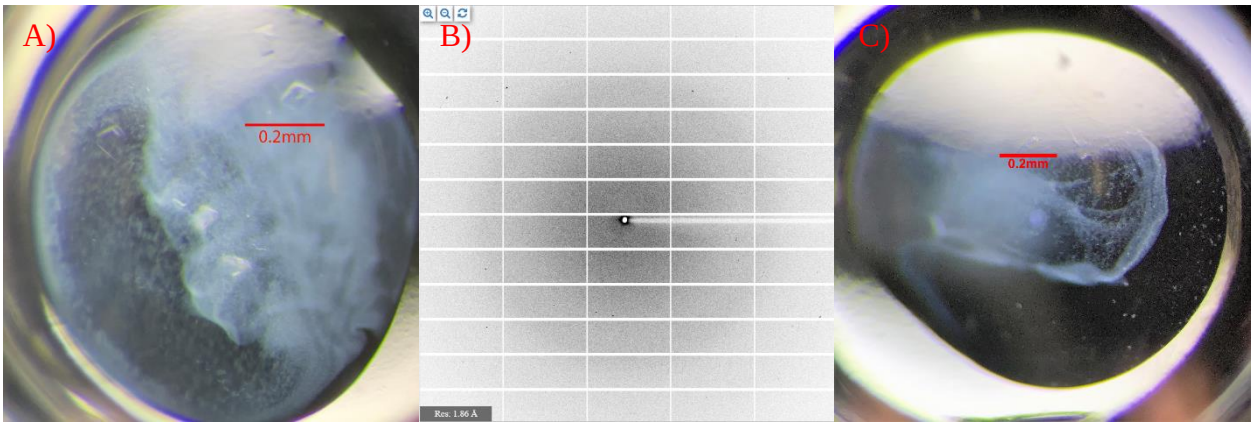


Figure 42. Crystallization of PGN1459 and Sample Diffraction Pattern. A) shows crystals generated in MCSG-II well E9 0.1M CHES:NaOH pH 9.5 and 1M sodium citrate, 0.3µL 15mg/mL protein + 0.3µL reservoir solution droplet, and the corresponding diffraction pattern generated from a crystal in this drop. C) shows crystal formed in MCSG-II well F12, 0.3µL 15mg/mL protein + 0.2µL reservoir solution (0.2M sodium tartrate dibasic; 20% (w/v) PEG3350) droplet.

Preliminary crystallization of PGN1461 in sitting-drop plates was achieved by a colleague (Katarina Mandić) and the conditions for subsequent hanging-drop expansion plates that yielded diffraction-quality crystals were co-designed. The preliminary statistics for diffraction of these crystals taken from these hanging-drop conditions can be seen in Appendix 7.

4.5 Mutagenesis of BT2857C and Truncation of DUF4959

Mutagenesis of BT2857C was initially undertaken as it was assumed that the DUF5000 domain would be the catalytically-active portion of the protein. Residues were identified both through multiple sequence-alignment of target proteins in the case of BT2857N and coordination of ethylene glycol molecules in X-ray diffraction structures in the case of BT2857C. These multiple sequence alignments and coordinated residues interacting with ethylene glycols can be seen in Figure 43 - 44.

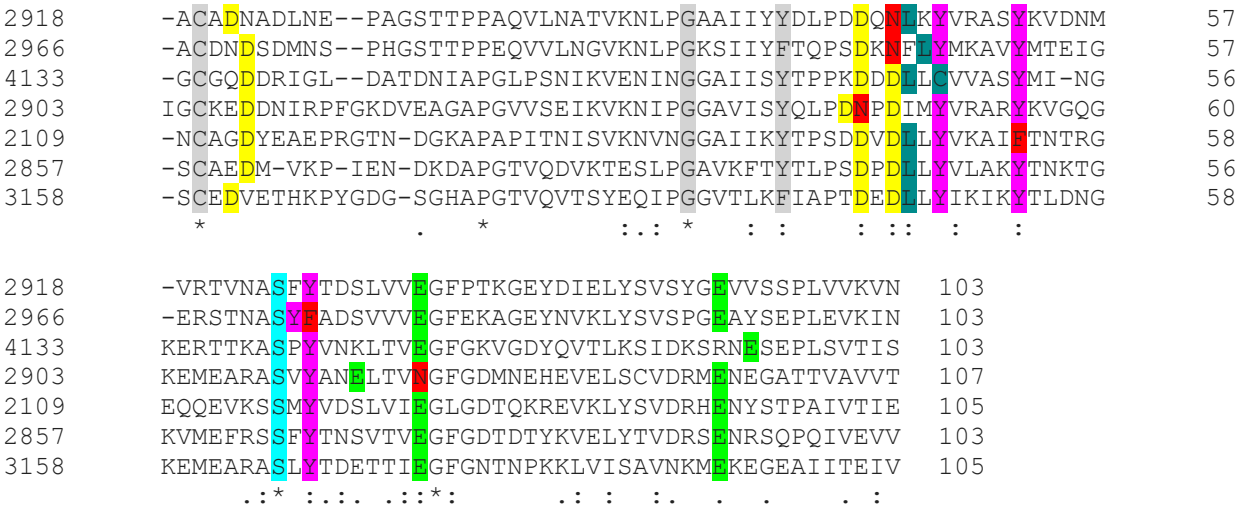


Figure 43. Multiple Sequence Alignment of DUF4959/5126 Domains from Conserved Putative Galactosidases from *B. thetaiotaomicron*. Sequences were aligned using the Clustal Omega online resource. Residues of high conservation and/or interest highlighted for clarity.

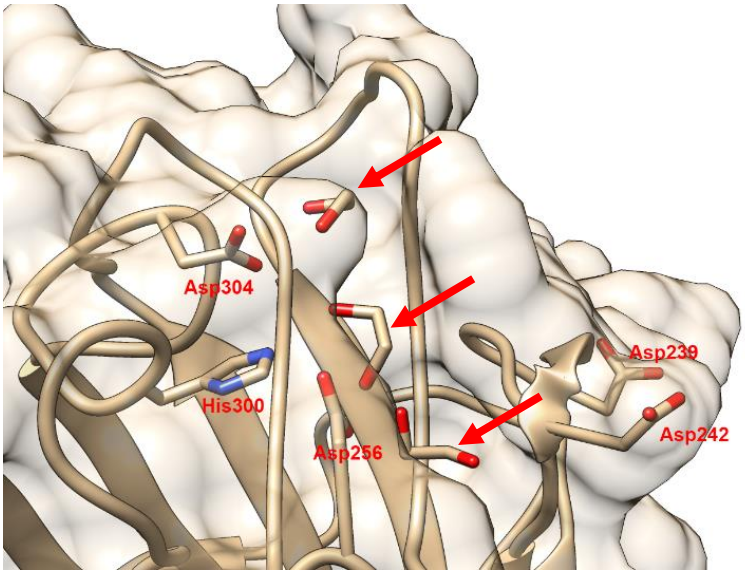


Figure 44. Coordinated Ethylene Glycol Molecules in BT2857C Crystal Structure. Protein structure figure generated in UCSF Chimera and residue labels added in Adobe Photoshop. Ethylene glycol molecules indicated by red arrows.

Residues were selected for mutagenesis in BT2857C based on them both lining the trench motif and being likely charged charged. Generation of the mutants BT2857C$^{D239A/D242A}$, BT2857C^{D256A}, and BT2857C$^{H300A/D304A}$ was done by Darryl Jones at Lethbridge Research and Development Centre, Agriculture and Agri-Food Canada. Mutagenesis of BT2857C to BT2857C^{D239}, BT2857C^{D242A}, BT2857C^{H300A}, BT2857C^{D304A}, and BT2857N to BT2857N^{D35A}, BT2857N^{D37A}, BT2857N^{D83A}, and BT2857N^{R84A} was pursued at Wilfrid Laurier University using Phusion High-Fidelity polymerase with the SOMA-type protocol. Forward and reverse primers used in mutagenic trials can be seen in Appendix 4. Following *Dpn*I digestion of PCR reactions agarose gel analysis was done to assess the success of the trial. Results of these analysis can be seen in Figure 45 and 46.

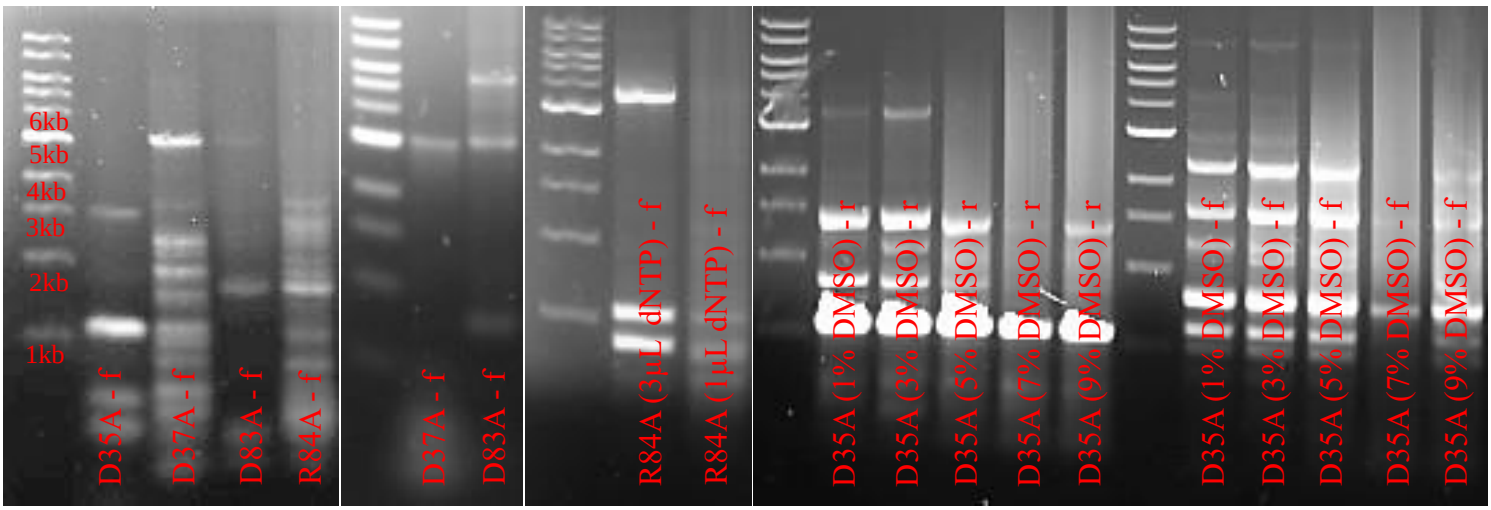


Figure 45. Agarose Gel Analysis of BT2857N SOMA Mutagenesis Post-DpnI Digestion. 0.7% (w/v) gels loaded with 6µL of sample containing 5µL PCR sample and 1µL NEB 6x Purple Gel Loading Dye. Details in parenthesis indicated deviations from standard procedure. "- f" and "- r" suffix indicates whether a forward(f) or reverse(r) primer was used in the reaction. Bands in the 5000 – 5500bp range indicate successful mutagenic replication of the plasmid. Ladder used was NEB Quick-Load Purple 1kb DNA Ladder.

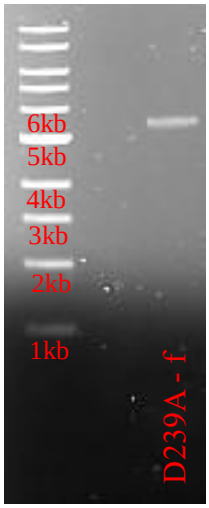


Figure 46. Agarose Gel Analysis of BT2857C SOMA Mutagenesis Post-DpnI Digestion. 0.7% (w/v) gels loaded with 6µL of sample containing 5µL PCR sample and 1µL NEB 6x Purple Gel Loading Dye. Band and the 5500bp range indicate successful mutagenic replication of the plasmid. Ladder used was NEB Quick-Load Purple 1kb DNA Ladder.

The D37A mutant for BT2857N showed single banding and was directly transformed into DH5α *E. coli* cells. BT2857C^{D35A}, BT2857N^{D83A}, and BT2857N^{R84A} mutant plasmids (Figure 44, 45) were extracted and purified from agarose gels with QIAquick gel purification kits prior to transformation. Transformed cells were streak plated on agar containing 100μg/mL ampicillin and colonies were picked to generate 5mL minicultures. Minicultures were miniprepped and plasmids sent to Robarts Research Institute (University of Western Ontario, London, ON, Canada) for bidirectional sequencing. BT2857N^{D37A} and BT2857N^{R84A} and BT2857C^{D239A} mutants were determined to have been successfully generated; amino acid sequence results can be seen in Appendix 5.

Truncation of DUF4959 from BT2857N and BT3158 were pursued, firstly in order to determine if this domain was the site of catalysis for the enzymes, and secondly for attempts at crystallization in the hopes that a full structural characterization could be made for the enzymes. The initial step in the truncation process involved amplifying the gene sequence from the plasmids that contain them. Bestaq 2x Mastermix was used with standard amplification protocol. PCR was performed with a primer that complemented the plasmid upstream of the inserted gene, and a primer that complemented the terminus of the DUF4959 sequence and contained the sequence for the *Nde*I cut site downstream of this terminus. Primers used in the initial amplification of the gene sequence can be seen in Appendix 4. Initial amplified targets can be seen in Figure 47. Successfully amplified targets were purified using QIAquick PCR Purification kit before digestion.

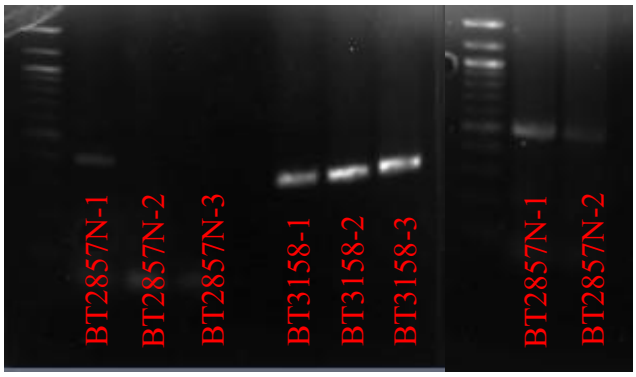


Figure 47. Agarose Gel Analysis of BT2857N/BT3158-DUF4959 Amplification. 0.7% (w/v) gels loaded with 6µL of sample containing 5µL PCR sample and 1µL NEB 6x Purple Gel Loading Dye. Ladder used was NEB Quick-Load Purple 100bp DNA Ladder.

Purified targets were digested using *Xho*I and *Nde*I, as was pET21b(+) with the addition of *Bam*HI, and purified using QIAquick PCR purification kit. Targeted insertion sites *Xho*I and *Nde*I on pET21b(+) can be seen in Appendix 6. Digested and purified products can be seen in Figure 48. pET21b(+) showed an apparent increase in base-pair size, which is like attributable to the linearization of the plasmid(124).

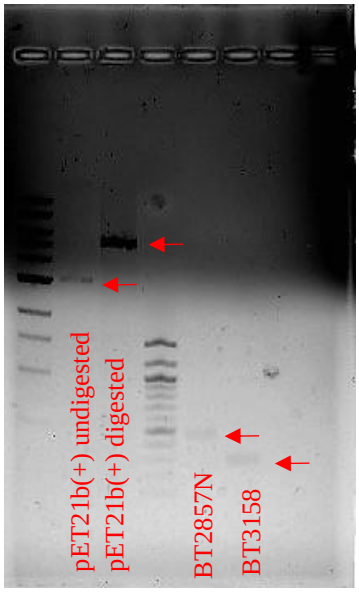


Figure 48. Agarose Gel Analysis of pET21b(+) and BT2857N/BT3158-DUF4959 Digestions. 0.7% (w/v) gels loaded with 6µL of sample containing 5µL PCR sample and 1µL NEB 6x Purple Gel Loading Dye. Ladders used were NEB Quick-Load Purple 100bp DNA Ladder and NEB Quick-Load Purple 1kb DNA Ladder.

BT3158-DUF4959, showing predicted decrease in apparent base-pair size after digestion, was then used in ligation with cut pET21b(+). Ligations were done at 1:3 and 1:5 vector:insert molar ratios. Transformation of ligation reactions into competent DH5α *E. coli* cells yielded 3 colonies on ampicillin-containing agar plates, but none of the colonies were able to grow in ampicillin-containing minicultures, indicating the ligations likely failed.

4.6 Preliminary Functional Analysis of PGN1459

As bioinformatics analysis and literature had indicated that PGN1459 was likely to interact with the genomic DNA of *P. gingivalis*, the protein was tested against several plasmid samples in order to assess this possibility. 50μL of DNA/buffer solution were initially made containing: 20ng of DNA, NEB CutSmart buffer, alkaline phosphatase buffer, and T4 Ligase buffer. 50μL samples were split into two 25μL samples before 1μL of 17.34mg/mL PGN1459 was added to one of the samples. Reaction were incubated for 18 hours, under agitation, at 16°C. Agarose analysis of the results of this can be seen in Figure 49. pET21b(+)-BT2857N, pET21b(+)-3158, and pET28a-BTLip lanes did not show banding for both the control and +protein samples and thus were not considered in analysis. Lanes that contained pET21b(+), pET15-BT2857, and pET21b(+)-BT2857C samples all showed disappearence in banding in wells that contained DNA+PGN1459 protein, suggesting some type of interaction between protein and DNA is occuring preventing the DNA from migrating on the gel. As the protein has a theoretical pI of 5.03 it ought to be negatively charged at the agarose gel pH of 8.3, and thus if it is bound to the DNA it should not cause migration towards the negative electrode. Further analysis will be needed to confirm the mode of interaction between protein and DNA.

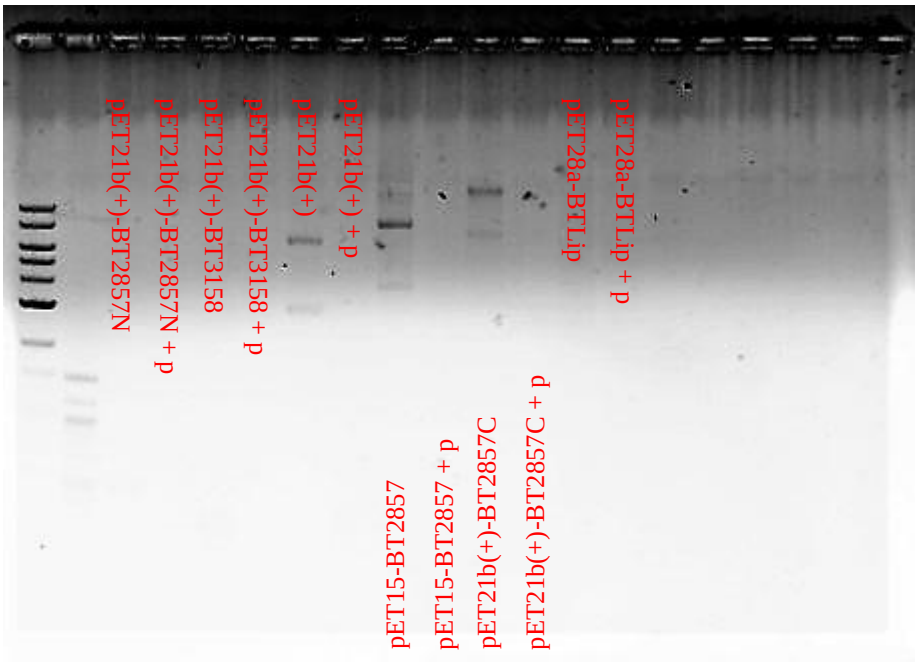


Figure 49. Agarose Gel DNA-Shift Analysis of PGN1459 and Selected Plasmids. 0.7% gels loaded with 12µL of sample containing 10µL PCR sample and 2µL NEB 6x Purple Gel Loading Dye. Ladders used were NEB Quick-Load Purple 100bp DNA Ladder and NEB Quick-Load Purple 1kb DNA Ladder. Labels containing the "+ p" designation indicate which lanes were loaded with samples that contained protein, while labels without "+ p" indicate protein-free controls.

CHAPTER 5 - DISCUSSION

5.1 Structural and Functional Characterization of BT2857 and BT3158

The first objective of this research was the structural and functional characterization of the homologous, putative carbohydrate-active enzymes from the symbiotic gut microbe *B. thetaiotaomicron* in order to assign function to the Domains of Unknown Function (DUF) that make up these proteins, determine active-site residues and add to both the understanding of the pathways that regulate carbohydrate metabolism in BT and general knowledge of protein structure and function. The specific targets chosen for study in this case were: BT2109, BT2857, BT2918, BT2966 and BT3158, all of which are conserved in position on various polysaccharide-utilization loci within the genome of *B. thetaiotaomicron*. Full-length constructs BT2109, BT2857, BT2966, and BT3158 were successfully expressed, purified, and concentrated for kinetic characterization with the substrate mimicking compounds 4-nitrophenyl-β-D-galactopyranoside and 4-nitrophenyl-α-D-galactopyranoside. From the generated Michaelis-Menten curve the V_{max} values for BT2857, BT2857N, and BT2857C were determined to be $2.32x10^{-7}$moles/min, $2.29x10^{-7}$moles/min, and $5.31x10^{-8}$moles/min, respectively at pH 8.4 and 22°C. k_{cat} for each construct was determined to be $0.046s^{-1}$ for both BT2857 and BT2857N, while BT2857C had a value less than ¼ the size, at $0.01s^{-1}$. K_M values for BT2857 and N/C-terminal truncations were determined to be 0.13mM, 0.07mM, and 0.047mM, respectively. Of note were the similar catalytic rates possessed by full-lenth BT2857 and its N-terminal truncation, suggesting that the C-terminal domain does not enhance enzyme activity against the pNP-β-Gal substrate analog. BT3158 generated much higher kinetic values from Michaelis-Menten analysis: $6.06x10^{-6}$moles/min, 0.052mM, and $1.212s^{-1}$ for V_{max}, K_M, and k_{cat} respectively at pH 8.0 and 22°C. Tuner cell-based expression of the constructs confirmed that the constructs possessed β-galactosidase activity as the particular cell line lacks any

genes coding for endogenous β-galactosidases that may cause background hydrolysis of the pNP-β-Gal substrate analog. Testing of the enzymes BT2857, BT2857 truncations, and BT3158 against the substrate analog 4-nitrophenyl-α-D-galactopyranoside did not show any increase in absorbance above background, demonstrating that the enzymes are selective for carbohydrates containing galactose in β-conformation. BT2109 and showed activity against the substrate pNP-β-D-Gal in preliminary testing, though Tuner-cell expressed-protein controls were not able to be run and as such catalysis from the endogenous β-galactosidase coded for by *E. coli* may have been present(125). V_{max}, K_M, and k_{cat} were determined to be $6.67x10^{-7}$moles/min, 0.94mM, and $0.513s^{-1}$ for this protein. Most striking from the various kinetic analysis was the apparent stimulation of catalytic activity in mutants of the BT2857 C-terminal truncation, with the double mutants D239A/D242A and H300A/D304A showing catalytic rates comparable to that of the full-length enzyme. While this domain of the enzyme was hypothesized to be a carbohydrate-binding module based on the initial comparison of construct catalytic rates, these mutants seem to suggest that the DUF5000 may have some suppressed catalytic activity against (R)-β-D-galactose-containing substrates. TLC analysis showed that both BT2857 and BT3158 have activity against the disaccharide α-lactose, which contains a galactose β-bonded to a glucose monomer. Analysis of BT2857 truncations showed that the N-terminal DUF4959/5126 domains had degradative activity against α-lactose, while the DUF5000-containing C-terminal truncation showed no degradative capability. Characterization of the products via NMR analysis revealed a likely ketose group in the product that migrated less quickly on the TLC plate. Contamination with TLC solvent was ruled out due to 1-butanol and acetic acid not containing ketose groups.

Initial crystallization of several BT proteins was achieved in sitting-drop trays although few conditions yielded diffraction-quality crystals in hanging-drop plates. High-quality crystals of

BT2857C and BT3158 were generated however, and high-resolution diffraction data for these proteins was collected.

Modification of the nucleotide sequences of genes coding for BT2857 and BT3158 through truncation and site-directed mutagenesis was pursued in order to elucidate active-site residues, confirm functionality of DUF4959, and build a full crystal structure of at least one of the proteins. Multiple sequence alignment of the DUF4959/5126 sequences from the homologous targets revealed several conserved charged residues, which were decided on as targets for mutagenesis. The amino acid residues in the putative substrate-binding trench of BT2857C were also selected as potential targets. Successful mutagenesis of aspartate 37, arginine 84, and aspartate 239 were achieved through single-oligonucleotide mutagenesis and validated at the Robarts Research Institute. Full truncation and expression of the DUF4959 domain was not fully successful, however it was shown that the amplification, restriction endonuclease digestion, and purification of the DUF4959 gene fragment was achieved.

5.2 Bioinformatics Analysis and Characterization of *P. gingivalis* Proteins

The second objective of this research was to characterize proteins from the oral pathogen *P. gingivalis* that were found to be upregulated during biofilm colonization in order to build a better model of periodontal disease, identify possible novel therapeutic targets to mitigate the virulence of *P. gingivalis*, and add to the collective knowledge on protein structure/function.

5.2.1 PGN1176

PGN1176 was assessed via bioinformatics analysis to very likely be cytoplasmic-localized 3-hydroxybutyryl-CoA dehydrogenase enzyme necessary for catalyzing the reaction: (S)-3-hydroxybutanoyl-CoA + NADP(+) ↔ 3-acetoacetyl-CoA + NADPH (126). As *P. gingivalis* is a microorganism that prefers to use amino acids as a source of carbon and energy(94), this enzyme

is likely not involved in energy metabolism, but rather in pathways necessary for the formation of short-chain carboxylic acids that will eventually be exported extracellularly(127). The export of these short-chain carboxyl species, particularly propionate and butyrate, is strongly associated with progression and relative severity of periodontal disease(128). Volatile fatty acids such as those produced by the pathways PGN1176 involved in, have been shown to cause suppression of the host immune system by inducing apoptosis in B- and T-cells(129, 130) as well as penetrate epithelial tissues and induce an inflammatory response(131). The FadB region in the protein also suggests that the protein may be involved in stationary-phase survival, as Fad genes are frequently associated with energy scavenging and production under starvation conditions in other organisms(132).

Several protein crystallisable conditions showing appropriate morphology for PGN1176 were identified through sitting-drop vapour-diffusion screening. Certain expanded conditions yielded several crystals that, via visual inspection, appeared to be of diffraction quality, while also not requiring submersion in cryoprotectant solution prior to freezing in liquid nitrogen, though these crystals diffracted at resolutions ~7Å and above.

5.2.2 PGN1459

PGN1459 was chosen as a target for investigation based on its relationship to PGN1461 and literature on the YicC domain it contains. YicC proteins (typically referred to as OrfX in previous literature(133, 134)) have been shown to be part of a complement of genes involved in stress-response essential for survival and normal cell morphology in stationary phase at high temperature (37°C - 45°C)(135). This fact is physiologically important as the human body falls within this temperature range and growth of microorganisms within a biofilm can be comparable to growth during stationary phase conditions in some ways(136). One knockout study on *P.*

gingivalis where Universal Stress Protein A (UspA) had been removed from the genome found that biofilm formation and general survivability under stress was reduced as compared to the parental strain of the organism(137). While the function of both of these proteins is unknown(132), it does illustrate that a stress-response protein can play a significant role in the persistence and pathogenicity of *P. gingivalis* during growth in a biofilm.

In the preliminary functional assay of PGN1459, lanes in the agarose gel loaded with sample that contained both DNA and the purified protein had banding disappear. This likely cannot be explained with the protein binding the DNA and migrating towards the negative electrode and off the gel as, like with DNA, PGN1459 should carry a negative charge at the pH of the TAE buffer (~8.3) due the protein having a theoretical pI of 5.03. The second attempt of this crude shift-type assay omitted the heat-arrest, and so precipitation of the DNA out of the solution along with the protein also does not explain the absence of banding in the protein-containing wells.

5.2.3 PGN1461

PGN1461 was assessed via bioinformatics analysis to be a trimeric gate protein localized to the cytoplasmic membrane, though the specific type of transporter, and thus the (type of) substrate the protein is specific to, could not be determined. Three candidates for the type of transporter were, however, able to be identified through sequence-based homology analysis: FeoB-type Fe^{2+} GTPase transporter, NupC-type concentrative nucleoside permease, and AbgT-type aminobenzoyl-glutamate transporter. The possibility of the protein functioning as an iron transporter is based on InterPro and BLASTp, analysis which indicated the presence of two FeoB-type gate domains in the protein. This is a significant possibility as FeoB transporters have a high affinity for Fe^{2+} (138) and iron uptake has been shown to be a high priority for *P. gingivalis*, which contributes greatly to the survival and pathogenicity of the organism(89, 139, 140). The second

possibility, with PGN1461 working as a concentrative nucleoside permease powered by proton symport(141) is based on the top two results for template possibilities according to Phyre^2 analysis. Literature is less clear regarding the importance of nucleosides in the survival and virulence of *P. gingivalis*, though it has been suggested previously they may be used as: A) precursors in amino acid biosynthesis, or B) carbon sources(95). Under certain conditions, genes involved in the scavenging, biosynthesis, and transmembrane transport of nucleotides are seen to be upregulated as well(86, 142), indicating that nucleotide/nucleoside metabolism likely plays some role in the survival of *P. gingivalis*. Thirdly, the possibility of AbgT-family membership was based on the top RaptorX template result (PDB ID: 4R1I) and #3 and #4 ranked Phyre^2 templates (PDB IDs: 4R1I, 4R0C). Members of this family of transporters have shown multitude of possible functions, including: importing aminobenzoyl-glutamate, exporting aminobenzoate, and operating as an efflux pump for sulfonamide antibiotics(143). If PGN1461 were to have any of these functions persistence and virulence would be affected: para-aminobenzoate and para-aminobenzoyl-glutamate are precursors used in the biosynthesis of essential folate and folate-based vitamins and elevated levels of these compounds have been shown to increase the survivability and colonization of *P. gingivalis* in mouse models(144); the presence of an antibiotic efflux protein increases the survivability of the organism(121).

PGN1461 was found to be more effectively bound by anion-exchange resin than cation-exchange resin during FPLC purification despite the theoretical pI of the protein being 8.97, as determined by the ExPASy ProtParam resource. This suggests the protein has a higher proportion of negatively-charged surface residues accessible by the resin, at operating pH of 7.5, than the theoretical pI would suggest. The apparent negative nature of the protein surface may be indicative

of preference for positively-charged species, as surface charge compensation has been previously be seen to be one of the mechanisms employed by certain metal transporters(145).

Successful crystal screening, optimization, cryoprotection, and diffraction of PGN1461 was conducted, with good preliminary data processing for the crystals (Appendix 7).

CHAPTER 6 - CONCLUSIONS AND FUTURE DIRECTIONS

6.1 *B. thetaiotaomicron* Proteins

Bioinformatics analysis predicted each protein in the selected *B. thetaiotaomicron* PULs would possess three conserved domains of unknown function: 4959, 5126, and 5000. DUF5000 was thought to be the most likely site of catalysis due to its homology to galactose-binding like domains. Through kinetic characterization the target proteins BT2109, BT2857, BT2966, and BT3158 all possessed catalytic activity against the substrate-analog 4-nitrophenyl-β-D-galactopyranoside which partially illuminates the functionality of these proteins. The proteins BT2857 and BT3158 were tested against the substrate 4-nitrophenyl-α-D-galactopyranoside, which showed no activity, confirming stereochemical-specificity of the proteins at least against galactose-containing disaccharides. For BT2857, a V_{max} of $3.19x10^{-7}$M/min, K_M of 1M, and k_{cat} of $0.001s^{-1}$ was calculated through Michaelis-Menten analysis, while BT3158 showed far more effective catalysis with values of $8.67x10^{-6}$M/min, 1M, and $0.029s^{-1}$ for V_{max}, K_M, and k_{cat}, respectively. The C-terminal truncation of BT2857 had values of $7.64x10^{-8}$M/min calculated for V_{max}, 1M for K_M, and $0.00025s^{-1}$ for k_{cat}, while the N-terminal truncation had values of $3.30x10^{-7}$M/min, 1M, and $0.001s^{-1}$. This revealed that contrary to the hypothesis that the C-terminal DUF5000 domain would be primarily responsible for catalysis in the protein, the N-terminal domains DUF4959 and DUF5126 appear to be responsible for breakdown of saccharides containing β-galactose. Mutations of BT2857C showed increased activity against the substrate pNP-β-Gal, possibly suggesting the DUF5000 domain may have some suppressed CAZyme activity. TLC analysis of BT2857 and BT3158 revealed degradative capability against the carbohydrate substrate α-lactose, a disaccharide which contains a galactose in β-conformation. The same analysis of BT2857N and BT2857C revealed that BT2857N is capable of degrading α-

lactose while BT2857C was not, further validating the notion that the DUF4959 and DUF5126 domains are primarily responsible for the breakdown of carbohydrates containing a β-conformed galactose monosaccharide.

Full structural validation of homologous galactosidase structures was not able to be achieved in the course of this research due to the poor electron density seen over the majority of the DUF4959 domain in BT3158 despite the high-resolution data (Appendix 7). Truncation and crystallization of this domain from BT3158 will be required to obtain a fully-validated structure for one of the selected BT proteins. Co-crystallization of the proteins with substrates and/or substrate analogs would also be assistive in the analysis of individual binding sites within crystallized BT proteins. Further TLC functional analysis of the proteins using trisaccharides should be done to further assess the degradative capabilities of the proteins, and Michaelis-Menten analysis of BT2857N point mutants would also identify if the targeted amino acid residues are catalytically important.

6.2 *P. gingivalis* Proteins

Bioinformatic and literature research was able to validate the putatively role of the likely biologically significant factors selected for study, with each protein selected being expected to serve to either increase the pathogenicity or general fitness of *P. gingivalis* in ways that fit within the context of previously characterized virulence pathways. PGN1176 likely plays a role in the production of fatty acids involved in the suppression of host immune response; PGN1461 in the uptake of metabolically necessary ions/compounds or efflux of antibiotics; PGN1459 in maintaining cellular health and morphology under heat and stress. Each protein was successfully expressed and purified to homogeneity for crystal tray plating, and some preliminary functional

analysis was conducted, in order to reveal the structural and functional characteristics of the targets.

Full structural validation of *P. gingivalis* target proteins was not able to be achieved over the course of this research, however significant progress towards this goal was made. PGN1176 was able to be crystallized from both fresh and frozen stocks of protein, and currently has several sitting-drop conditions that may be optimized in hanging drop plates, such that crystals that diffract at high resolution may be generated, including those that are co-crystallized with NAD(P)+ and/or acetyl-CoA. PGN1461 was managed to be crystallized and diffracted at acceptable resolutions with good preliminary statistics. Due to the protein having low similarity with any other previously characterized construct, experimental phasing will need to be pursued in order to solve the structure. For experimental phasing both expression of the protein in selenomethioine-containing media and soaking of native protein crystals with heavy atoms (particularly iron) should be pursued. Co-crystallization of proteins with putative substrates, such as nucleosides and p-aminobenzoate/aminobenzoyl-glutamate should also be pursued. PGN1459 yielded a one crystallization condition that appeared to be in fact sodium citrate crystals, and one condition containing mainly PEG3350 as the precipitating agent. This PEG-containing condition permitted the growth of only a single small crystal and should be further optimized in future attempts to solve the structure of this unique, and poorly studied type of protein. Should optimization of PEG-based crystallization be successful experimental phasing will be required to solve the structure of this protein, as well as co-crystallization of protein with nucleotides/nucleosides.

All of the various *P. gingivalis* targets selected for study in will all require experimental validation of function in order to confirm activity of the proteins and generate any kinetic constants that may be associated with the type of reaction. In the case of PGN1176, function as either a 3-

hydroxybutyryl-CoA dehydrogenase (EC: 1.1.1.157) or 3-hydroxyacyl-CoA dehydrogenase (EC: 1.1.1.35) will need to be determined, along with parameters such as K_M and k_{cat}. For PGN1461, activity as either a Fe^{2+}, nucleoside, or aminobenzoyl-glutamate transporter will need to be determined in order to fit it into an improved model of periodontal disease being built. PGN1459 is a protein of truly unknown function, with little research into its mode of action or mechanism. Based on previous literature and genomic context however, it likely interacts with DNA in some fashion. Confirmation of protein-DNA interaction, and elucidation of the type/mechanism of function will be required to determine what the role of the protein is in the cell.

REFERENCES

1. Herman, M., Marchessault, R., Caruthers, M., Dorn, K., Ringsdorf, H., Fitzer, E., Ward, I., Jerome-Jackson, W., Stille, J., Dhingra, A., Relles, H., Wegner, G., Baughman, R., Economy, J., Eisenberg, A., Besso, E., Harris, F., Gupta, R., Yeager, H., Steck, A., and Pearce, E. (1984) *Contemporary Topics in Polymer Science* (Vandenberg, E. J. ed), Springer US
2. Saliba, R. C., and Pohl, N. L. B. (2016) Designing sugar mimetics: Non-natural pyranosides as innovative chemical tools. *Curr. Opin. Chem. Biol.* **34**, 127–134
3. McNaught, A. D. (1996) Nomenclature of Carbohydrates. *Pure Appl. Chem.* **68**, 1919–2008
4. Colley, K., Varki, A., and Kinoshita, T. (2017) *Cellular Organization of Glycosylation In: Essentials of Glycobiology*, 3rd Ed. (Varki A, Cummings RD, Esko JD, et al. ed), Cold Spring Harbor, New York, 10.1101/glycobiology.3e.004
5. Berg, J., Tymoczko, J., and Stryer, L. (2007) Glycolysis Is an Energy-Conversion Pathway in Many Organisms. in *Biochemistry*, pp. 117–144, W.H. Freeman and Company, 10.1007/978-3-8274-2989-6
6. Kuntz, I. D., Davies, D. R., Dickerson, R. E., Doolittle, R. F., Feldmann, R. J., Hermans, J., Martin, Y. C., Prestegard, J. H., Rossky, P. J., Scheraga, H. A., Smith, D. H., Sweeley, C. C., Tinoco, I., Rosen, W. G., and Walton, S. (1987) *Computer Assisted Modeling: Contributions of Computational Approaches to Elucidating Macromolecular Structure and Function Committee on Computer-Assisted Modeling, National Research Council*
7. Chen, L., Cheung, L. S., Feng, L., Tanner, W., and Frommer, W. B. (2015) Transport of Sugars. *Annu. Rev. Biochem.* 10.1146/annurev-biochem-060614-033904
8. Weaver, R. (2011) *Molecular Biology*, 978-0-07-352532-7
9. Nelson, D. L., and Cox, M. M. (2008) *Lehninger Principles of Biochemistry*, 10.1016/j.jse.2011.03.016
10. Corfield, A. (2017) Eukaryotic protein glycosylation: a primer for histochemists and cell biologists. *Histochem. Cell Biol.* **147**, 119–147
11. Piotrowski, M., Wulta, D., Obuch-woszczaty, P., and Pituch, H. (2019) Fructooligosaccharides and mannose affect Clostridium difficile adhesion and biofilm formation in a concentration-dependent manner. *Eur. J. Clinincal Microbiol. Infect. Dis.*
12. Cummings, R. D. (2019) Stuck on sugars – how carbohydrates regulate cell adhesion, recognition, and signaling. *Glycoconj. J.* **36**, 241–257
13. Beauvais, A., and Latge, J. (2014) Functional duality of the cell wall. *Curr. Opin. Microbiol.* **20**, 111–117
14. Mistou, M., Sutcliffe, I. C., and Sorge, N. M. Van (2016) Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria. *FEMS Microbiol. Rev.* **40**, 464–479
15. Jurtshuk, P. (1996) Bacterial Metabolism. in *Medical Microbiology*, 4th Ed. (Baron, S. ed), University of Texas Medical Branch, Galveston
16. Muñoz-elías, E. J., and Mckinney, J. D. (2006) Carbon metabolism of intracellular bacteria. *Cell. Microbiol.* **8**, 10–22
17. Herget, S., Toukach, P. V., Ranzinger, R., Hull, W. E., Knirel, Y. A., and Von Der Lieth, C. W. (2008) Statistical analysis of the bacterial carbohydrate structure data base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with

mammalian glycans. *BMC Struct. Biol.* 10.1186/1472-6807-8-35
18. Ndeh, D., Rogowski, A., Cartmell, A., Luis, A. S., Baslé, A., Gray, J., Venditto, I., Briggs, J., Zhang, X., Labourel, A., Terrapon, N., Buffetto, F., Nepogodiev, S., Xiao, Y., Field, R. A., Zhu, Y., O'Neill, M. A., Urbanowicz, B. R., York, W. S., Davies, G. J., Abbott, D. W., Ralet, M.-C., Martens, E. C., Henrissat, B., and Gilbert, H. J. (2017) Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*. 10.1038/nature21725
19. André, I., Potocki-Véronèse, G., Barbe, S., Moulis, C., and Remaud-Siméon, M. (2014) CAZyme discovery and design for sweet dreams. *Curr. Opin. Chem. Biol.* **19**, 17–24
20. Cantarel, B. I., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238
21. Davies, G., and Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases. *Structure*. **3**, 853–859
22. Rye, C. S., and Withers, S. G. (2000) Glycosidase mechanisms. *Curr. Opin. Chem. Biol.* **4**, 573–580
23. Sinnott, M. L. (1990) Catalytic Mechanisms of Enzymic Glycosyl Transfer. *Chem. Rev.* **90**, 1171–1202
24. Lairson, L. L., Henrissat, B., Davies, G. J., and Withers, S. G. (2008) Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **77**, 521–555
25. Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* **432**, 437–444
26. Aragunde, H., Biarnés, X., and Planas, A. (2018) Substrate recognition and specificity of chitin deacetylases and related family 4 carbohydrate esterases. *Int. J. Mol. Sci.* **19**, 1–30
27. Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., and Henrissat, B. (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels*. **6**, 1
28. BORASTON, A. B., BOLAM, D. N., GILBERT, H. J., and DAVIES, G. J. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769–781
29. Busk, P. K., and Lange, L. (2013) Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs. *Appl. Environ. Microbiol.* **79**, 3380–3391
30. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 10.1371/journal.pcbi.1000605
31. Jones, D. R., Thomas, D., Alger, N., Ghavidel, A., Douglas Inglis, G., and Wade Abbott, D. (2018) SACCHARIS: An automated pipeline to streamline discovery of carbohydrate active enzyme activities within polyspecific families and de novo sequence datasets. *Biotechnol. Biofuels*. **11**, 1–15
32. McDonald, D., Birmingham, A., and Knight, R. (2015) Context and the human microbiome. *Microbiome*. **3**, 52
33. Lander, E. S. . L., Birren, L. M. ., Nusbaum, B., Zody, C., Baldwin, M. C. ., Devon, J., Dewar, K., Doyle, K., FitzHugh, M., Funke, W., Gage, R., Harris, D., Heaford, K., Howland, A., Kann, J., Lehoczky, L., LeVine, J., McEwan, R., and McKernanKevin., P.

(2001) Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921
34. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007) The Human Microbiome Project. *Nature*. **449**, 804–810
35. Fields, C., Adams, M., White, O., and Venter, J. C. (2013) How Many Genes Are in the Human Genome? *Nat. Genet.* **7**, 345–346
36. Dunham, I. (2000) The Gene Guessing Game. *Yeast*. **1**, 218–224
37. Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. (2018) Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv*. 10.1101/332825
38. Martín, R., Miquel, S., Langella, P., and Bermúdez-Humarán, L. G. (2014) The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*. **5**, 413–23
39. Cho, I., and Blaser, M. (2017) The Human Microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270
40. Funktionen, C., and Repräsentationsfunktion, I. (2003) Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*. **486**, 1049–1058
41. Torres, P. J., Thompson, J., Mclean, J. S., Kelley, S. T., and Edlund, A. (2019) Discovery of a Novel Periodontal Disease-Associated Bacterium. *Microb. Ecol.* **77**, 267–276
42. Sender, R., Fuchs, S., and Milo, R. (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, 1–14
43. Kaoutari, A. El, Armougom, F., Gordon, J. I., Raoult, D., and Henrissat, B. (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.* **11**, 497–504
44. Cantarel, B. L., Lombard, V., and Henrissat, B. (2012) Complex carbohydrate utilization by the healthy human microbiome. *PLoS One*. **7**, 1–10
45. Ravcheev, D. A., Godzik, A., Osterman, A. L., and Rodionov, D. A. (2013) Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics*. **14**, 1–17
46. Wilkens, C., Busk, P. K., Pilgaard, B., Zhang, W. J., Nielsen, K. L., Nielsen, P. H., and Lange, L. (2017) Diversity of microbial carbohydrate-active enzymes in Danish anaerobic digesters fed with wastewater treatment sludge. *Biotechnol. Biofuels*. 10.1186/s13068-017-0840-y
47. Mahowald, M. A., Rey, F. E., Seedorf, H., Turnbaugh, P. J., Fulton, R. S., Wollam, A., Shah, N., Wang, C., Magrini, V., Wilson, R. K., Cantarel, B. L., Coutinho, P. M., Henrissat, B., Crock, L. W., Russell, A., Verberkmoes, N. C., Hettich, R. L., and Gordon, J. I. (2009) Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc. Natl. Acad. Sci.* **106**, 5859–5864
48. Larsson, E., Tremaroli, V., Lee, Y. S., Koren, O., Nookaew, I., Fricker, A., Nielsen, J., Ley, R. E., and Bäckhed, F. (2012) Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut Microbiota*. **61**, 1124–1131
49. Zocco, M. A., Ainora, M. E., Gasbarrini, G., and Gasbarrini, A. (2007) Bacteroides thetaiotaomicron in the gut: Molecular aspects of their interaction. *Dig. Liver Dis.* **39**, 707–712

50. Hooper, L. V., Midtvedt, T., and Gordon, J. I. (2002) How host-microbial relationships shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283–307
51. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005) Host-bacterial mutualism in the human intestine. *Science (80-.).* **307**, 1915–20
52. Hooper, L. V., Stappenbeck, T. S., Hong, C. V., and Gordon, J. I. (2003) Angiogenins: A new class of microbicidal proteins involved in innate immunity. *Nat. Immunol.* **4**, 269–273
53. Hooper, L. V. (2001) Molecular Analysis of Commensal Host-Microbial Relationships in the Intestine. *Science (80-.).* **291**, 881–884
54. Wexler, H. M. (2007) Bacteroides: The good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20**, 593–621
55. Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., and Forano, E. (2012) Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*. **3**, 289–306
56. Terrapon, N., Lombard, V., Gilbert, H. J., and Henrissat, B. (2015) Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*. **31**, 647–655
57. Teravest, M. A., He, Z., Rosenbaum, M. A., Martens, E. C., Cotta, M. A., Gordon, J. I., and Angenent, L. T. (2014) Regulated expression of polysaccharide utilization and capsular biosynthesis loci in biofilm and planktonic Bacteroides thetaiotaomicron during growth in chemostats. *Biotechnol. Bioeng.* **111**, 165–173
58. Terrapon, N., Lombard, V., Drula, É., Lapébie, P., Al-Masaudi, S., Gilbert, H. J., and Henrissat, B. (2018) PULDB: The expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res.* **46**, D677–D683
59. Martens, E. C., Chiang, H. C., and Gordon, J. I. (2009) Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe*. **4**, 447–457
60. Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V, and Gordon, J. I. (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science (80-.).* **299**, 2074–6
61. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30
62. Comstock, L. E., and Coyne, M. J. (2003) Bacteroides thetaiotaomicron: A dynamic, niche-adapted human symbiont. *BioEssays*. **25**, 926–929
63. Kyrpides, N. C. (1999) Genomes Online Database (GOLD 1.0): A monitor of complete and ongoing genome projects world-wide. *Bioinformatics*. **15**, 773–774
64. Niehaus, T. D., Thamm, A. M., de Crécy-Lagard, V., and Hanson, A. D. (2015) Proteins of unknown biochemical function - A persistent problem and a roadmap to help overcome it. *Plant Physiol.* **169**, pp.00959.2015
65. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I. M. A., Kyrpides, N. C., and Reddy, T. B. K. (2019) Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Res.* **47**, D649–D659
66. Schwede, T., Diemand, A., Guex, N., and Peitsch, M. C. (2000) Protein structure computing in the genomic era. *Res. Microbiol.* **151**, 107–112
67. Mills, C. L., Beuning, P. J., and Ondrechen, M. J. (2015) Biochemical functional predictions for protein structures of unknown or uncertain function. *Comput. Struct. Biotechnol. J.* **13**, 182–191
68. Bateman, A., Coggill, P., and Finn, R. D. (2010) DUFs: Families in search of function. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1148–1152

69. Goodacre, N. F., Gerloff, D. L., and Uetz, P. (2014) Protein Domains of Unknown Function Are Essential in Bacteria. *MBio*. **5**, e00744-13
70. Iyer, L. M., Aravind, L., Bork, P., Hofmann, K., Mushegian, A. R., Zhulin, I. B., and Koonin, E. V. (2001) Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol.* **2**, 1–11
71. Roberts, R. J. (2004) Identifying protein function - A call for community action. *PLoS Biol.* **2**, 293–294
72. Kilian, M., Chapple, I. L. C., Hannig, M., Marsh, P. D., Meuric, V., Pedersen, A. M. L., Tonetti, M. S., Wade, W. G., and Zaura, E. (2016) The oral microbiome - An update for oral healthcare professionals. *Br. Dent. J.* **221**, 657–666
73. Yost, S., Duran-Pinedo, A. E., Teles, R., Krishnan, K., and Frias-Lopez, J. (2015) Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome Med.* **7**, 1–19
74. Deng, Z. L., Szafrański, S. P., Jarek, M., Bhuju, S., and Wagner-Döbler, I. (2017) Dysbiosis in chronic periodontitis: Key microbial players and interactions with the human host. *Sci. Rep.* **7**, 1–13
75. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators (2017) Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. **390**, 1211–1259
76. Kinane, D. F., Stathopoulou, P. G., and Papapanou, P. N. (2017) Periodontal diseases. *Nat. Rev. Dis. Prim.* **3**, 1–14
77. Eke, P., Dye, B., Wei, L., Thornton-Evans, G., and Genco, R. (2012) Prevalence of Periodontitis in Adults in the United States: 2009 and 2010. *J. Dent. Res.* **91**, 914–920
78. Tonetti, M. S., Jepsen, S., Jin, L., and Otomo-Corgel, J. (2017) Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. *J. Clin. Periodontol.* **44**, 456–462
79. Darveau, R. P. (2010) Periodontitis: A polymicrobial disruption of host homeostasis. *Nat. Rev. Microbiol.* **8**, 481–490
80. Zhang, Y., Wang, X., and Yan, F. (2018) Human oral microbiota and its modulation for oral health. *Biomed. Pharmacother.* **99**, 883–893
81. Endo, A., Watanabe, T., Ogata, N., Nozawa, T., Aikawa, C., Arakawa, S., Maruyama, F., Izumi, Y., and Nakagawa, I. (2015) Comparative genome analysis and identification of competitive and cooperative interactions in a polymicrobial disease. *ISME J.* **9**, 629–642
82. Bostanci, N., and Belibasakis, G. N. (2012) Porphyromonas gingivalis: an invasive and evasive opportunistic oral pathogen. *FEMS Microbiol. Rev.* **333**, 1–9
83. How, K. Y., Song, K. P., and Chan, K. G. (2016) Porphyromonas gingivalis: An overview of periodontopathic pathogen below the gum line. *Front. Microbiol.* **7**, 1–14
84. Lamont, R. J., and Jenkinson, H. F. (1998) Life below the gum line: pathogenic mechanisms of Porphyromonas gingivalis. *Microbiol. Mol. Biol. Rev.* **62**, 1244–63
85. Yilmaz, Ö. (2008) The chronicles of Porphyromonas gingivalis: The microbium, the human oral epithelium and their interplay. *Microbiology*. **154**, 2897–2903
86. Tan, K. H., Seers, C. A., Dashper, S. G., Mitchell, H. L., Pyke, J. S., Meuric, V., Slakeski, N., Cleal, S. M., Chambers, J. L., McConville, M. J., and Reynolds, E. C. (2014) Porphyromonas gingivalis and Treponema denticola Exhibit Metabolic Symbioses. *PLoS*

Pathog. 10.1371/journal.ppat.1003955
87. Kuboniwa, M., Tribble, G. D., Hendrickson, E. L., Amano, A., and Richard, J. (2013) Insights into the virulence of oral biofilms: discoveries from proteomics. *Expert Rev. Proteomics*. **9**, 311–323
88. Sakanaka, A., Takeuchi, H., Kuboniwa, M., and Amano, A. (2016) Dual lifestyle of Porphyromonas gingivalis in biofilm and gingival cells. *Microb. Pathog.* **94**, 42–47
89. Hajishengallis, G. (2008) Immune evasion strategies of Porphyromonas gingivalis. *J. Oral Biosci.* **23**, 1–7
90. Sánchez, M. C., Romero-Lastra, P., Ribeiro-Vidal, H., Llama-Palacios, A., Figuero, E., Herrera, D., and Sanz, M. (2019) Comparative gene expression analysis of planktonic Porphyromonas gingivalis ATCC 33277 in the presence of a growing biofilm versus planktonic cells. *BMC Microbiol.* **19**, 1–11
91. Yamamoto, R., Noiri, Y., Yamaguchi, M., Asahi, Y., Maezono, H., and Ebisu, S. (2011) Time course of gene expression during Porphyromonas gingivalis strain ATCC 33277 biofilm formation. *Appl. Environ. Microbiol.* **77**, 6733–6736
92. Shah, H. N., and Williams, R. A. D. (1987) Utilization of glucose and amino acids by Bacteroides intermedius and Bacteroides gingivalis. *Curr. Microbiol.* **15**, 241–246
93. Ohara-Nemoto, Y., Rouf, S. M. A., Naito, M., Yanase, A., Tetsuo, F., Ono, T., Kobayakawa, T., Shimoyama, Y., Kimura, S., Nakayama, K., Saiki, K., Konishi, K., and Nemoto, T. K. (2014) Identification and characterization of prokaryotic dipeptidyl-peptidase 5 from porphyromonas gingivalis. *J. Biol. Chem.* **289**, 5436–5448
94. Nemoto, T. K., and Ohara-Nemoto, Y. (2016) Exopeptidases and gingipains in Porphyromonas gingivalis as prerequisites for its amino acid metabolism. *Jpn. Dent. Sci. Rev.* **52**, 22–29
95. Nelson, K. E., Fleischmann, R. D., Deboy, R. T., Paulsen, I. T., Fouts, D. E., Eisen, J. A., Sean, C., Dodson, R. J., Durkin, A. S., Haft, D. H., Kolonay, J. F., Nelson, W. C., Mason, T., Tallon, L., Gray, J., Granger, D., Tettelin, H., Dong, H., Galvin, J. L., Margaret, J., Dewhirst, F. E., Fraser, C. M., Daugherty, S. C., Gwinn, M., and Duncan, M. J. (2003) Complete Genome Sequence of the Oral Pathogenic Bacterium Porphyromonas gingivalis Strain W83. *J. Bacteriol.* **185**, 5591–5601
96. Pomowski, A., Usón, I., Nowakowska, Z., Veillard, F., Sztukowska, M. N., Guevara, T., Goulas, T., Mizgalska, D., Nowak, M., Potempa, B., Huntington, J. A., Potempa, J., and Gomis-Rüth, F. X. (2017) Structural insights unravel the zymogenic mechanism of the virulence factor gingipain K from Porphyromonas gingivalis, a causative agent of gum disease from the human oral microbiome. *J. Biol. Chem.* **292**, 5724–5735
97. Kuramitsu, H. K. (2001) Virulence Properties of Oral Bacteria: Impact of Molecular Biology. *Issues Mol. Biol.* **3**, 35–36
98. Bostanci, N., and Belibasakis, G. N. (2012) Porphyromonas gingivalis: An invasive and evasive opportunistic oral pathogen. *FEMS Microbiol. Lett.* **333**, 1–9
99. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, 457–462
100. Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and Mering, C. Von (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, 362–368
101. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and

Bairoch, A. (2005) Protein Identification and Analysis Tools on the ExPASy Server. *Proteomics Protoc. Handb.* 10.1385/1592598900
102. Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., Brinkman, F. S. L., and Va, C. (2005) Sequence analysis subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*. **21**, 617–623
103. Dellaire, G., Farrall, R., and Bickmore, W. A. (2003) The Nuclear Protein Database (NPD): Sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res.* **31**, 328–330
104. *Predicting Secretory Proteins with SignalP* (2017) 1611th Ed., Methods in Molecular Biology
105. Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., and Madden, T. L. (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*. 10.1186/1745-6150-7-12
106. Enright, A. J., Dongen, S. Van, and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584
107. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015) The Phyre2 web portal for protein modeling , prediction and analysis. *Nat. Protoc.* **10**, 845–858
108. Ma, J., Wang, S., Zhao, F., and Xu, J. (2013) Protein threading using context-specific alignment potential. *Bioinformatics*. **29**, 257–265
109. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., Thompson, J. D., Higgins, D. G., Mcwilliam, H., Remmert, M., and So, J. (2011) Fast , scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 10.1038/msb.2011.75
110. Keenleyside, W. J., Clarke, A. J., and Whitfield, C. (2001) Identification of Residues Involved in Catalytic Activity of the Inverting Glycosyl Transferase WbbE from Salmonella enterica Serovar Borreze. *J. Bacteriol.* **183**, 77–85
111. He, Y., Macauley, M. S., Stubbs, K. A., Vocadlo, D. J., and Davies, G. J. (2010) Visualizing the Reaction Coordinate of an O-GlcNAc Hydrolase. *J. Am. Chem. Soc.*
112. Wiggins, C., and Munro, S. (1998) Activity of the yeast MNN1 α-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 7945–7950
113. Veiga, L. A., and L, C. E. (1967) Detection and Differentiation of Sugars on Paper by the p-Anisidinel Periodate Reaction. *Anal. Biochem.* **20**, 419–422
114. Duus, J. Ø., Gotfredsen, C. H., and Bock, K. (2000) Carbohydrate Structural Determination by NMR Spectroscopy: Modern Methods and Limitations. *Chem. Rev.* **100**, 4589–4614
115. Matthiesen, R. (2013) *Mass Spectrometry Data Analysis in Proteomics. Methods in Molecular Biology*. (Matthiesen, R. ed), Humana Press, 10.1007/978-1-62703-392-3
116. Holcomb, J., Spellmon, N., Zhang, Y., Doughan, M., Li, C., and Yang, Z. (2017) Protein crystallization: Eluding the bottleneck of X-ray crystallography. *AIMS Biophys.* **4**, 557–575
117. Moon, A. F., Mueller, G. A., Zhong, X., and Pedersen, L. C. (2010) A synergistic approach to protein crystallization: Combination of a fixed-arm carrier with surface entropy reduction. *Protein Sci.* **19**, 901–913
118. Veeraraghavan, S., Baleja, J. D., and Gilbert, G. E. (1998) Structure and topography of the

membrane-binding C2 domain of factor VIII in the presence of dodecylphosphocholine micelles. *Biochem. J.* **332**, 549–555
119. Ash, M. R., Guilfoyle, A., Clarke, R. J., Guss, J. M., Maher, M. J., and Jormakka, M. (2010) Potassium-activated GTPase reaction in the G protein-coupled ferrous iron transporter B. *J. Biol. Chem.* **285**, 14594–14602
120. Hirschi, M., Johnson, Z. L., Lee, S., and Carolina, N. (2017) Visualizing multistep elevator-like transitions of a nucleoside transporter. *Nature*. **545**, 66–70
121. Su, C.-C., Reddy-Bolla, J., Kumar, N., Radhakrishnan, A., Long, F., Delmar, J., Chou, T.-H., Rajashankar, K., Shafer, W., and Yu, E. (2015) Structure and function of Neisseria gonorrhoeae MtrF illuminates a class of antimetabolite efflux pumps. *Cell Rep.* **11**, 61–70
122. Johnson, Z., Cheong, C.-G., and Lee, S.-Y. (2017) Crystal structure of a concentrative nucleoside transporter from vibrio cholerae at 2.4Å. *Nature*. **483**, 489–493
123. Hung, K.-W., Chang, Y.-W., Eng, E., Chen, J.-H., Chen, Y.-C., Sun, Y.-J., Hsiao, C.-D., Dong, G., Spasov, K., Unger, V., and Huang, T.-H. (2010) Structural fold, conservation and Fe(II) binding of the intracellular domain of prokaryote FeoB. *J. Struct. Biol.* **170**, 501–512
124. Oppenheim, A. (1981) Separation of closed circular DNA from linear DNA by electrophoresis in two dimensions in agarose gels. *Nucleic Acids Res.* **9**, 6805–6812
125. Juers, D. H., Matthews, B. W., and Huber, R. E. (2012) LacZ β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance. *Protein Sci.* **21**, 1792–1807
126. Taylor, R. C., Brown, A. K., Singh, A., Bhatt, A., and Besra, G. S. (2010) Characterization of a beta-hydroxybutyryl-CoA dehydrogenase from Mycobacterium tuberculosis. *Microbiology*. **156**, 1975–1982
127. Niederman, R. (1996) The relationship of gingival crevicular fluid short chain carboxylic acid concentration to gingival inflammation. *J. Clin. Periodontol.* **23**, 743–749
128. Sato, M., Yoshida, Y., Nagano, K., Hasegawa, Y., Takebe, J., and Yoshimura, F. (2016) Three CoA transferases involved in the production of short chain fatty acids in Porphyromonas gingivalis. *Front. Microbiol.* **7**, 1–13
129. Kurita-Ochiai, T., Fukushima, K., and Ochiai, K. (1997) Butyric acid-induced apoptosis of murine thymocytes, splenic T cells, and human Jurkat T cells. *Infect. Immun.* **65**, 35–41
130. Kurita-Ochiai, T., Ochiai, K., and Fukushima, K. (1998) Volatile fatty acid, metabolic by-product of periodontopathic bacteria, induces apoptosis in WEHI 231 and RAJI B lymphoma cells and splenic B cells. *Infect. Immun.* **66**, 2587–2594
131. Singer, R. E., and Buckner, B. A. (1981) Butyrate and propionate: Important components of toxic dental plaque extracts. *Infect. Immun.* **32**, 458–463
132. Farewell, A., Diez, A. A., DiRusso, C. C., and Nyström, T. (1996) Role of the Escherichia coli FadR regulator in stasis survival and growth phase-dependent expression of the uspA, fad, and fab genes. *J. Bacteriol.* **178**, 6443–6450
133. Andersen, J. T., and Jensen, K. F. (1989) Molecular and mutational analysis of three genes preceding pyrE on the Escherichia coli chromosome. *Mol. Microbiol.* **3**, 393–404
134. Lundegaard, C., and Jensen, K. F. (1994) The DNA Damage-Inducible dinD Gene of Escherichia coli Is Equivalent to orfY Upstream of pyrE. *J. Bacteriol.* **176**, 3383–3385
135. Poulson, P., and Jensen, K. F. (1991) Three genes preceding pyrE on the Escherichia coli chromosome are essential for survival and normal cell morphology in stationary culture and at high temperature. *Res. Microbiol.* **142**, 283–288

136. Christopher, A. B., Arndt, A., Cugini, C., and Davey, M. E. (2010) A streptococcal effector protein that inhibits Porphyromonas gingivalis biofilm development. *Microbiology*. **156**, 3469–3477
137. Chen, W., Honma, K., Sharma, A., and Kuramitsu, H. K. (2006) A universal stress protein of Porphyromonas gingivalis is involved in stress responses and biofilm formation. *FEMS Microbiol. Lett.* **264**, 15–21
138. Velayudhan, J., Hughes, N. J., McColm, A. A., Bagshaw, J., Clayton, C. L., Andrews, S. C., and Kelly, D. J. (2000) Iron acquisition and virulence in Helicobacter pylori: A major role for FeoB, a high-affinity ferrous iron transporter. *Mol. Microbiol.* **37**, 274–286
139. Lewis, J. (2010) Metal uptake in host-pathogen interactions: Role of iron in Porphyromonas gingivalis interactions with host organisms. *Periodontol. 2000*. **52**, 94–116
140. Olczak, T., Simpson, W., Liu, X., and Genco, C. A. (2005) Iron and heme utilization in Porphyromonas gingivalis. *FEMS Microbiol. Rev.* **29**, 119–144
141. Patching, S. G., Baldwin, S. A., Baldwin, A. D., Young, J. D., Gallagher, M. P., Henderson, P. J. F., and Herbert, R. B. (2005) The nucleoside transport proteins, NupC and NupG, from Escherichia coli: Specific structural motifs necessary for the binding of ligands. *Org. Biomol. Chem.* **3**, 462–470
142. Lewis, J. P., Iyer, D., and Anaya-Bergman, C. (2009) Adaptation of Porphyromonas gingivalis to microaerophilic conditions involves increased consumption of formate and reduced utilization of lactate. *Microbiology*. **155**, 3758–3774
143. Delmar, J. A., and Yu, E. W. (2016) The AbgT family: A novel class of antimetabolite transporters. *Protein Sci.* **25**, 322–337
144. Kuboniwa, M., Houser, J., Hendrickson, E. L., Wang, Q., Alghamdi, S., Sakanaka, A., Miller, D., Hutcherson, J., Wang, T., Beck, D., Whiteley, M., Amano, A., Wang, H., Marcotte, E., Hackett, M., and Lamont, R. (2017) Metabolic crosstalk regulates Porphyromonas gingivalis colonization and virulence during oral polymicrobial Infection. *Nat. Microbiol.* **2**, 1493–1499
145. Grewer, C., Zhang, Z., Mwaura, J., Albers, T., Schwartz, A., and Gameiro, A. (2012) Charge compensation mechanism of a Na+-coupled, secondary active glutamate transporter. *J. Biol. Chem.* **287**, 26921–26931

APPENDICES

Appendix 1: List of Bioinformatics Tools

Program	Information Provided	Web Address
NCBI BLASTp	Runs a query sequence against a database of other sequences to find local similarities as well as annotated domains	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
EMBL-EBI InterPro	Runs a query sequence and predicts domains, families, superfamilies, signal regions, and motifs through comparison to sequence databases	https://www.ebi.ac.uk/interpro/
UniProt	Provides amino acid sequences and links to associated publications, database entries, and various bioinformatics tools	https://www.uniprot.org/
CAZy Database	Collection of information on hypothetical and validated CAZymes Links to characterized CAZymes within enzyme families	http://www.cazy.org/
Kyoto Encyclopedia of Genes and Genomes	Database of annotated genomes for various organism and the genes contained within them Links from genes/genomes to other bioinformatics resources/databases	https://www.genome.jp/kegg/
ExPASy ProParam	Computes various biophysical properties of submitted amino acid sequences including: molecular weight, theoretical pI, extinction coefficient, and instability index score	https://web.expasy.org/protparam/
Phyre^2	Automated homology-modelling of amino acid sequences to find likely 3-dimensional structure; ranked list of previously-characterized template proteins	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
RaptorX	Automated homology-modelling of amino acid sequences to find likely 3-dimensional structure	http://raptorx.uchicago.edu/
RCSB Protein Data Bank	Experimentally-determined 3-dimensional protein structures and related publications/database links	https://www.rcsb.org/
STRING	Generates interactive maps of gene/protein interactions around a central node gene	https://string-db.org/
SignalP4.1	Predicts the presence and location of signal peptides within submitted protein sequence	http://www.cbs.dtu.dk/services/SignalP-4.1/
ExPASy TMPred	Predicts areas of a protein sequence likely to traverse a membrane bilayer	https://embnet.vital-it.ch/software/TMPRED_form.html
PSORTb	Predicts the subcellular localization of proteins based on amino acid sequence	https://www.psort.org/psortb/
Clustal Omega	Aligns multiple amino acid sequences based on the shared similarity	https://www.ebi.ac.uk/Tools/msa/clustalo/

Appendix 2: Amino Acid Sequences of Protein Constructs

BT2109 - pET15b - amp-resistant	MKNIAFILLILAAVCFANCAGDYEAEPRGTNDGKAPAPITNISVKNVNGGAIIKYTPSDD VDLLYVKAIFTNTRGEQQEVKSSMYVDSLVIEGLGDTQKREVKLYSVDRHENYSTPAIVT IEPLTPAVVQTQESLKVTESFGGFFLDFNNAARSALSFYIYKWVDEHNEYEIHDVYSSQQ EEGRLTIKGLEHKLLKLAIFVRDKWENTSDTLYAEITPWKESLLDKTKFQLVKVMDDVSW DYHEGYHSRLWDGQIGGWNWGHTEYPIPFPHAFSIDLNVNVRLSQFQFWQRLDSQDLLYA HGAPKYFKLYGCEEGKDLQNPDNWVVLFDGEMKKPSGGAFGDALTQEDIEEAQRGHVFTL NQDVPFIRYFRFQSLMSWSGMETSVMSEITLWGDIQGEE
BT2857 - pET15b - amp-resistant	MISRKFTIWACGLLFVFGSCAEDMVKPIENDKDAPGTVQDVKTESLPGAVKFTYTLPSDP DLLYVLAKYTNKTGKVMEFRSSFYTNSVTVEGFGDTDTYKVELYTVDRSENRSQPQIVEV VPLTPPILSCYESLSMVSDFGGMTFEMDNKFKSDLAIYVCTPDELGDMVLAETFYSAREE IVYSVRGYDAVPRRFGIFLQDKYGNETDTLFTELTPIYERELDKTKFREMYLQNDSRVDS YDGKMEYVWNGRISKDGDSGGVGLHTGTGTKDGPAVFTFDLGVLAKLSRFALWAIQDEKH FYNDMSPRRYEVWGCATEPNPDGSWDQWVKLLDMENVKPSGSPIGILTEDDIEAAKIGDQ ANVPLDMPRVRYIRIKCLKNWSNNYNICFTELTFWGNDNEE
BT2857 N-terminal truncation - pET21b(+) - amp-resistant	VKPIENDKDAPGTVQDVKTESLPGAVKFTYTLPSDPDLLYVLAKYTNKTGKVMEFRSSFY TNSVTVEGFGDTDTYKVELYTVDRSENRSQPQIVEVVPLTPPILSCYESLSMVSDFGGMT FEMDNKFKSDLAIYVCTPDELGDMVLAETFYSAREEIVYSVRGYDAVPRRFGIFLQDKYG NETDTLFTELTPIYERELDKT
BT2857 C-terminal truncation - pET21b(+) - amp-resistant	ELDKTKFREMYLQNDSRVDSYDGKMEYVWNGRISKDGDSGGVGLHTGTGTKDGPAVFTFD LGVLAKLSRFALWAIQDEKHFYNDMSPRRYEVWGCATEPNPDGSWDQWVKLLDMENVKPS GSPIGILTEDDIEAAKIGDQANVPLDMPRVRYIRIKCLKNWSNNYNICFTELTFWGND
BT2918 - pET15b - amp-resistant	MKKKLYKYGLLYIGVALAACADNADLNEPAGSTTPPAQVLNATVKNLPGAAIIYYDLPDD QNLKYVRASYKVDNMVRTVNASFYTDSLVVEGFPTKGEYDIELYSVSYGEVVSSPLVVKV NPDTPPYQKVRGTLVSAETFGGIRVNFDNPEKAKLGLGVIKKQAEGIWTQVYMHYTEAKG GDFYVRGLDAVTTDFGIFVRDRWGHLSDTLYVTETPLYEEQCDKSLFRKMALPTDSYECH SWNEVTKGNDMTRLWDGITDADPCFQTKTTTVMPQWFTFDMGENYKLSRFVMVSRYYPGK YGNTFKAGHPKHFEIWGSINPNPDGSFDDSWVLLSEYESVKPSGGGVNDALTAEDQEAAK NGENFIIPDNAPAVRYIRFRTNNTWGNTRYMHLHELTFFGAKHN
BT2966 - pET28a - kan-resistant	MKKIIYSSLLVLAITLLPSACDNDSDMNSPHGSTTPPEQVVLNGVKNLPGKSIIYFTQPS DKNFLYMKAVYMTEIGERSTNASYFADSVVVEGFEKAGEYNVKLYSVSPGEAYSEPLEVK INPEEPPYITAYKEMEIKPDFAGIRIKTSNTSNETLTFYVYRKDATGKWTEAGALYTKKQ EINEPIRGMEAVPSEFSVVVKDRWGHLSESKEISLTPYYEEEVDKKKMGYLAIGEYKGYL APNANTPKNLYDGIIGSNNTFMTLTTAGYDFTKPSSVTLDLGKKFKLSRMIVYGRRNGTD YSSIFDGLYPKEIEIWGRNDNNVTKFDPENDEGWVRLYQGVLPRADGSVIPAAIVPLTDA DKELARDGNELEFSVDLDAYRYVTFVCIKNYNVGTSRINVAELTFFGTPEDKIIQ
BT3158 - pET15b - amp-resistant	MKKRILYYIWIACAAVLLHSCEDVETHKPYGDGSGHAPGTVQVTSYEQIPGGVTLKFIAP TDEDLLYIKIKYTLDNGKEMEARASLYTDETTIEGFGNTNPKKLVISAVNKMEKEGEAII TEIVPGKPAYLTAIEEIEVNPTFGGIYIHTTNGGRNYLIFDVSTKDSKGNWNIEHTEYTS VKNIGFTLRGFSAEPHDFKVRVRDLYDNQSEEYLTTLTPLYEEKLDLTKFKTFYLANDIK MDNAGHTLESLFNGDHGLNSWNYAHGYDFNPSEFPVWFTFDMGQTAQLSRFTSWQRSMGG SYYYRAGAIKEWEVWGRSDLPSSDGSWDGWTKLADCESIKPSGWPTGSNSEEDITYASKG EEFEFLADIPPVRYIRFKILSTHDGAGLVVMQQLWFYGTPIQ
PGN1176 - pET21b(+) - amp-resistant	MKIAVVGTGAMGSGIVQCIAQNGLDVLMKGRSETSISKSLKGIEKNLNRLVEKGKMSADE QQAVMARIRTTMDFNDLKDAGVVIEAIAEEMEIKKQALSAVAEVVSEQCIIATNTSSLSI TALATELPTPSRVIGMHFFNPVPVMKLVEVISAQQTSDEVMKLVLDLCEMLKKTAVRVNE APGFVVNRLLIPMINEGIGAYADGVASIEEIDQAMMLGANHPMGPLALGDFIGLDVCLAI MEVLFTEFGDAKYRPHPLLRKMVRAGQLGRKTGKGFYDYSK

PGN1459 - pET21b(+) - amp-resistant	MTGFGKHTIQFADKKITATIKSLNSKQFDLSTKISSRYRDRELPLRALVSADIRRGKVDL SLFLEESGSTDVVSQIFDTEKMSSYFFQLKEFGEKMGVEVPAGGWFEQLLRIPGVIQIDE DKEEEIPEDEWAVVIETCRRALDQLIGFREQEGAMLEQVFTEKISNISSLLLQIEQYEPE RIQRIKERIEENLTKISEKDYDKNRFEQEMIYYIEKLDVNEEKHRLRNHLSYFLQTMQGE PGQGKKLSFIAQEIGREINTLGSKSNHVEMQQIVVQMKDELEQIKEQVLNVL
PGN1461 - pET21b(+) - amp-resistant	MVLNYIFVAFFLVSLIVALVRLIAFGDVSVFDEMVQASFSQSKNAFEIALALTGVLALWL GLMKIAERSGLISKLAQISSPVLSRLFPSVPKGHPALGNIFMNISANLLGLDNAATPLGI RSMESLQTINQQKDKASDAMIMFLAINASGLTLIPSSIMAYRLQAGAANPADIFVPVLIA TFASTLVAILAVGVKQRIDFFQKPLLLFLLSSLLFIGGVIYAGRTLPPETFSSISTAFAS ILLFGIMCGFVVSGLRARINVYDAFIEGAKDGFGTAVTIIPYLLAMLVGIGIFRASGAMD LITEGLRGLVGFWGGDMKFVEGIPTMLMKPLSGSGARGLMVDAMQTYGADSFVGRLCSIV QGSSDTTFYVVALYYGAVKIRNTRYTVSYSLLADLAGALTAIAVTYIFFG

Appendix 3: ProtParam Outputs for Protein Constructs

Construct	**Molecular Weight (Da)**	**Residue Number**	**Isoelectric Point**	**Extinction Coefficient ($M^{-1}cm^{-1}$)**
BT2109	45588.19	399	4.81	87320
BT2857	45578.32	401	4.59	74175
BT2857N	22774.60	201	4.42	19370
BT2857C	20327.78	178	4.87	48930
BT2918	45473.12	404	5.47	81250
BT2966	46565.60	415	4.99	62230
BT3158	45511.95	402	4.98	94770
PGN1176	30362.61	281	5.76	5960
PGN1459	34049.74	292	5.03	21430
PGN1461	43658.65	410	8.97	28880

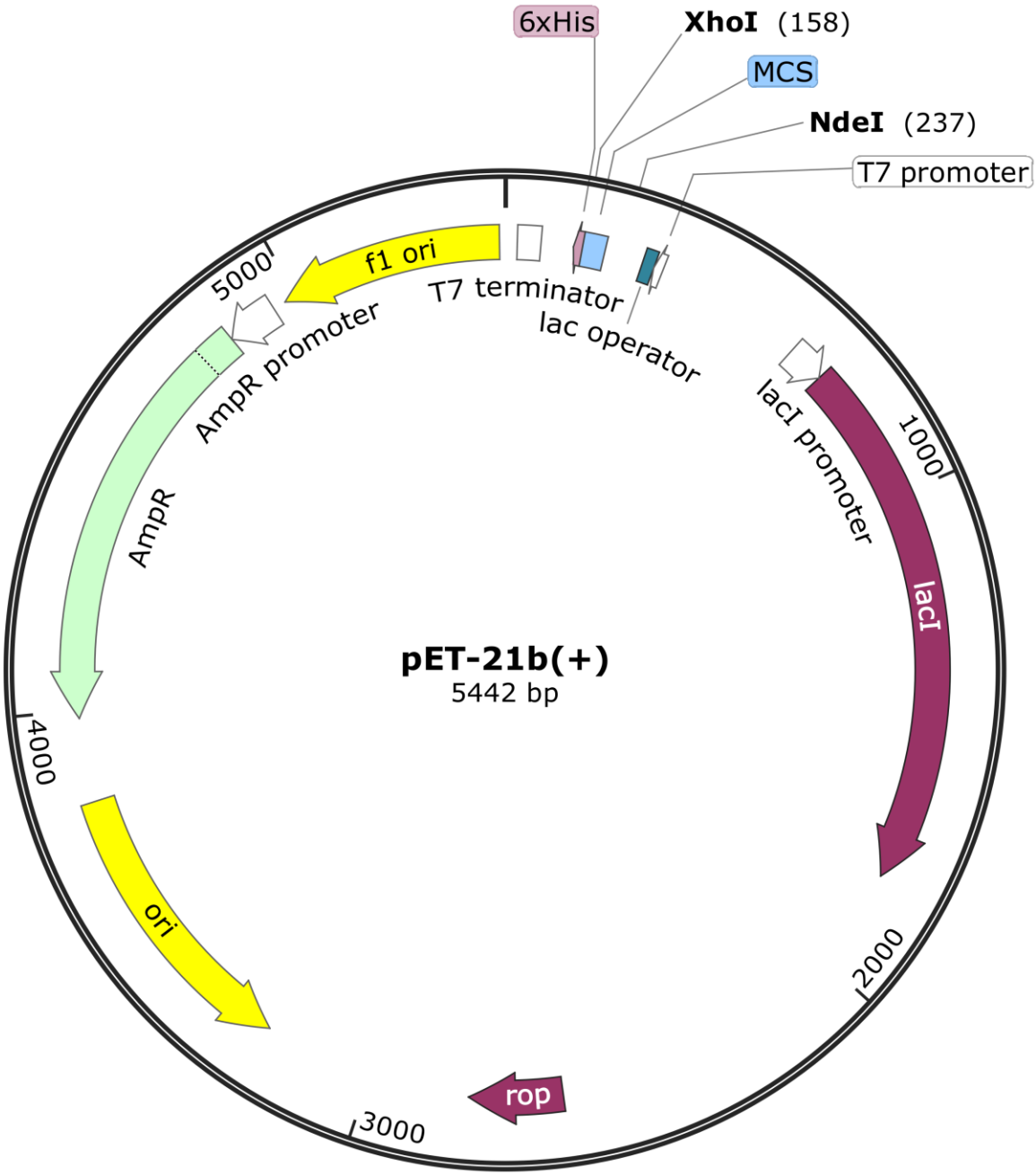
Appendix 4: List of Primers for Site-Directed Mutagenesis and Truncation

BT2857N D35A	
Forward Primer	5′-ACCCTGCCGTCTGCGCCGGATCT-3′
Reverse Primer	5′-CAGCAGATCCGGCGCAGACGGCA-3′
BT2857N D37A	
Forward Primer	5′-CGTCTGATCCGGCGCTGCTGTATGTT-3′
Reverse Primer	5′-AACATACAGCAGCGCCGGATCAGACG-3′
BT2857N D83A	
Forward Primer	5′-TACACCGTTGACGCCTCTGAAAACCG-3′
Reverse Primer	5′-CGGTTTTCAGAGGCGTCAACGGTGTA-3′
BT2857N R84A	
Forward Primer	5′-TACACCGTTGACGCCTCTGAAAACCGT-3′
Reverse Primer	5′-ACGGTTTTCAGAGGCGTCAACGGTGTA-3′
BT2857C D239A	
Forward Primer	5'-TTTTACCGTCGTAAGAGGCAACACGAGAGTCGTTC-3'
Reverse Primer	5'-GAACGACTCTCGTGTTGCCTCTTACGACGGTAAAA-3'
BT2857C D242A	
Forward Primer	5′-ATTTTACCGGCGTAAGAGTCAACACGA-3′
Reverse Primer	5′-GACTCTTACGCCGGTAAAATGGAATACGTT-3′
BT2857C H300A	
Forward Primer	5′-GGTACCGGTCGGCAGACCAACGCCA-3′
Reverse Primer	5′-GTTGGTCTGGCCACCGGTACCGGCA-3′
BT2857C D304A	
Forward Primer	5′-ATAACGACGCGGAGACATGGCGTTGTAGAA-3′
Reverse Primer	5′-TTCTACAACGCCATGTCTCCGCGTCGTTAT-3′
BT2857N-DUF4959 Truncation	
BT2857N 4959 FlankR	5′-CCGCTCACAATTCCCCTATAGTGAGTCG-3′
BT2857N 4959 GeneR	5′-CTCGAGCGGAACAACTTCAACGATCTG-3′
BT3158-DUF4959 Truncation	
BT3158 4959 NF	5′-CATATGTGCGAAGATGTAGAAACTCATAAA-3′
BT3158 4959 CF	5′-CTCGAGGGTACTATTTCGGTGATGATA-3′

Appendix 5: Amino Acid Sequence Results for Mutant BT2857C and BT2857N Constructs

BT2857N D37A - 5	MVKPIENDKDAPGTVQDVKTESLPGAVKFTYTLPSDPALLYVLAKYTNKTGKVMEFRSSFYTNSV TVEGFGDTDTYKVELYTVDRSENRSQPQIVEVVPLTPPILSCYESLSMVSDFGGMTFEMDNKFKS DLAIYVCTPDELGDMVLAETFYSAREEIVYSVRGYDAVPRRFGIFLQDKYGNETDTLFTELTPIY ERELDKTLEHHHHHH
BT2857N R84A - 2	MVKPIENDKDAPGTVQDVKTESLPGAVKFTYTLPSDPDLLYVLAKYTNKTGKVMEFRSSFYTNSV TVEGFGDTDTYKVELYTVDASENRSQPQIVEVVPLTPPILSCYESLSMVSDFGGMTFEMDNKFKS DLAIYVCTPDELGDMVLAETFYSAREEIVYSVRGYDAVPRRFGIFLQDKYGNETDTLFTELTPIY ERELDKTLEHHHHHH
BT2857C D239A - 4	MELDKTKFREMYLQNDSRVASYDGKMEYVWNGRISKDGDSGGVGLHTGTGTKDGPAVFTFDLGVL AKLSRFALWAIQDEKHFYNDMSPRRYEVWGCATEPNPDGSWDQWVKLLDMENVKPSGSPIGILTE DDIEAAKIGDQANVPLDMPRVRYIRIKCLKNWSNNYNICFTELTFWGNDLEHHHHHH

Appendix 6: pET21b(+) Vector with XhoI and NdeI Cut Sites Labelled to Insertion of Truncated BT2857 and BT3158 Constructs



Appendix 7: Crystal Diffraction Preliminary Statistics for Each Protein Construct

BT2857C

Statistic	**BT2857C 3-7**	**BT2857Clacb3**
Resolution (Å)	46.69 – 1.45 (1.53 – 1.45)	46.87 – 1.45 (1.53 – 1.45)
Space Group	$P2_1$	$P2_1$
Unit Cell Dimensions: a, b, c, (β)	59.14, 60.94, 61.63 (101.39)	59.30, 61.08, 61.76 (101.59)
Measured Reflections	501995 (50476)	495287 (67903)
Unique Reflections	74234 (9268)	76475 (11085)
Multiplicity	5.4 (3.9)	6.5 (6.1)
Completeness (%)	99.3 (85)	99.9 (97.6)
Mosaicity	0.39	0.12
R_{merge}	2.3 (6.5)	10.2 (28.1)
I/σ(I)	24.1 (16.0)	11.4 (4.7)
$CC_{1/2}$(%)	1.000 (0.995)	0.997 (0.967)

BT3158

Statistic	**BT3158 7-8**	**BT3158GAL_01**
Resolution (Å)	46.96 – 1.53 (1.61 – 1.53)	40.66 – 1.80 (1.90 – 1.80)
Space Group	C121	C121
Unit Cell Dimensions: a, b, c, (β)	59.14, 60.94, 61.63 (101.39)	75.14, 48.96, 113.22 (103.76)
Measured Reflections	424568 (50476)	138416 (19994)
Unique Reflections	64687 (9268)	37206 (5362)
Multiplicity	6.6 (5.4)	3.7 (3.7)
Completeness (%)	99.8 (98.8)	99.8 (99.9)
Mosaicity	0.39	0.44
R_{merge}	0.025 (0.058)	0.72 (0.389)
I/σ(I)	50.1 (24.1)	11.1 (3.2)
$CC_{1/2}$(%)	1.000 (0.997)	0.996 (0.869)

BT3158-SelenoMet

Statistic	**BT3158 7-16**
Resolution (Å)	40.27 – 2.01 (2.12 – 2.01)
Space Group	C121
Unit Cell Dimensions: a, b, c, (β)	76.18, 49.12, 115.21, (105.03)
Measured Reflections	102912 (15076)
Unique Reflections	27426 (3967)
Multiplicity	3.8 (3.8)
Completeness (%)	99.3 (98.9)
Mosaicity	0.29
R_{merge}	0.053 (0.270)
I/σ(I)	15.9 (5.0)
$CC_{1/2}$(%)	0.999 (0.944)

PGN1176 Screening Summary

	Pgn1176 6-8	**Pgn1176 6-9**
Resolution (Å)	6.7	6.8
Space Group	P1	P1
Unit Cell Dimensions: a, b, c, (β)	149.00, 158.42, 158.42, (90)	152.54, 161.55, 161.89 (90)
Multiplicity	2.0	2.0
Completeness (%)	99.5	99.5
Mosaicity	0.32	0.30
I/σ(I)	0.0	0.1

PGN1461

	Pgn1461 4-7	**Pgn1461 4-5**
Resolution (Å)	47.99 - 2.10 (2.21 – 2.10)	47.99 – 2.16 (2.28 – 2.16)
Space Group	$P2_12_12_1$	$P2_12_12_1$
Unit Cell Dimensions: a, b, c, (β)	122.13, 128.20, 171.39 (90)	122.07, 128.07, 171.10 (90)
Measured Reflections	2064190 (280695)	1917127 (281916)
Unique Reflections	156796 (22686)	144252 (20851)
Multiplicity	13.2 (12.4)	13.3 (13.5)
Completeness (%)	99.9 (100.0)	99.9 (100.0)
Mosaicity	0.13	0.16
R_{merge}	0.259 (1.062)	0.234 (0.840)
I/σ(I)	9.1 (2.4)	9.7 (3.2)
$CC_{1/2}$ (%)	0.999 (0.903)	0.996 (0.943)