Mechanistic insights into N-glycan degradation by the human gut microbiota

Justina Briliūtė

A thesis submitted for the degree of Doctor of Philosophy

September 2018

Institute for Cell and Molecular Biosciences

Faculty of Medical Sciences

Newcastle University

Abstract

The gut microbiota, a dense microbial community present in the human large intestine, plays a significant role in the maintenance of human health. Apart from digestion of complex carbohydrates, these beneficial mutualists compete with invading pathogens and are involved in the modulation of human immune system.

Despite being home to hundreds of different species of bacteria, two bacteria phyla dominate the human gut – Gram-positive *Firmicutes* and Gram-negative *Bacteroidetes*. *Bacteroides spp.*, prominent members of this microbial ecosystem, degrade a vast range of dietary and host-derived glycans by utilizing a complex trans-envelope machinery known as Sus-like systems encoded by co-regulated clusters of genes known as polysaccharide utilization loci (PUL). Each *Bacteroides* spp. can have >100 PULs, each specialised in targeting and degrading different glycans, with majority of glycans having 1 to 2 dedicated PULs.

Glycans are the major nutrients available to the human colonic microbiota and come from both dietary and host sources. While dietary plant glycans are the most prominent in terms of quantity, the ability to access host mucin O-glycans appears to be important for gut survival and is a trait shared by many members of the microbiota. Various N-glycosylated dietary and host proteins are also found in the human gut in significant quantities. These include antibodies and the terminal domains of mucins, however, it is currently unclear whether these N-glycans are also accessed as nutrients by the gut microbiota and if so, how they are broken down.

In this study, several prominent *Bacteroides* spp. were shown to be capable of utilising complex N-glycans as a sole carbon source. Transcriptomic analysis identified the genes upregulated in one of these, *B. thetaiotaomicron,* during the growth on complex N-glycans. These include 16 predicted carbohydrate-active enzymes encoded by multiple discrete loci, some of which, such as BT0455-0461, do not fit classical PUL model due to the lack of SusC/D pair, SGBP and a regulator. Biochemical characterisation of the activated enzymes provided an insight into their role in N-glycan breakdown. Four GH20-family β-hexosaminidases with overlapping specificities were characterised and were found to employ a novel CBM domain to bind N-glycan structures. The crystal structure of one of these GH20s, BT0459^{GH20}, suggests this glycosidase could utilize a novel strategy to target N-glycan substrates.

The combination of the cellular localization, gene deletion analysis and biochemical characterization data allowed us to propose a model for N-glycan degradation where the structure is degraded sequentially and requires a cooperative activity of numerous glycosidases encoded by multiple discrete loci. The N-glycan degradation begins at the cell surface with a cleavage of the complex N-glycan structures off the protein backbones by endo-glycosidase BT1044^{GH18}. This step was found to be critical in complex bi-antennary N-glycan utilization. The liberated N-glycan structures are then imported into the cell where they are sequentially broken down into monosaccharides by the specialised exo-glycosidases.

Variants of this intricate N-glycan utilization apparatus were identified in a number of other members of *Bacteroides spp.*, suggesting this model of degradation could be widely employed by the human gut *Bacteroides*.

Acknowledgements

I would like to thank my supervisor Dr. David Bolam for his support and excellent advice throughout the course of my PhD.

I would also like to acknowledge and offer a heartful thanks to Dr. Lucy Crouch for the collaboration, helpful advice and support. I would also like to thank all of the members of the Bolam lab (M2035), past and present, for their optimism, support and advice throughout the last four years. I especially would like to thank Carl Morland for his help and witty sense of humour that kept the working atmosphere light and pleasant.

I would also like to thank Dr. Arnaud Baslé for his help with crystallography experiments.

And finally, I would like to thank my mother, Reda, without whom I would have never made it.

Contents

List of figures

Chapter 1: Introduction

Page

Chapter 2: Materials and Methods

Chapter 3: Complex N-Glycan metabolism by members of human gut microbiota

Chapter 4: Characterization of the key members of N-glycan utilization apparatus in *B. thetaiotaomicron*

Chapter 5: Proposed model of N-glycan utilization by *B. thetaiotaomicron*

Chapter 6: Final discussion

List of tables

Chapter 2: Materials and Methods

Chapter 3: Characterization of the key members of N-glycan utilization apparatus in *B. thetaiotaomicron*

Chapter 5: Proposed model of N-glycan utilization by *B. thetaiotaomicron*

Appendix

Abbreviations

B. theta/Bt	*Bacteroides thetaiotaomicron*
BHI	Brain Heart Infusion
BLAST	Basic Local Alignment Search Tool
CAZy	Carbohydrate Active Enzyme database
CBM	Carbohydrate-Binding Module
DNA	Deoxyribonucleic acid
DUF	Domain of Unknown Fuction
E. coli	*Escherichia coli*
GIT	Gastrointestinal tract
GH	Glycoside Hydrolase
GalNAc	N-Acetylgalactosamine
GlcNAc	N-Acetylglucosamine
HEPES	4-(2-Hydroxyethyl) piperazine-1-ethanesulfonic acid
HGM	Human gut microbiota
HPLC	High-performance liquid chromatography
IM	Inner membrane
ITC	Isothermal Titration Calorimetry
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KDO	3-deoxy-D-manno-2-octulosonic acid
KDN	2-keto-3-deoxy-D-glycero-D-galacto-nononic acid
KOD	DNA polymerase enzyme
MM	Minimal media
MST	Microscale thermophoresis
MW	Molecular weight
ORF	Open reading frame
PCR	Polymerase chain reaction
PFAM	protein family, annotation, multiple sequence alignment database
PGM III	Porcine gastric mucin type III (bound sialic acid, 0.5-1.5%)
pNP	Paranitrophenol
PUL	Polysaccharide utilization locus
RNA	Ribonucleic acid
SDS-PAGE	Sodium dodecyl sulfate – polyacrylamide gel electrophoresis
SGBP	Sodium glycan binding protein
SP	Signal peptide
Sus	Starch utilization system
TEMED	N,N,N',N'-Tetramethylethylenediamine
TLC	Thin layer chromatography
TYG	Tryptone Yeast Extract Glucose Media

Chapter 1: Introduction

1.1. The Human Gut Microbiota

The human body is home to a vast collection of microorganisms that have co-evolved with the host to create a highly complex, accomodating, habitat-specific relationships. These include bacteria, viruses, fungi, archaea and protists (Wang *et al.,* 2016). The mucosal surfaces of the human gastrointestinal tract (GIT) are home to one of the largest populations of microorganisms that usually maintain mutualistic relationship with the host (**Figure 1.1**). This vast and highly competitive microbial ecosystem is known as the gut microbiota (Rothschild *et al.,* 2018). The human gut microbiota is established shortly after birth and is incredibly dense, diverse and dynamic. It is estimated that there are trillions of organisms present in the human colon and they collectively encode more genes than the host, the majority of which are responsible for metabolic functions required to derive nutrients from the food we consume (Ley *et al.,* 2008; Marchesi *et al.,* 2016; Sender et al., 2016).

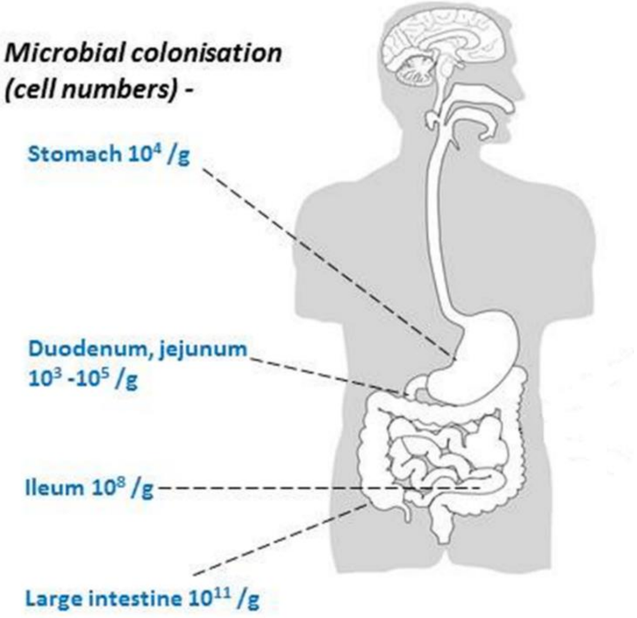


Figure 1.1: A representative model of human gut microbial colonisation. Cell numbers are displayed. Large intestine harbours the highest population of microorganisms. Figure was adapted from Griffiths *et al.,* 2015.

1.1.1. The acquisition and diversity of the human gut microbiota

During the last decade the rise of low-cost and high-throughput sequencing methods has given rise to numerous projects aimed at investigating the human microbiota of a number of bodily niches, including the gastrointestinal tract. One of these projects is the US based Human Microbiome Project (HMP), which aimed to characterize the genetic and metabolic functions of microbial communities co-existing with the human host. The combined data from the HMP and the European Metagenomics of the Human Intestinal Tract (MetaHIT) studies allowed for hundreds of bacterial species and their functions to be identified, in turn influencing human health and potential risk of developing a disease (Turnbaugh *et al.,* 2007; Marchesi *et al.,* 2016; Thursby *et al.,* 2017).

The human gut microbiota is established shortly after birth and plays a major role in metabolic and immunologic pathways throughout the human life. The human microbiome co-evolves with the host and the disruption of the microbial colonization process was shown to negatively influence the human health by increasing the susceptibility to disease during life (Rodriguez *et al.,* 2015). Culture-based techniques, high-throughput sequencing and metagenomics approaches have identified the composition of the human gut microbiota throughout the different stages of life (Koenig *et al.,* 2011). The exposure to the maternal microbiota allows for the formation of the first infant gut inoculum. The composition of the infant gut microbiota rapidly develops in response to dietary stimuli and stabilises, forming an adult-like microbiota by the age of 3-5 years (**Figure 1.2**) (Rodriguez *et al.,* 2015). Influenced by the environmental factors, such as diet, the composition of the human gut microbiota converges as we age and reaches the peak complexity by adulthood (Ottman *et al.,* 2012).

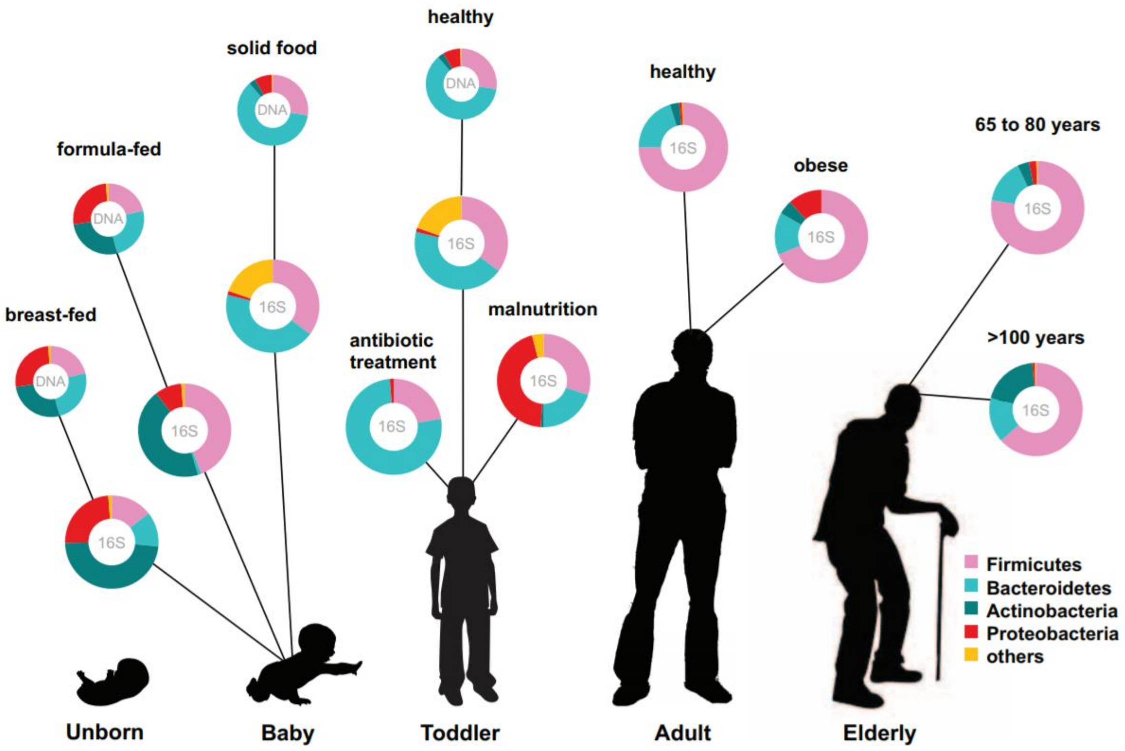


Figure 1.2: The composition of the human gut microbiota throughout life. The schematic gives a global overview of relative abundance of key bacterial phyla present in the human gut through different life stages. The data was obtained using metagenomics approaches (DNA) or 16S RNA sequencing. Figure was adapted from Ottoman *et al.,* 2012.

Despite being home to hundreds of different species of bacteria, two bacteria phyla dominate the healthy human gut – Gram-positive *Firmicutes* and Gram-negative *Bacteroidetes* (**Figure 1.3**), both of which have a significant impact on human health (Arumugam *et al.,* 2011; Ley *et al.,* 2008; Thursby *et al.,* 2017). The benefits of human gut microbiota activity to the host are undeniable; however, there is no clear evidence which microbes are the key species or whether cooperative activity of the microbial community is more important than any individual. It is worth noting that some strains of the same bacterial species can be beneficial, yet others are pathogens. For instance, one strain of *Escherichia coli* is used as a probiotic while the other was shown to cause inflammatory bowel disease (IBD) (Marchesi *et al.,* 2016).

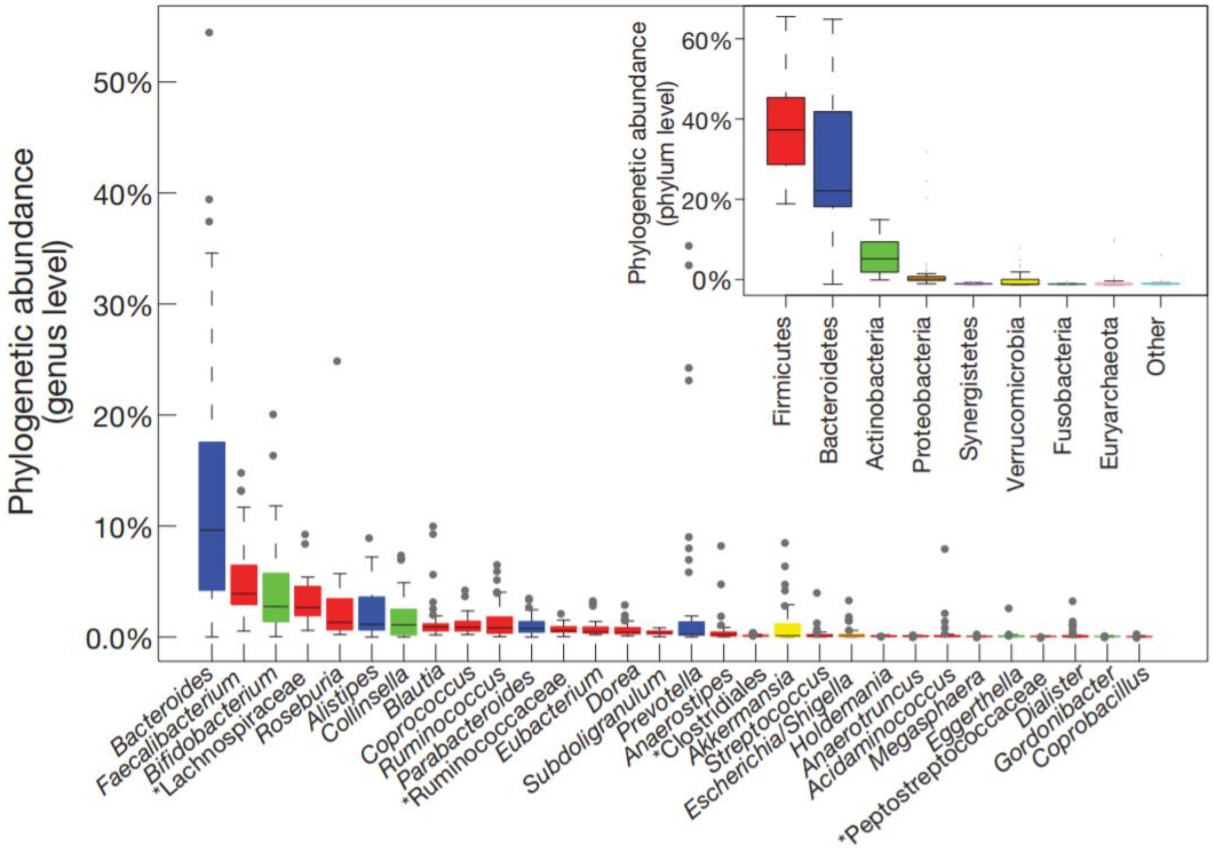


Figure 1.3: Phylogenetic profile of human gut microbiome. The genetic abundance box plot of 30 most abundant genera in human gut microbiota. Samples were collected from 35 individuals. Reference-genome-based mapping with 85% and 65% sequence similarity cut-off was used to calculate the phylum and genus abundances. Figure was adapted from Arumugam *et al.,* 2011.

The bacterial colonisation can differ greatly among the individual people. Comparative studies indicate that diet is a primary environmental factor influencing the inter-individual gut microbiota variation. The so-called Western diet, which includes fast food high in saturated fats and sugar content, results in accumulation of higher numbers of *Firmicutes* and was shown to play a key role in dramatic increase of prevalence of inflammatory intestinal diseases such as Crohn's disease (CD) in 21st century (Agus *et al.,* 2016). It was also observed that type 2 diabetes drug metformin modifies gut microbiome which in turn mediates the effect of the drug (Wu *et al.,* 2017; Koropatkin *et al.,* 2017).

1.1.2. The role of host-microbe interactions in human health and disease

The microbial community that inhabits the human gut is incredibly vast and complex, so it is not surprising that it has a huge impact on human health. The *Bacteroidetes* and the *Firmicutes* are the two phyla of beneficial bacteria that usually dominate the human gut. However, it was shown that the changes in the microbial composition of the GIT can have major consequences to host wellbeing (Ley *et al.,* 2005). Normally, microorganisms residing in the human gut are under constant selective pressure from both the microbe competitors and the host, which usually results in a stable and balanced microbial ecosystem that exists in homeostasis with the human body (Arumugam *et al.,* 2011). Although diversity is predominantly thought to be advantageous for the stability of gut microbiota, some bacterial species occur in higher abundance than others; most likely thanks to the huge repertoire of cellular responses to stress they possess (Backhed *et al.,* 2005; Arumugam *et al.,* 2011). There is also evidence that depletion of a single species from the gut microbiota community can have a negative impact, for example a depletion of the major member of *Firmicutes*, *Faecalibacterium prausnitzii*, has been associated with inflammatory bowel disease (IBD) (Marchesi *et al.,* 2016). It is clear that this continuously fluctuating microbial community diversity can be both beneficial and disadvantageous to the human host.

Modulation of gastrointestinal metabolites is one of the most important functions performed by the gut microbiota. To benefit the host, gut microbiota is usually organized in a cooperative "food web" structure in order to efficiently break down food and produce the metabolites the host requires (Valdes *et al.,* 2018). Generated short chair fatty acids (SCFAs) mediate a variety of important functions, such as regulation of immune system responses and development (**Figure 1.4**) (Blacher *et al.,* 2017). In the human GIT, the immune system plays an important role in protecting the body from invading pathogens and keeping a fine balance by maintaining tolerance to commensal gut microbiota. This is possible because antigen presenting cells (APCs) have co-evolved with microbiota, some developing specific phenotypes rendering them inactive against commensal microbes under homeostatic conditions (Bessman *et al.,* 2016). One of the most important roles of commensal gut microbiota in immune system homeostasis is the induction of CD4+ T cell differentiation, required to regulate the immune response and subsequently control the infection (Wu *et al.,* 2012). Along with the dietary polysaccharide degradation and immune system modulation, the microbes in the human gut are also capable of synthesizing and supplying vitamins required by their host. It was shown that the human gut microbiota is capable of synthesizing B vitamins such as riboflavin, nicotinic acid, biotin, folates, cobalamin and thiamine along with the vitamin K (LeBlanc *et al.,* 2013).

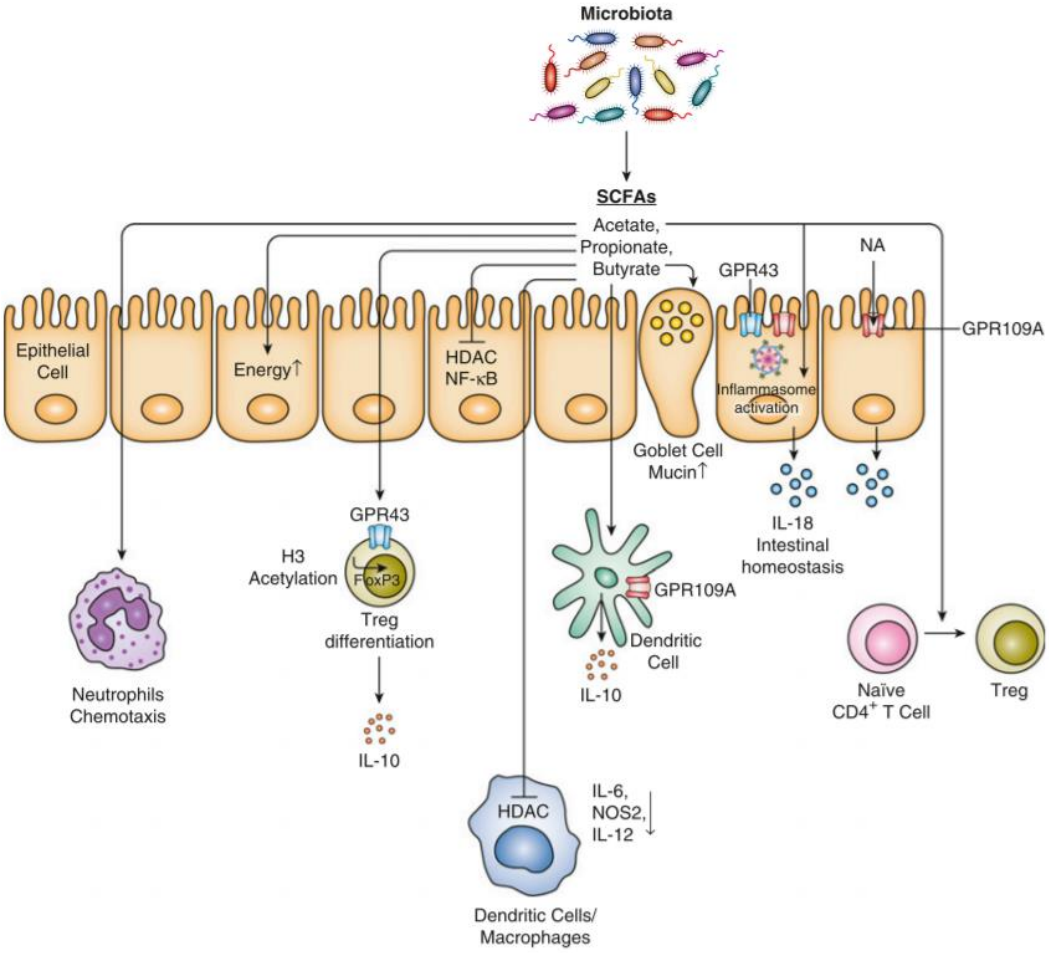


Figure 1.4: The effects of microbiota-associated metabolites to host health. Metabolites, such as acetate, propionate and butyrate, produced by the fermentation of carbohydrates by the microbiota contribute to the host-microbiome network of communications as: 1) an energy source to epithelial cells; 2) in differentiation of naïve T cells to produce regulatory T cells (Treg); 3) to neutrophils chemotaxis; 4) as histone deacetylase (HDAC) inhibitors and to overall intestinal homeostasis. Figure was adapted from Blacher *et al.,* 2017.

Under normal conditions, gut microbiota functions in cooperation with the host, maintaining a mutualistic relationship. However, it has been recognised that even the slightest disruption to the composition of normal gut microbiota can result in imbalance of the microbe-host relationship, which could potentially lead to alterations in the microbiota composition known as dysbiosis (Cerf-Bensussan *et al.,* 2010) (**Figure 1.5**).

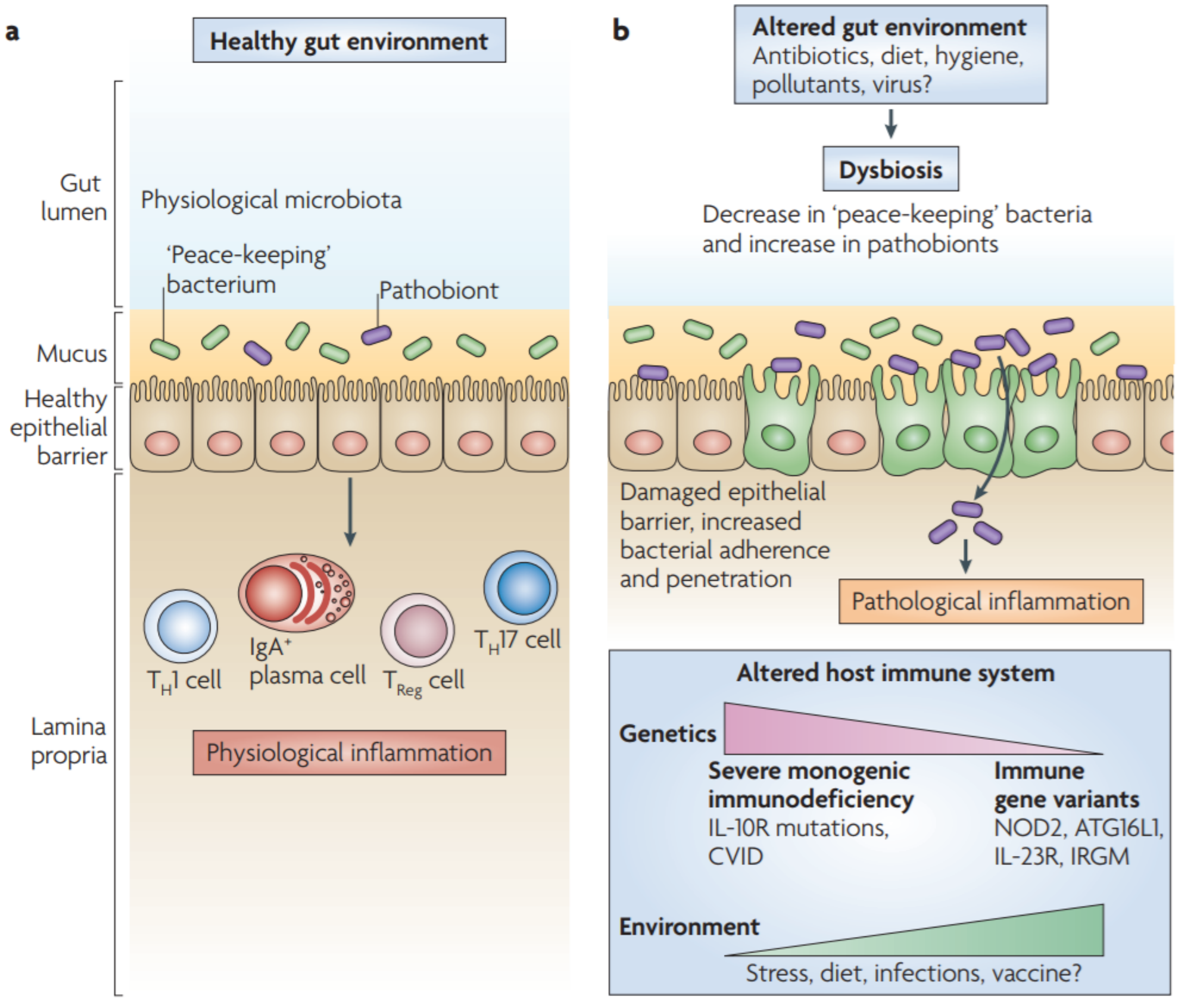

a
Healthy gut environment
Gut lumen
Physiological microbiota
'Peace-keeping' bacterium
Pathobiont
Mucus
Healthy epithelial barrier
Lamina propria
T_H1 cell
IgA^+ plasma cell
T_{Reg} cell
T_H17 cell
Physiological inflammation
b
Altered gut environment
Antibiotics, diet, hygiene, pollutants, virus?
Dysbiosis
Decrease in 'peace-keeping' bacteria and increase in pathobionts
Damaged epithelial barrier, increased bacterial adherence and penetration
Pathological inflammation
Altered host immune system
Genetics
Severe monogenic immunodeficiency
IL-10R mutations, CVID
Immune gene variants
NOD2, ATG16L1, IL-23R, IRGM
Environment
Stress, diet, infections, vaccine?

Figure 1.5: Host-microbiome interaction comparison in healthy and inflamed gut. A) Healthy gut environment is composed of a balanced, physiological microbiota. Efficient immune barrier is maintained and supported by the commensal 'Peace-keeping' bacteria that release anti-inflammatory metabolites. Pathogenic (pathobiont) bacteria numbers are low, thus maintaining intestinal homeostasis. **B)** Altered gut environment leads to dysbiosis. It can be caused by use of antibiotics, changes in diet or hygiene, pollutants and viruses. The numbers of the 'peace-keeping' bacteria decrease and the numbers of pathobionts increase. Due to damaged epithelial barrier, bacterial adherence and penetration increases ultimately leading to pathological inflammation. Immunodeficient patients are highly susceptible to dysbiosis. Figure was adapted from Cerf-Bensussan *et al.,* 2010.

Currently the majority of host–microbiome relationship research is focused on generating data derived from faecal samples. This technique is limited because it can only detect culturable intestinal microbes. Hence, utilising novel experimental techniques that exploit the genetic machinery, which microbiota have evolved to survive in the gut is of utmost importance in studying the relationship in detail (Marchesi *et al.,* 2016).

1.1.3. The mucus layer of the gastrointestinal (GI) tract

Human epithelial surfaces are lined with a layer of mucus that forms a protective barrier against various environmental factors (Koropatkin *et al.,* 2012). Intestinal mucus is secreted by the goblet cells and contains a mixture of enzymes, peptides, proteoglycans and glycoproteins, such as secretory IgA antibody, which collectively function to limit infection and protect epithelial cells from damage (Li *et al.,* 2015).

Studies on mucus composition of the mammalian gastrointestinal (GI) tract revealed that it is composed of two layers in stomach and colon whereas only one layer covers the small intestine. The inner, thinner layer is secreted by the goblet cells and is firmly attached to the epithelium whereas the upper, much thicker loosely attached layer is home to numerous gut microbes (**Figure 1.6**) (Johansson *et al.,* 2011). This thick inner layer provides a physical barrier against invading pathogens while the looser outer layer provides lubrication to allow for a smooth passage for digested food (Li *et al.,* 2015).

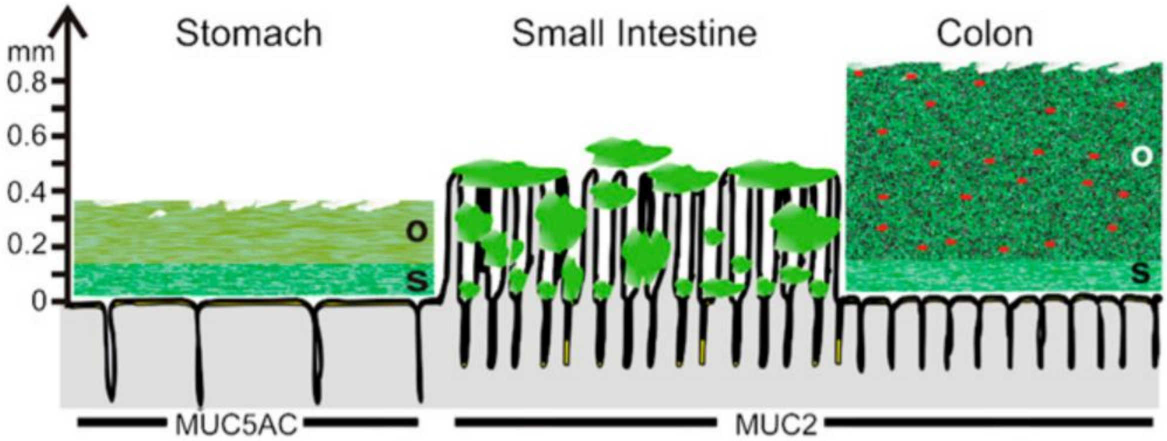


Figure 1.6: Schematic representation of how mucus layers are composed in the gut. The mucus layer is composed of outer (O) and inner, stratified (S) layers that differ in thickness and composition in stomach, small intestine and colon. Mucins forming the gel-like mucus layer in stomach are encoded by MUC5AC gene whereas MUC2 encodes mucins in small intestine and colon. The thick outer mucus layer in colon provides a niche glycan source for specialized gut microbes, such as *B. thetaiotaomicron* (red dots). Figure is taken from Johansson *et al*., 2011.

Several studies have associated thinning of the colonic mucus layer with the reduced availability of dietary fibre (Earle *et al.,* 2015; Hedemann *et al.,* 2009). Some prominent members of the human gut microbiota, such as *B. thetaiotaomicron*, were observed to shift from metabolising dietary polysaccharides to utilising mucus glycans (Sonnenburg *et al.,* 2005). Regular consumption of dietary fibre, such as plant-derived glycans, helps to maintain healthy gut microbiota composition and prevent erosion of the colonic mucus layers. A study of a synthetic gut microbiota assembled in germ-free mice using 14 commensal members of human gut microbiota and a pathogen observed that a low-fibre diet promotes the growth of colonic mucin-utilising bacteria (**Figure 1.7**) (Desai *et al.,* 2016). The dietary fibre deprivation changes the physiological composition of a gut microbiota and promotes the growth of mucus-degrading bacteria, thus thinning the mucus layer and promoting greater epithelial access that ultimately increases the risk of infection by enteric pathogens (Makki *et al.,* 2018; Desai *et al.,* 2016).

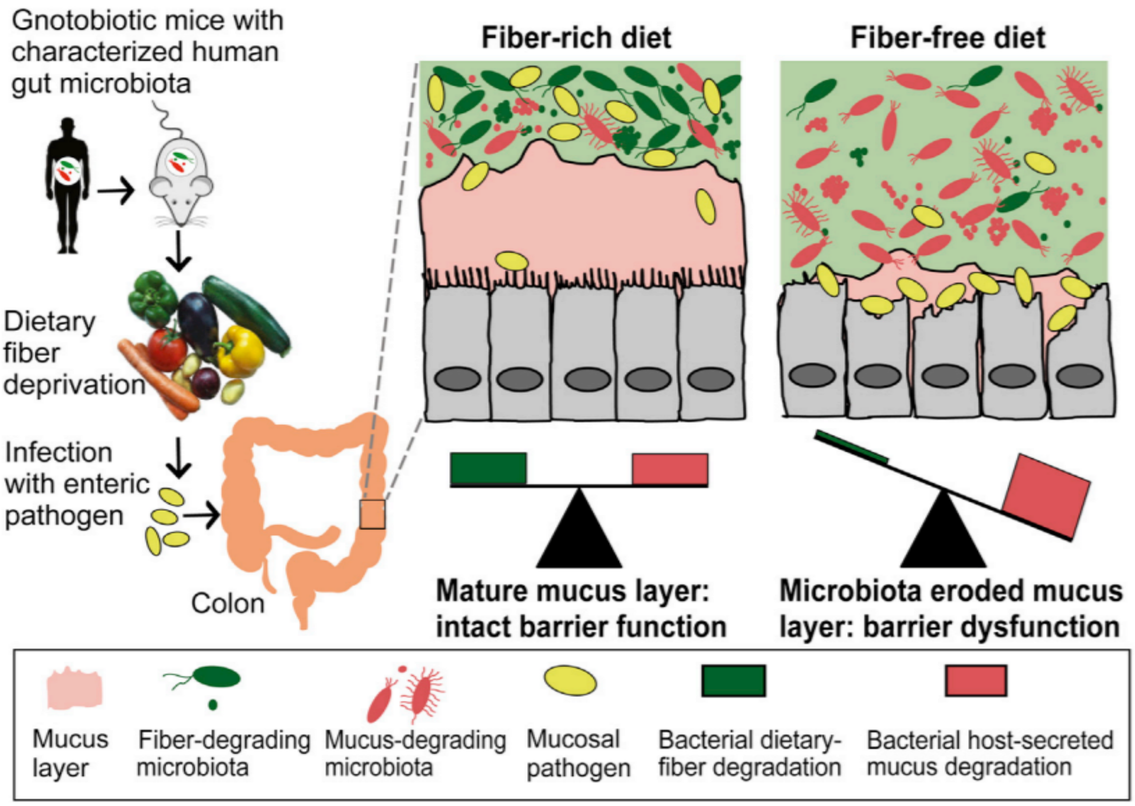

Gnotobiotic mice with characterized human gut microbiota
Dietary fiber deprivation
Infection with enteric pathogen
Colon
Fiber-rich diet
Fiber-free diet
Mature mucus layer: intact barrier function
Microbiota eroded mucus layer: barrier dysfunction
Mucus layer
Fiber-degrading microbiota
Mucus-degrading microbiota
Mucosal pathogen
Bacterial dietary-fiber degradation
Bacterial host-secreted mucus degradation

Figure 1.7: Fibre-deprived gut microbiota degrades colonic mucus layer and promotes enteric pathogen infection. The diagram shows the impact fibre-rich and fibre-free diet has on the thickness of the mucus layer in mice. Fibre-free diet leads to growth of mucin-utilising bacteria, which in turn thins the mucus layer and reduces resistance to infection by mucosal pathogens (Desai *et al*. 2016).

1.2. Nutrition Sources for The Human Gut Microbiota

Fermentable polysaccharides are the main energy source for mammalian hosts and their resident gut microbiota. Numerous plant-associated polysaccharides are commonly found in the human diet, such as fructans, starches, cellulose, hemicelluloses and pectins. Most of these plant glycans cannot be digested by the host and degradation of these structures relies entirely on cooperative activity of intestinal microorganisms. The gut microbiota has a vast repertoire of metabolic machinery enabling them to utilize various dietary and host-derived glycans, including plant polysaccharides, milk oligosaccharides, mucin O-Glycans and N-glycans (**Figure 1.8**) (Koropatkin *et al.,* 2012; Sonnenburg *et al.,* 2005).

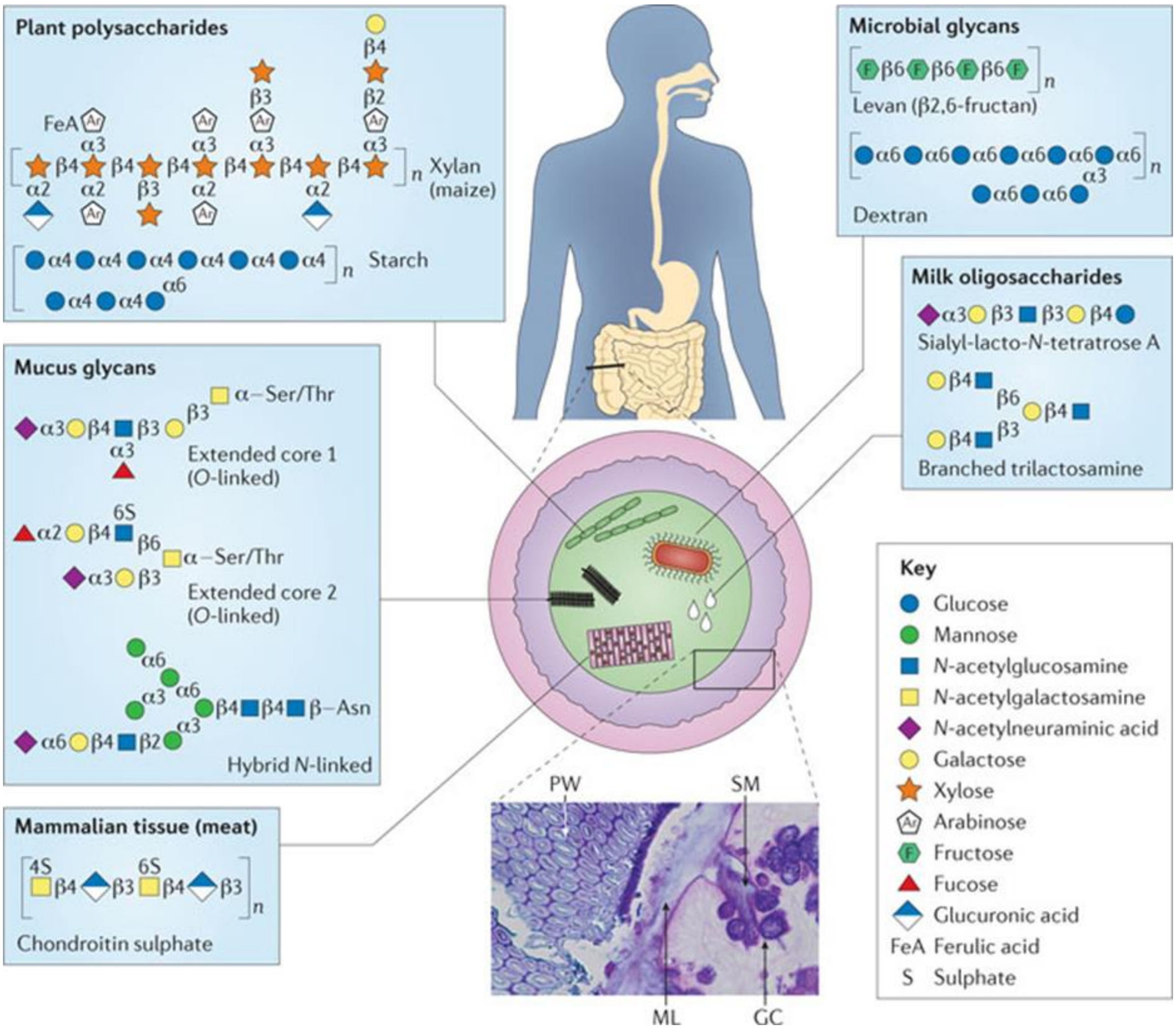


Figure 1.8: Nutrition sources of human gut microbiota. The schematic illustration of the sources and chemical composition of glycans available in the gut. The centre illustrates an intestine and five sources of glycans: plant polysaccharides, mucus glycans (O- and N-linked), milk oligosaccharides, endogenous microbe glycans and dietary mammalian tissue glycans. The structures shown can be much more complex and diverse in nature. The lower panel shows a stained section of germ-free colon highlighting fragments of plant cell wall (PW), host mucus-secreting goblet cells (GC), mucus layer (ML) and secreted mucus (SM). Figure was adapted from Koropatkin *et al.,* 2012.

Although humans derive only a small amount of dietary energy (up to ~10%) through the gut microbiota activities, the impact they have on human health is undeniable and can be heavily influenced by the carbohydrate content the intestinal microbial communities are exposed to (Flint *et al.,* 2012). Gut microbiota has adapted and nutritionally-specialized to degrade and utilize specific dietary carbohydrates the host consumes into short-chain fatty acids (SCFAs) and gases that can also be used as an energy source for other specialist gut microorganisms. However, it has been shown that drastic changes in host diet can result in decrease in one species of bacteria and increase in others, better specialized to metabolize new carbohydrate structures, which could have negative effects on health and potentially lead to disease (Sonnenburg *et al.,* 2005, Flint *et al.,* 2012, Marchesi *et al.,* 2016).

1.2.1. Mucins

The majority of biochemical properties of mucus layer come from mucins secreted by the epithelial cells. These glycoproteins are composed of polypeptide chains rich in threonine and serine amino acids that are extensively glycosylated with small, negatively charged chains of glycans (Johansson *et al.,* 2011). The heavy glycosylation and intertwining of these huge thread-like polymers is what gives the mucus bilayer the viscoelastic texture and protective abilities. Mucus glycoproteins are also normally heavily sialylated, making them polyanionic and enabling interactions with positively charged immune system effector molecules (Li *et al.,* 2015). Mucins are either membrane-bound or secreted extracellular glycoproteins that all share similar characteristics. Membrane-bound mucins are biologically important structural building blocks of mucosal glycocalyx and are involved in cell signalling, cell-cell and cell-matrix interactions whereas secreted mucins are essential for protection against pathogens and they can also act as nutrient source to several prominent members of gut microbiota due to their complex glycosylation patterns (**Figure 1.9**) (Tailford *et al.,* 2015; Li *et al.,* 2015).

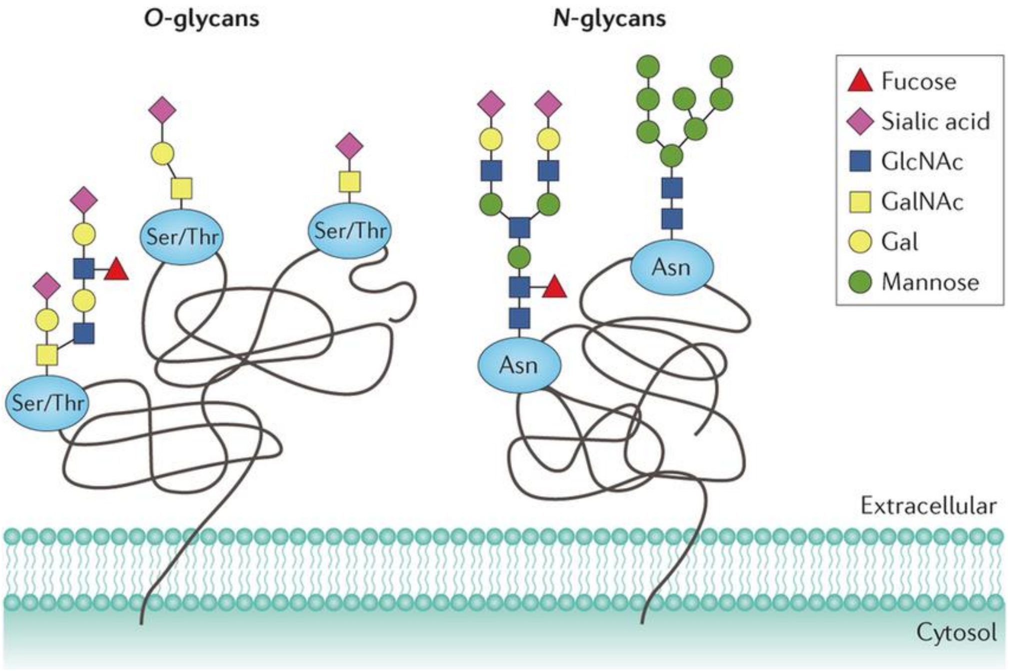
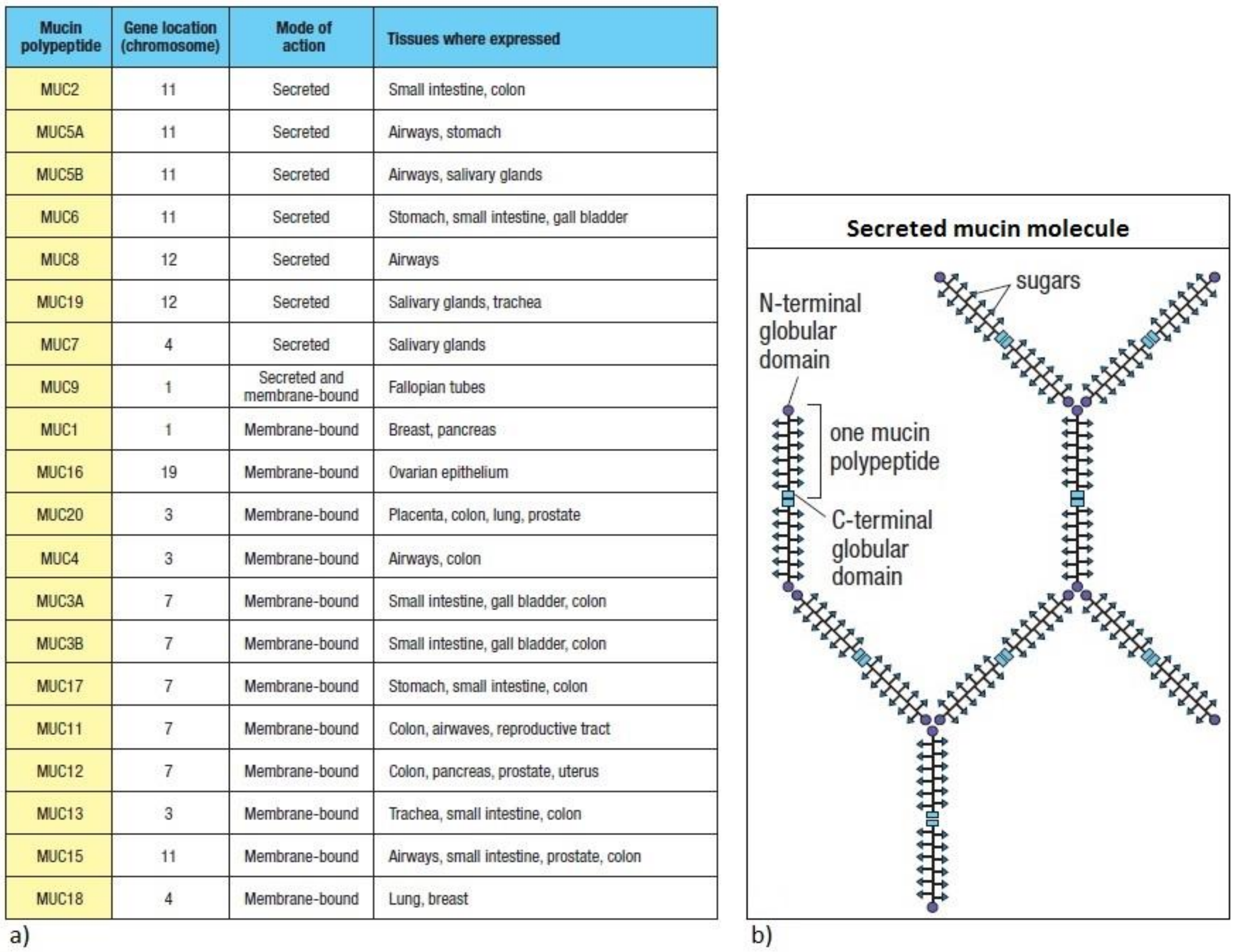


Figure 1.9: Schematic representation of mammalian O- and N-linked glycosylation of mucins. O-glycosylation occurs via serine or threonine (Ser/Thr) residues whereas N-glycosylation occurs through aspargine (Asn). Figure is adapted from Munkley *et al.,* 2016.

There are 21 known genes in the human genome encoding mucins: 7 genes encode secreted mucin glycoproteins and 13 genes encode the membrane-bound mucin polypeptides (**Figure 1.10-A**). The mucin expression profile varies greatly among the host tissues, with majority of genes encoding mucins localised and expressed within the gastrointestinal (GI) tract (Tailford *et al.,* 2015; Parkham *et al.,* 2011). The main characteristic of mucin proteins is a central domain made of proline and threonine rich tandem repeats (**Figure 1.10-B**). MUC2 mucin is the major mucin found in both small and large intestinal mucus. The MUC2 protein is around 550kDa in size and is made of two heavily O-glycosylated PTS domains (Proline/Threonine/Serine) (van der Post *et al.,* 2014).



Mucin polypeptide	Gene location (chromosome)	Mode of action	Tissues where expressed
MUC2	11	Secreted	Small intestine, colon
MUC5A	11	Secreted	Airways, stomach
MUC5B	11	Secreted	Airways, salivary glands
MUC6	11	Secreted	Stomach, small intestine, gall bladder
MUC8	12	Secreted	Airways
MUC19	12	Secreted	Salivary glands, trachea
MUC7	4	Secreted	Salivary glands
MUC9	1	Secreted and membrane-bound	Fallopian tubes
MUC1	1	Membrane-bound	Breast, pancreas
MUC16	19	Membrane-bound	Ovarian epithelium
MUC20	3	Membrane-bound	Placenta, colon, lung, prostate
MUC4	3	Membrane-bound	Airways, colon
MUC3A	7	Membrane-bound	Small intestine, gall bladder, colon
MUC3B	7	Membrane-bound	Small intestine, gall bladder, colon
MUC17	7	Membrane-bound	Stomach, small intestine, colon
MUC11	7	Membrane-bound	Colon, airwaves, reproductive tract
MUC12	7	Membrane-bound	Colon, pancreas, prostate, uterus
MUC13	3	Membrane-bound	Trachea, small intestine, colon
MUC15	11	Membrane-bound	Airways, small intestine, prostate, colon
MUC18	4	Membrane-bound	Lung, breast

Figure 1.10: A list of currently known mucin-encoding human genes and a typical structural organization of a secreted mucin. a) The list of secreted and membrane-bound mucins encoded by human genome. **b)** A structure of a characteristic mucin polymer. Long, heavily glycosylated polypeptide chains are cross-linked to form large polymeric structures that give mucus its characteristic properties. Figure is adapted from Parkham *et al.,* 2011.

1.2.2. O-linked glycans

Mucin function is heavily influenced by their glycosylation pattern. It contributes to protein conformation, developmental processes and regulation of protein activities. Numerous studies have concluded that human glycoproteins can have two types of glycosylation patterns: O-glycosylation and N-glycosylation (Staudacher *et al.,* 2015). O-glycosylation of mucins is the most common and one of the most important post-translational modifications. It was shown that alteration of N- and O-glycosylation patterns in cancer has detrimental effects to protein function (Munkley *et al.,* 2016). These tumor-associated alterations in glycosylation patterns can result in changes in glycoprotein conformation, cell turnover and oligomerization. Most notable altered O-glycans include Tn antigens, sialylated Tn antigens and sialylated Lewis antigens that have been associated with altered cell signalling pathways and metastasis (Stowell et al., 2015). Interestingly, an oral pathogen *Fusobacterium nucleatum* was shown to enable immune evasion and accelerate colorectral adenocarcinoma (CRC) tumor progression by recognizing and binding Gal-GalNac moieties of O-glycans, produced by tumor-associated cells in abnormally high levels (Rubinstein *et al*., 2013; Abed *et al.,* 2017).

O-glycan structures of mammalian mucus glycoproteins are very complex and diverse, consisting mainly of alpha- and beta-linked GlcNAc (N-acetylglucosamine), GalNAc (N-acetylgalactosamine) or Galactose sugar residues covalently attached to the serine (Ser) or threonine (Thr) amino acids (**Figure 1.11**) (Jensen *et al.,* 2010). The biosynthesis of an O-glycan structure begins by the addition of GalNAc residue to the protein backbone, followed by an assembly of other monosaccharides, forming various structures and linkages. The core 1-4 mucin-type structures are further elongated and often modified by addition of sialic acid or fucose residues. Core 1 (Gal-β1,3-GalNAc-α1-Ser/Thr) and core 2 (Gal-β1,3-(GlcNAc-β1,6)-GalNAc-α1-Ser/Thr) structures are commonly found on duodenal and gastric mucins whereas the core 3 (GlcNAc-β1,3-GalNAc-α1-Ser/Thr) and core 4 (GlcNAc-β1,6-(GlcNAc-β1,3)-GalNAc-α1-Ser/Thr) are predominantly found in colon (Tailford *et al.,* 2015).

Any changes in O-glycosylation patterns of mucins have been linked with a number of debilitating conditions in humans, such as colonic cancer, inflammatory bowel disease, Crohn's disease and ulcerative colitis (Leroy *et al.,* 2006; Theodoratou *et al.,* 2014; Simurina *et al.,* 2018)

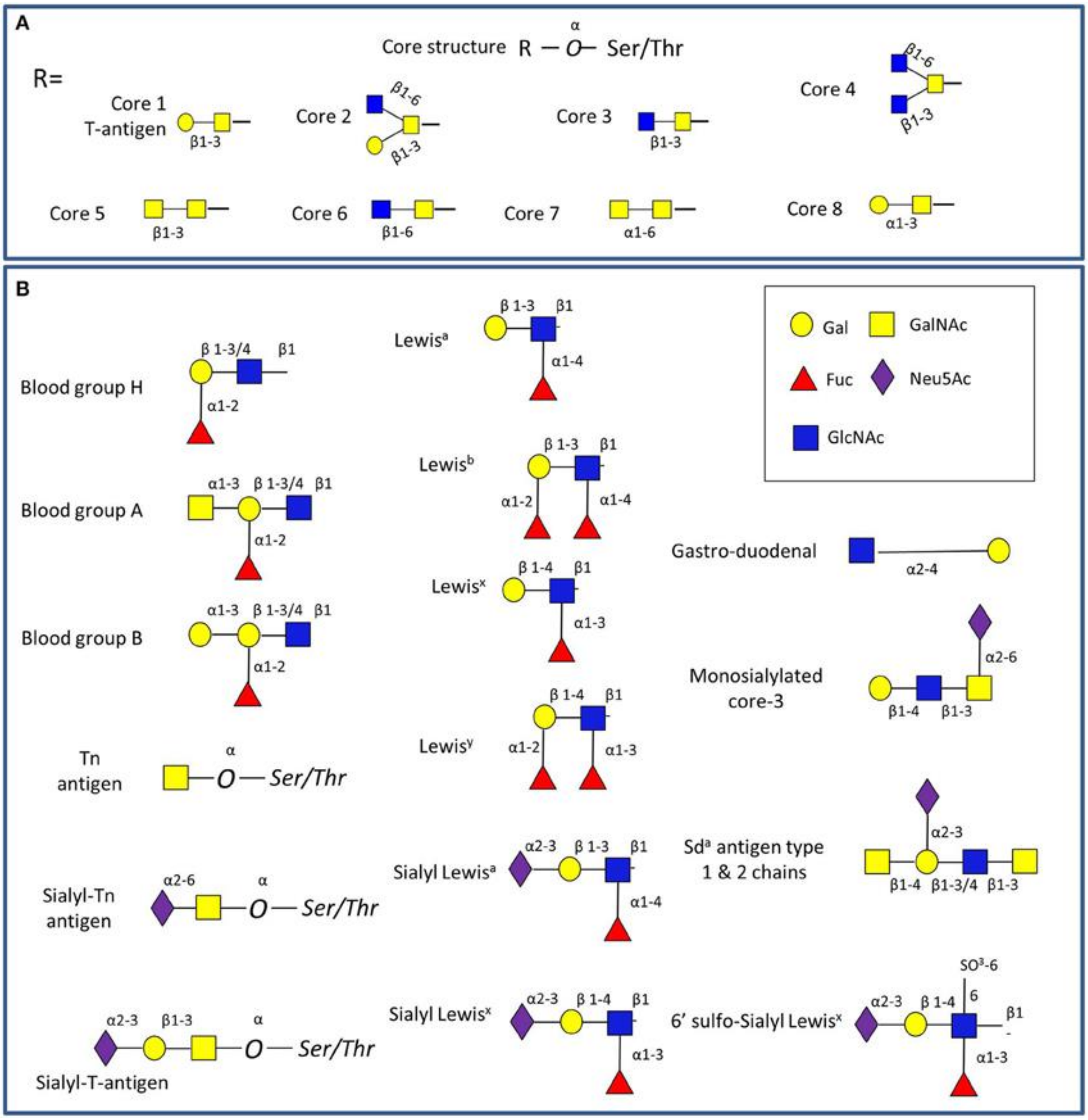


Figure 1.11: The structures of glycans found in the GI tract. A) Main O-glycan core structures found in gastrointestinal (GI) tract. **B)** Common glycan epitopes found in the human gut. Human blood classification system is based on the expression of glycoproteins known as Lewis antigens: Lewis^a and Lewis^b. These structures can be further glycosylated on O-glycans to form Lewis^x, Lewis^y and sialylated forms that are important in cell-to-cell recognition. In adults, overexpression of Lewis^Y has been linked with epithelial cell cancer. Tn antigen is commonly found on gastroduodenal mucins. It is also a tumor-associated antigen that is not normally found in blood cells or peripheral tissues. Figure is adapted from Tailford *et al.,* 2015.

1.2.3. N-linked glycans

N-glycosylation is one of the most important types of post-translational modifications of secreted and membrane-bound glycoproteins; it is involved in the regulation of protein activity, cellular targeting and stability enhancement (Jensen *et al.,* 2010). N-linked glycans are oligosaccharides formed by linking an N-acetylglucosamine (GlcNAc) residue to an amino group of an asparagine (Asn) in a certain Asn-X-Ser/Thr sequon in ER (Staudacher *et al.,* 2015). N-glycosylation is a non-templated process that relies on using a range of specialised enzymes to add or remove monosaccharides from the glycan chains, producing a diverse range of structures. All N-glycans share a common tri-mannosyl chitobiose core, Man-α1,6(Man-α1,3)Man-β1,4GlcNAc-β1,4GlcNAc-β1-Asn, and can be divided into three main classes depending on the capping structures added to this core structure: high mannose, complex and hybrid structures (**Figure 1.12**). High mannose N-glycans contain only α1,2/3/6-linked mannose residues attached to the core structure whereas the typical complex N-glycan structures usually contain β1,2/4/6-linked GlcNAc, β1,4-linked galactose and α2,3/6-linked sialic acid residues, but these structures can vary greatly depending on the type of the glycoprotein they are found on. Hybrid N-glycans contain only mannose residues on one arm whereas the other arm is made of GlcNAc and galactose residues (Bieberich *et al.,* 2014). Complex N-glycan and hybrid structures can be further elongated with a bisecting GlcNAc-β1,4-, where GlcNAc residue is attached to the central mannose residue of the tri-mannosyl core residue (Man-β1,4-GlcNAc). Depending on the function of the glycoprotein, complex N-glycan structures can be incredibly diverse and be further glycosylated with galactose, sialic acid, core and antennary fucose or GalNAc residues (Sethi *et al.,* 2015). Altered expression of branched fucosylated N-glycans has been observed in cancer cells (Stowell et al., 2015).

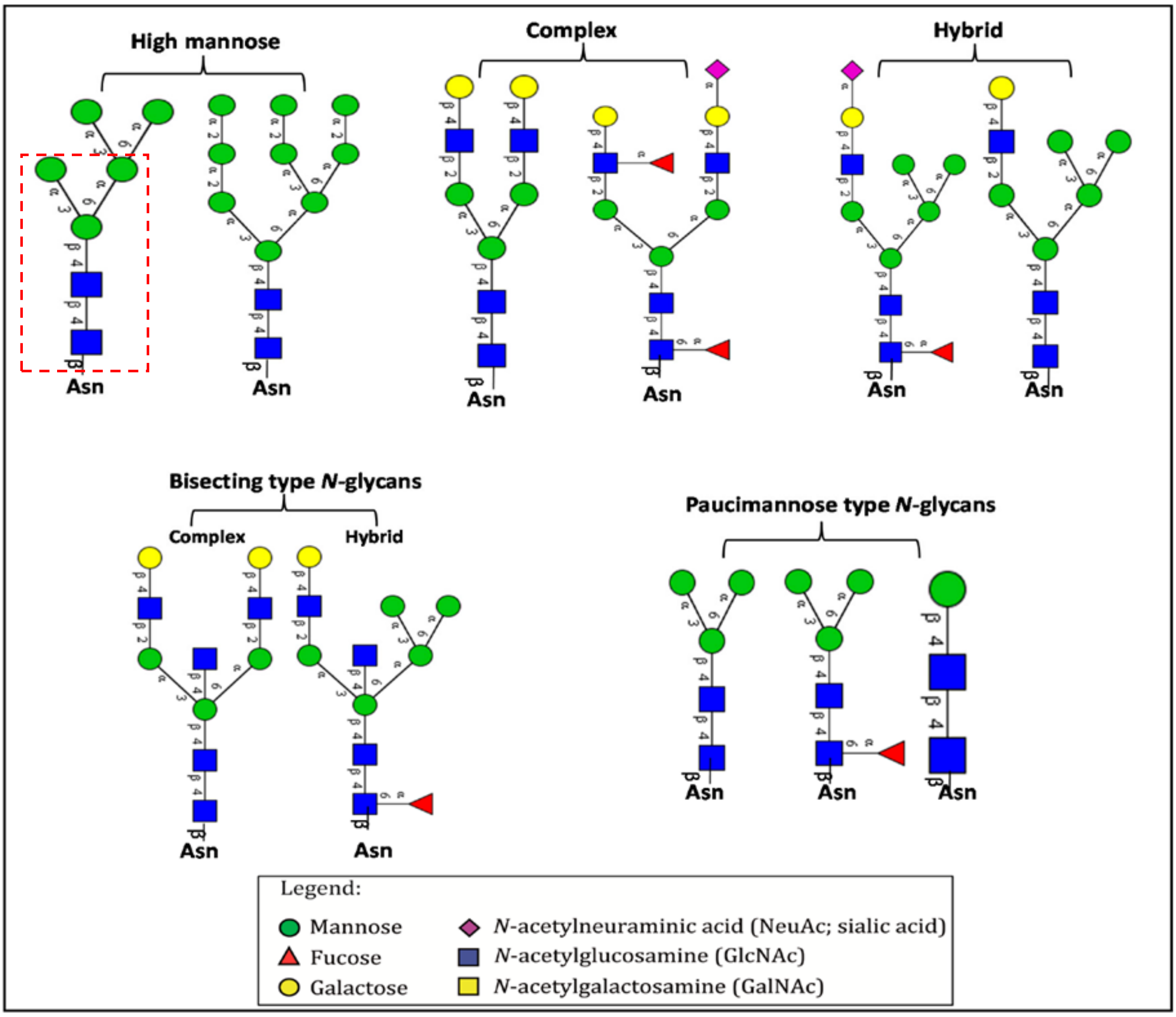

High mannose
Complex
Hybrid
Asn
Bisecting type N-glycans
Complex
Hybrid
Paucimannose type N-glycans
Legend:
● Mannose
▲ Fucose
● Galactose
◆ N-acetylneuraminic acid (NeuAc; sialic acid)
■ N-acetylglucosamine (GlcNAc)
■ N-acetylgalactosamine (GalNAc)

Figure 1.12: Types of N-glycan structures. High mannose, complex and hybrid are the main three N-glycan structures found on N-glycoproteins. They all share a common trimannosyl-chitobiose core (highlighted in red). The unusual paucimannose type N-glycan structures can be found in pathogen-infected humans (e.g. lungs). Figure is adapted from Sethi *et al.,* 2015.

N-glycans are very abundant and can be found on various human, animal, yeast, bacteria, archaea, insect and plant glycoproteins. Depending on the function of the glycoprotein, the N-glycan structures found in these organisms can vary greatly (**Figure 1.13**) (Wang *et al.,* 2017). The complex N-glycan structures are the most common type of N-glycans found in mammalians and plants whereas the high-mannose N-glycans are the most prevalent in yeast (Strasser *et al.,* 2016). Furthermore, human glycoproteins can only be capped with N-acetylneuraminic acid (Neu5Ac) caps whereas animal N-glycans contain both Neu5Ac caps and N-glycolyneuraminic acid (Neu5Gc) caps (Dell *et al.,* 2010).

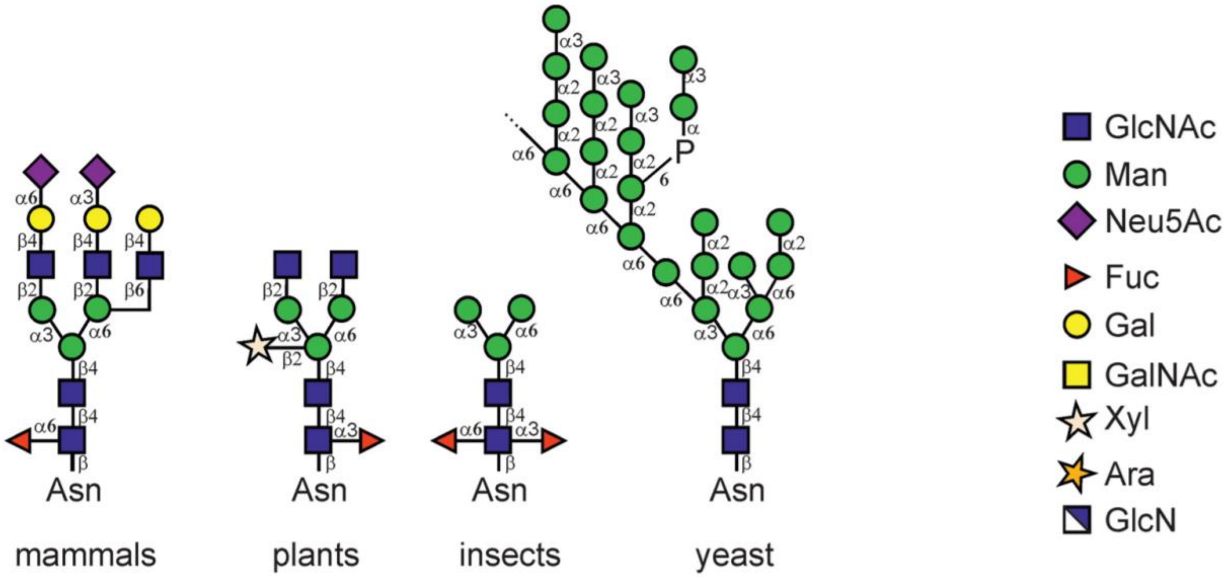


Figure 1.13: The schematic representation and comparison of types of N-glycans produced in different organisms. Common N-glycan structures found in mammals, plants, insects and yeast are shown. Mammalian N-glycan structures are usually complex and capped with terminal sialic acid residues (Neu5Ac) Plant N-glycans typically contain core N-glycan structure capped with GlcNAc residues and further glycosylated with a fucose and xylose. Insects typically produce a double-fucosylated core N-glycan structure whereas yeast typically produce oligomannose N-glycan structures (Strasser *et al.,* 2016).

In the human gut, the microbiota can scavenge on N-glycan structures acquired from a diverse range of sources. These include host-derived N-glycoproteins such as mucins and antibodies (e.g. antibody IgA) (Munkley *et al.,* 2016). There is also an abundance of dietary-derived complex N-glycans from animal (e.g. eggs, milk, meat) and plant sources. These include complex bi-antennary and tetra-antennary N-glycan structures (Valdes *et al.,* 2018). Furthermore, human gut microbiota was shown to be capable of scavenging high-mannose N-glycans present on yeast mannan (Cuskin *et al.,* 2015). Although it appears to be clear that N-glycans are prominent in the human gastrointestinal (GI) tract, the influence of the abundance of N-glycan sources on the proliferation of the symbiotic gut microbiota members is lacking.

1.3. Carbohydrate-Active Enzymes (CAZymes)

Fermentable carbohydrates are the primary energy source for mammalian hosts and their resident gut microbiota. Along with the maintenance of a variety of biosynthetic processes in the body, metabolites derived from the diet are involved in regulation of host-microbe interactions (Cantarel *et al.,* 2009). Carbohydrate structures found in nature are extremely varied, composed of different inter-sugar linkages and complex glycosylation patterns. Numerous separate monosaccharide units can be joined together via various glycosidic linkages to form complex polysaccharides. These can be broadly classified into simple (e.g. glucose, galactose and fructose) and complex (e.g. starch and cellulose) dietary carbohydrates (Bhattacharya *et al.,* 2015). Many of these glycans cannot be digested by the host and degradation of these structures relies entirely on cooperative activity of intestinal microorganisms. To bind, catabolise and utilize these carbohydrates, the gut microbiota produces a huge number of carbohydrate-active enzymes (CAZymes) that, based on their amino acid sequence, are typically assigned to three broad classes: polysaccharide lyases (PLs), carbohydrate esterases (CE) and glycoside hydrolases (GHs). Glycoside hydrolases (GHs) and polysaccharide lyases (PLs) are the most abundantly produced CAZymes by *Bacteroides* species (**Figure 1.14**) (Kaoutari *et al.,* 2013). All currently known carbohydrate-active enzymes and their associated non-catalytic modules are listed in the Carbohydrate Active enzyme (CAZy) database (www.cazy.org).

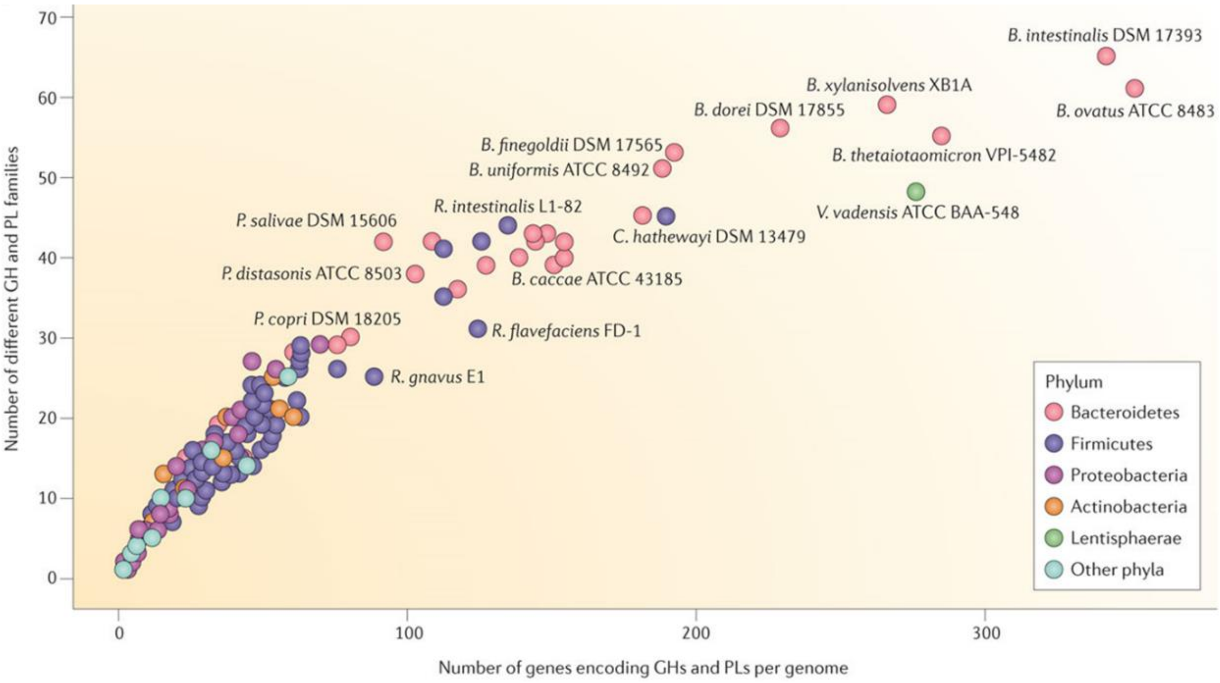


Figure 1.14: Comparison of a number of genes encoding carbohydrate-active enzymes versus a number of different glycoside hydrolases (GH) and polysaccharide lyases (PLs) families encoded by members of human gut microbiota. The Bacteroidetes genomes generally contain significantly more enzyme-coding genes and more different Cazy families than the rest of the phyla present, suggesting they can utilise a wider range of glycan substrates. Figure is taken from Kaoutari *et al.,* 2013.

1.3.1. Glycoside Hydrolases (GHs)

Glycoside hydrolases are the major class of enzymes specialised in cleaving the glycosidic bonds. Based on their amino acid sequence, glycoside hydrolases are classified into different GH families on the CAZy database (Berlemont *et al.,* 2016). To date, there are a total of 153 GH families. The amino acid sequences of predicted carbohydrate-active enzymes enable for the specific GH domains and conserved catalytic residues to be identified and overall mechanistic information to be derived, which allows for the prediction of their evolutionary relationships and basic substrate specificity (Henrissat *et al.,* 1995). For example, presence of the GH20 domains predicts a β-hexosaminidase activity for the enzyme (Liu *et al.,* 2012). The relationship between the amino acid sequence and folding similarities also allows for the prediction of structural features of novel enzymes. However, a full biochemical characterisation is required to fully understand their roles.

Based on the research done using a mini-microbiome, the majority of GHs encoded by the dominant members of the gut microbiota are predicted to be involved in breakdown of plant glycans such as starch, fructans and plant cell wall polysaccharides, with only a small proportion involved in animal glycan degradation (**Figure 1.15**).

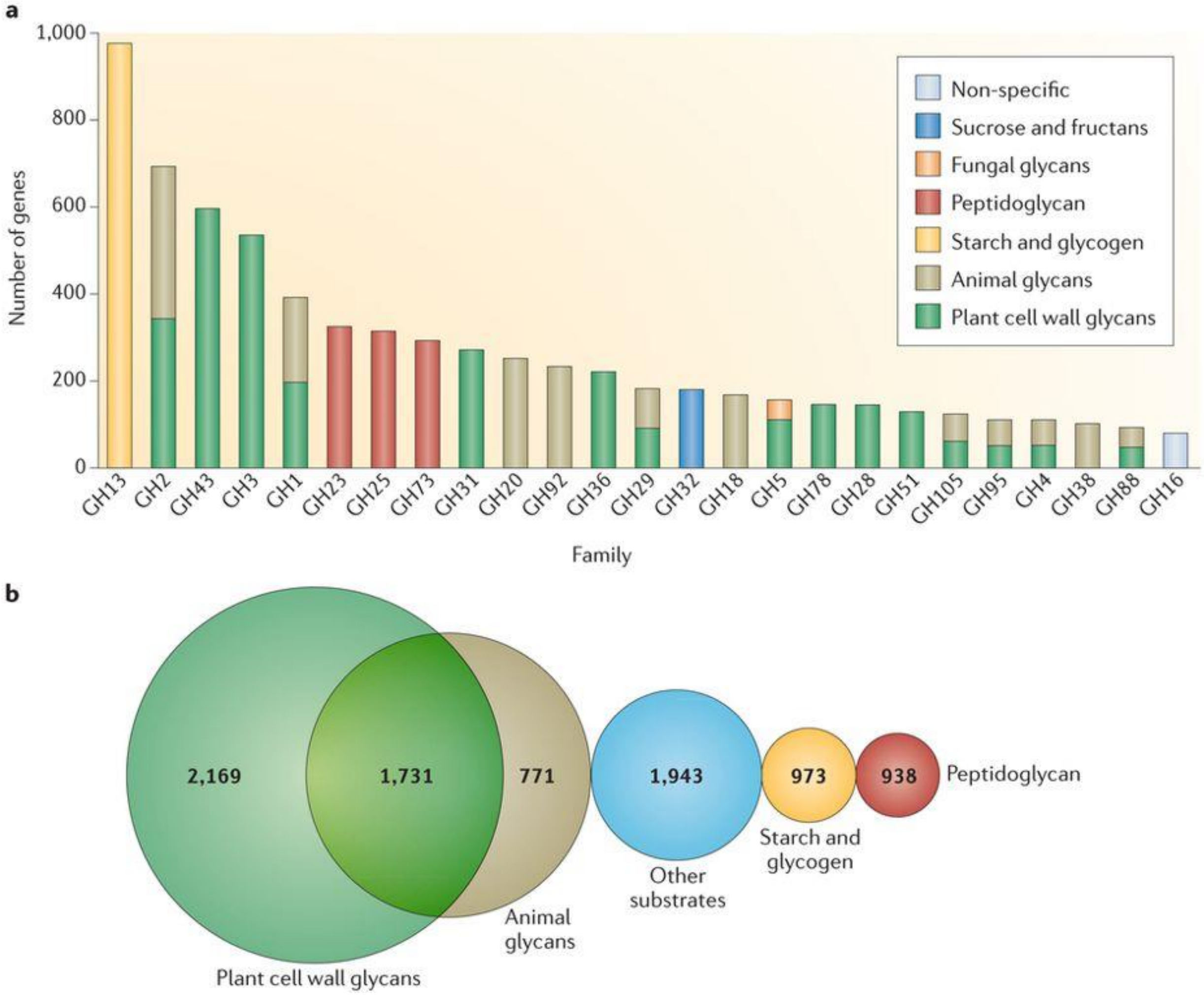


Figure 1.15: Variation of the glycoside hydrolase (GH) families in mini-microbiome.
a) A graph showing the most abundant GH families. The enzymes from GH18, GH20, GH38 and GH92 families are only involved in the breakdown of screened animal glycans. **b)** Venn diagram comparison of substrate specificities of glycoside hydrolases (GHs) and polysaccharide lyases (PLs). Mini-microbiome refers to genes encoded by the dominant members of human gut microbiota, such as Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria. Out of these, Bacteroidetes encode the largest and most diverse number of glycoside hydrolases (GHs) and polysaccharide lyases (PLs). Figure is adapted from Kaoutari *et al.,* 2013.

1.3.1.1.Catalytic mechanisms of glycoside hydrolases

In most cases, the classification system of the glycoside hydrolase (GH) families enables for a relatively easy prediction of catalytic residues and catalytic mechanisms because they are usually conserved in majority of characterised GH enzymes (Henrissat *et al.,* 1995). Depending on the change in the configuration of the anomeric oxygen during the reaction, Koshland *et al.* (1953) grouped GH families into two categories: retaining and inverting. A classic retaining glycoside hydrolase employs a two-step action and requires a nucleophile and a catalytic acid/base catalyst to form a covalent intermediate whereas an inverting glycosidase needs both a catalytic base and a catalytic acid residue (**Figure 1.16**). In both cases, the mechanism of action is heavily dependent on the catalytic residues (Vuong *et al.,* 2010).

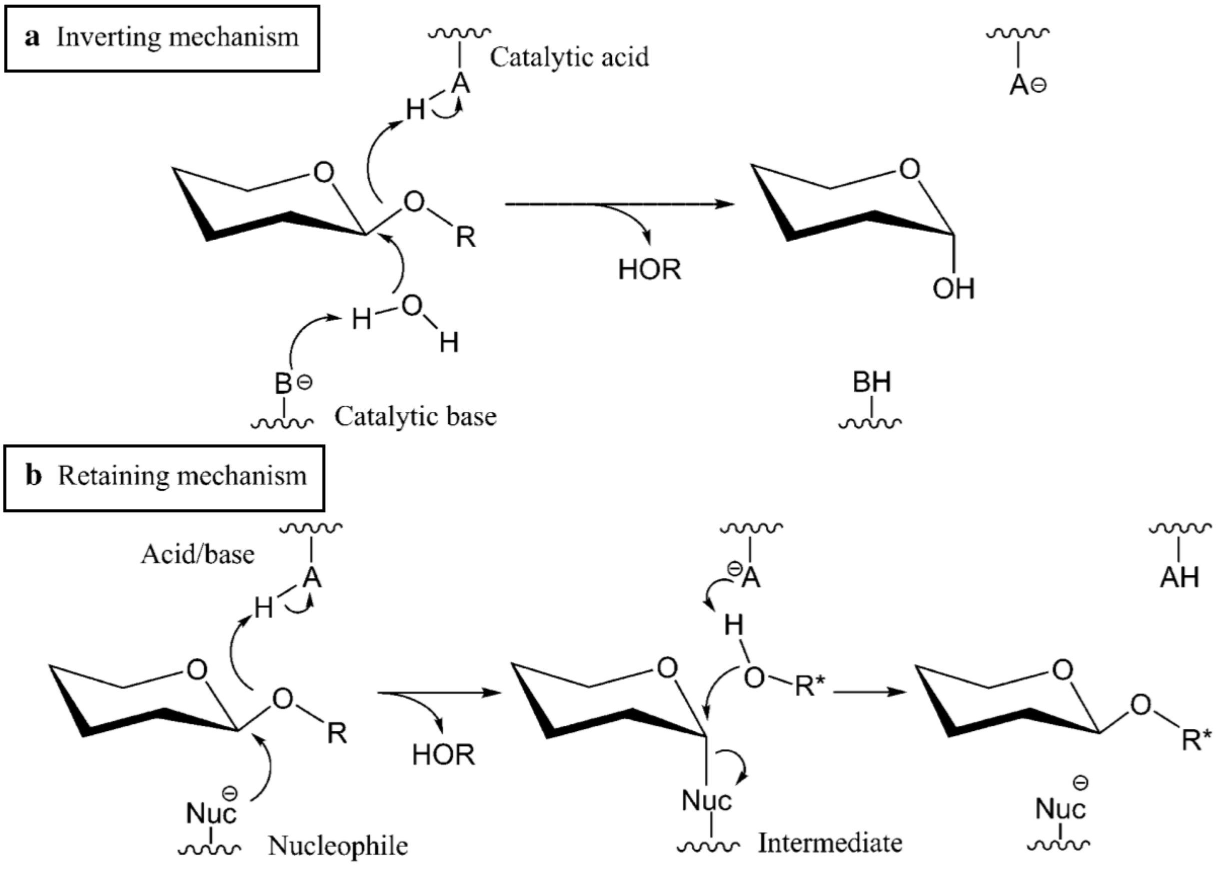


Figure 1.16: A schematic diagram of main two mechanisms of glycoside hydrolases. B$^-$: a catalytic base residue; AH: a catalytic acid residue; Nuc: a nucleophile; HOR: an exogenous nucleophile (e.g. H_2O) and R: carbohydrate derivative. Figure is taken from Vuong *et al.,* 2010.

Some enzymes rely on a substrate-assisted catalysis (SAC) where a functional group of the substrate is required for the catalysis by the enzyme. This type of catalysis was first demonstrated in serine proteases, GTPases and hexose-1-phosphate uridylyl transferases (Dall'Acqua *et al.,* 2000). It depends on the stereochemical, mechanistic and structural properties of the glycosidase. The substrate-assisted catalysis was also observed in GH20 β-hexosaminidase from *S. plicatus SpHEX* (**Figure 1.17**). Here the catalytic mechanism of the β-glycosyl hydrolase depends on the C2-acetamido group present on the substrate (**Figure 1.17-B**). This group replaces enzyme nucleophile residue to form a stable covalent, cyclic intermediate which is then hydrolysed at the anomeric centre in a manner similar to the double displacement mechanism (Mark *et al.,* 2000).

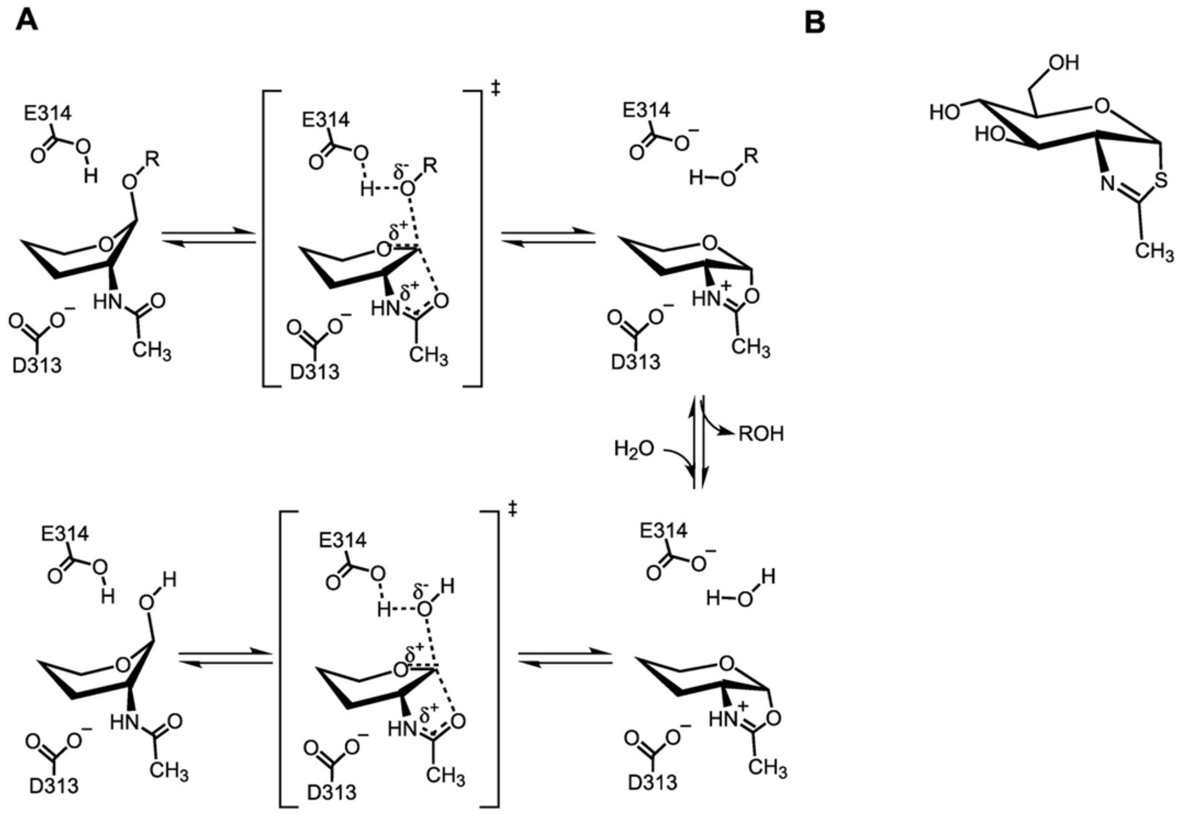


Figure 1.17: **Proposed catalytic mechanism for the *S. plicatus* β-hexosaminidase *SpHEX*. A)** The steps of the substrate-assisted catalysis of the NAG-thiazoline by *SpHEX*. **B)** The cyclic intermediate analogue NAG-thiazoline structure. The predicted general acid/base (Glu^{314}) and the residue (Asp^{313}) are shown. For clarity, C6 and hydroxyl groups were removed from the pyranose ring. Figure was adapted from Mark *et al.,* 2000.

Glycoside hydrolases can be further classified into exo- and endoglycosidases (Kaoutari *et al.,* 2013). Exoglycosidases are enzymes that have an active site pocket topology and release monosaccharide residues from the terminal ends of the glycans whereas the endoglycosidases have an open cleft topology that enables them to target the internal linkages of the sugar chains. Some exo-glycosidases, such as cellobiohydrolases, can have a tunnel topology (**Figure 1.18-A**; Davies and Henrissat, 1995). Endoglycosidases are specialised to cleave N-linked glycans off protein backbones (**Figure 1.18-B**). The examples include the PNGaseF endoglycosidase that cleaves between the core GlcNAc and aspargine residue of the complex, high-mannose and hybrid N-glycans; and GH18-family endo-β-N-acetylglucosaminidases that cleave the chitobiose (GlcNAc-β1,4-GlcNAc) linkages of the N-glycan core structures (Kobata *et al.,* 2013).

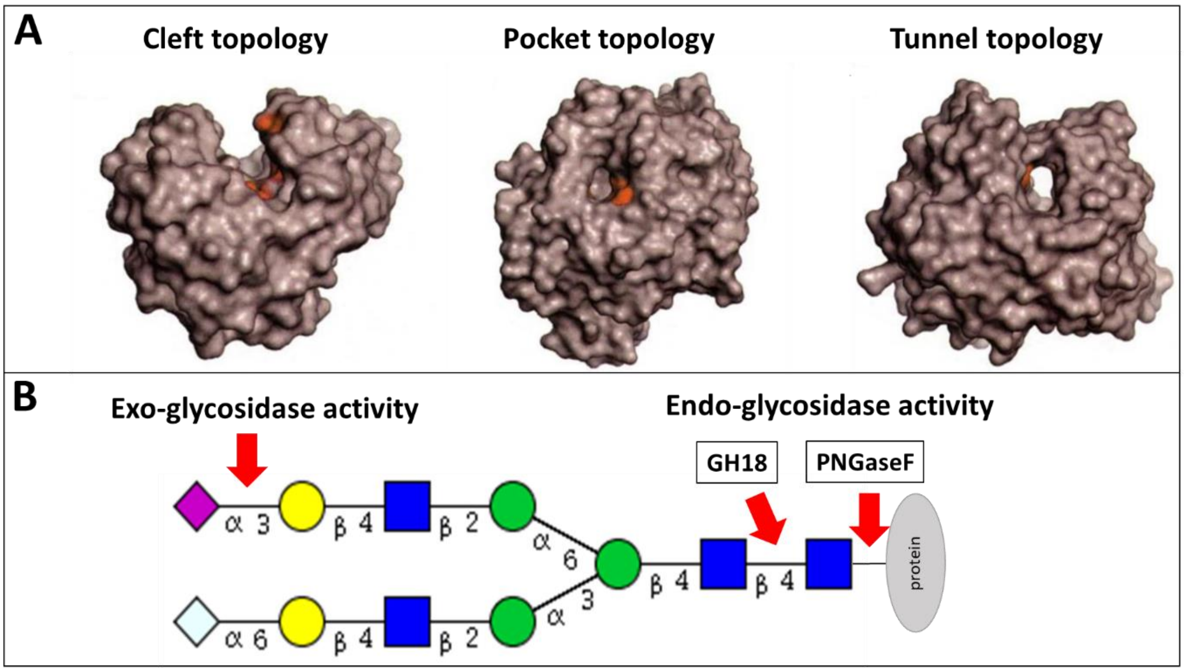


Figure 1.18: The comparison of the exo- and endo-glycosyl hydrolases.
A) The GH active site topology. Cleft, pocket and tunnel topologies are displayed with the active sites highlighted in red. Figure was adapted from Davies and Henrissat, 1995. **B)** Examples of exo-glycosidase and endo-glycosidase cleavage points on a complex bi-antennary N-glycan. Two types of endo-glycosidases are displayed: GH18 and PNGaseF. Figure was made using a GlycanBuilder tool (Expasy).

1.3.1.2. Sub-sites of Glycoside Hydrolases

The substrate binding sites of glycoside hydrolases (GHs) can be divided into sub-sites and numbered. The nomenclature was proposed by Henrissat and Davies (1997) and divides the sub-sites into negative (donor) and positive (acceptor) sub-sites, depending which side of the cleavage point they are located (**Figure 1.19**). Sub-sites on the non-reducing side are denoted by negative (-1, -2, -3) numbers while sub-sites located at the reducing end are positively labelled (+1, +2, +3). The number of sub-sites varies from enzyme to enzyme and depends on the substrate binding pathway (Bissaro *et al.,* 2015).

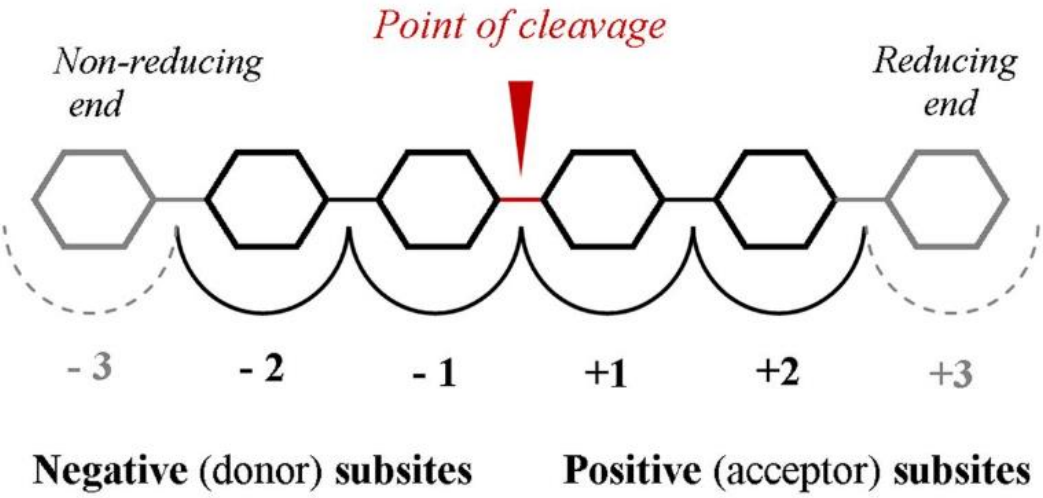


Figure 1.19: Glycoside hydrolase glycan binding sub-sites.
Enzyme point of cleavage is located between -1 and +1 sub-sites and is denoted with a red arrow. Positive (acceptor) sub-sites are located on reducing end whereas negative (donor) sub-sites are located at the non-reducing end. Figure was adapted from Bissaro *et al.,* 2015.

1.3.2. Carbohydrate esterases (CEs)

Carbohydrate esterases are enzymes that catalyse the de-O or de-N-acetylation of ester bonds of glycans. CEs are widely spread in nature – they are produced by both prokaryotes and eukaryotes. Most of CEs are general-acting and based on their substrate specificities are divided into 16 families on the CAZy database. The removal of steric ester-based modifications of glycans by carbohydrate CEs enhances the access of enzymes to these sugar structures, accelerating the glycan degradation. Thus, esterases are widely used in a range of biological and biotechnological applications (Nakamura *et al.,* 2017). The majority of the carbohydrate esterases utilize the Asp-His-Ser catalytic triad (**Figure 1.20**) (Hakulinen *et al.,* 2000). However, some CEs were observed to rely only the His-Ser diad or the Zn$^+$ ion (Adesioye *et al.* 2016).

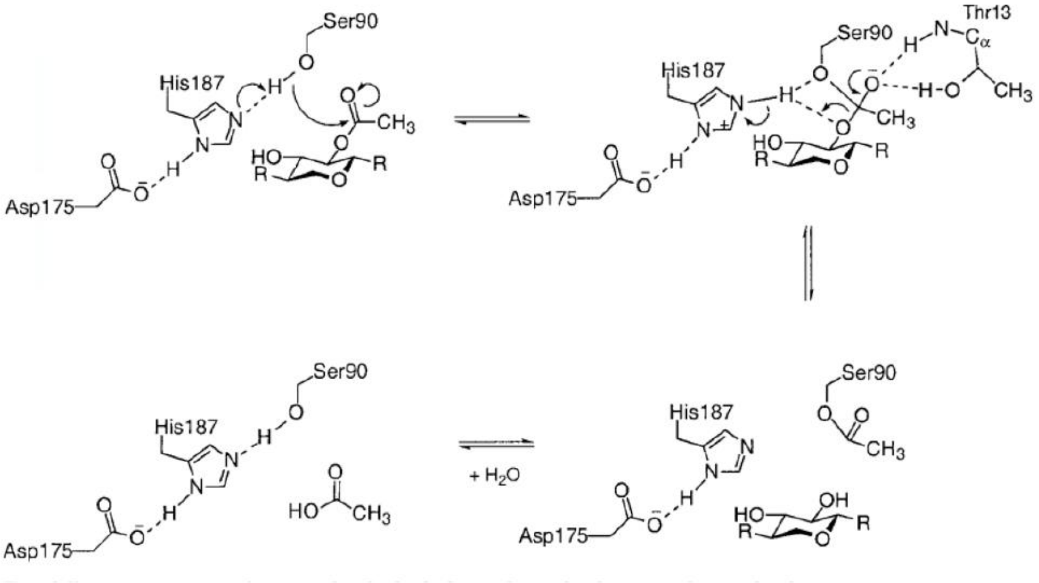


Figure 1.20: The schematic diagram of carbohydrate esterase catalytic mechanism.
The carbohydrate esterase requires Asp-His-Ser triad for its activity. The aspartate (Asp) activates the serine (Ser) residue which results in nucleophilic attack on the acetyl group, resulting in tetrahedral oxyanion transition state. Enzyme oxyanion hole stabilises the transition state while the histidine (His) utilises the water molecule to hydrolyse the cleavage of the acetyl group. Figure is from Hakulinen *et al.,* 2000.

Sialic-acid-specific esterases are of particular importance due to the prevalence of acetylation modification of sialic acids found on glycans produced by mammalian cells. 9-O-acetylation is the most common modification of the sialic acids, such as Neu5Ac. For example, Neu5,9Ac is commonly found in ocular tear film-secreted mucins and human colonic mucins (Corfield *et al.,* 2005). These decorations sterically inhibit the access of most sialidases and thus, require a sialic-acid-specific 9-O-acetylesterase to remove. Several sialic-acid-specific 9-O-acetylesterases are produced by human gut microbiota (Zhu *et al.,* 2004). It was recently observed that the sialic-acid specific O-acetyl esterase EstA from *B. fragilis* enhances the efficiency of liberation of sialic acids from bovine submaxillary mucin (BSM) by *B. fragilis* sialidase NanH and *B. thetaiotaomicron* sialidase NanH (BT0455), thus mediating the cross-species foraging of sialoglycans by the members of gut microbiota (Robinson *et al.,* 2017). Sialic-acid-specific 9-O-acetylesterase, NanS, was also observed to enhance the sialic acid release by the sialidase NanH produced by oral pathogen *Tannerella forsythia* (Phansopa *et al.,* 2015). Interestingly, the increase of production of sialoglycoproteins containing Neu5,9Ac has also been observed in hematopoietic cells from children suffering from acute lymphoblastic leukemia (Ghosh *et al*., 2007).

1.3.3. Carbohydrate binding modules (CBMs)

Many carbohydrate-active enzymes have a complex molecular architecture comprised of multiple discrete modules. This domain composition commonly includes a catalytic domain and a carbohydrate-binding module (CBM). Traditionally, CBMs were associated with plant cell wall-degrading enzymes. However, novel CBMs are increasingly being identified in many host-derived glycan-degrading enzymes. The CBMs are involved in binding of the glycan and anchoring it in the close proximity of the catalytic domain, promoting catalysis (Shoseyov *et al.,* 2006). In enzymes specialised in plant cell wall degradation, the CBM domains are attached to the catalytic domains via a long, flexible linker domain where they function to anchor the enzyme to their target polysaccharide substrates. Meanwhile, the CBM domains present in enzymes specialised in host-derived glycan degradation are usually much more closely associated with the catalytic domain, promoting the efficient degradation of their target glycans (Gilbert *et al*., 2013). Examples of CBMs present in the starch-degrading enzyme SusG from *B. thetaiotaomicron* and mucin glycan-targeting enzyme NagJ from *C. perfringens* are displayed in **Figure 1.21** (Arnal *et al.,* 2018; Ficko-Blean *et al.,* 2009).

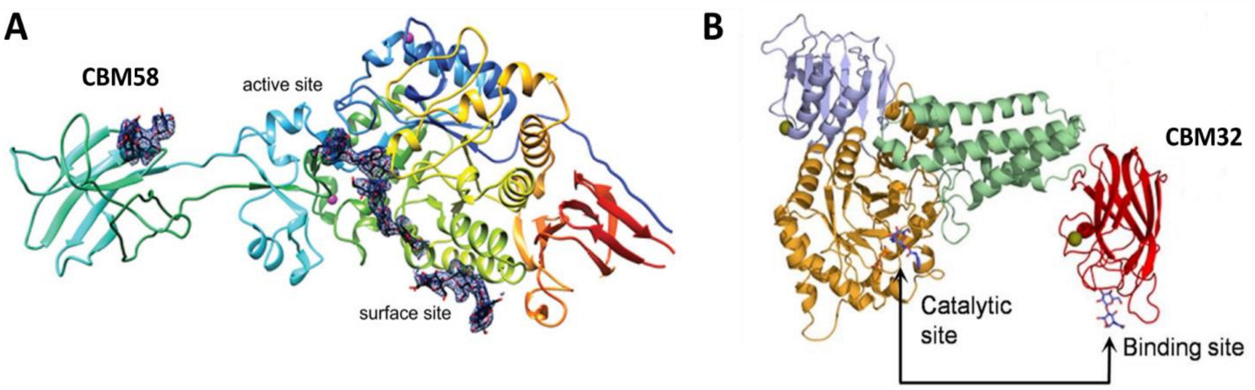


Figure 1.21: Examples of enzymes that contain carbohydrate-binding modules (CBMs).
A) A structure of amylase SusG from *B. thetaiotaomicron* showing the binding of CBM58 domain to pullulan oligosaccharide (black and red sticks). Figure was adapted from Arnal *et al.,* 2018. **B)** A structure of O-GlcNAcase NagJ (GH84C) from *C. perfringens* showing the binding of the CBM32 domain to N-Acetyllactosamine (blue and red sticks), a disaccharide commonly found on mucin glycans. Figure was adapted from Ficko-Blean *et al.,* 2009.

The first carbohydrate-binding module (CBM) described was involved in binding of crystalline cellulose and was initially called cellulose binding domain (CBDs) (Gilkes *et al.,* 1988). Since then, numerous CBMs have been identified and were shown to display a great variation in substrate specificity. Based on their properties and sequence similarities, the carbohydrate-binding modules (CBMs) have been classified into 81 families on the CAZy database (Cantarel *et al.,* 2009). However, because the CAZy classification is based on the sequence similarities, it does not always accurately predict substrate specificity of the CBMs. Some families were shown to be poly-specific, for example CBM4 binds xylan, cellulose and various β-glucans (Abou-Hachem *et al.,* 2000). Based on their mode of ligand recognition and conformation of the ligand and CBM binding site, CBM families can also be subdivided into surface-binding (type A), endo-type (type B) and exo-type (type C) binding CBMs (**Figure 1.22**; Boraston *et al.,* 2004; Gilbert *et al.,* 2013). In type A, the polysaccharide interacts with the hydrophobic surface-binding site of the CBM. In type B, the ligand-binding site of the CBM has a cleft topology that allows for internal accommodation of the glycan chains. In type C, the CBM has an exo- pocket-like binding site (Gilbert *et al.,* 2013).

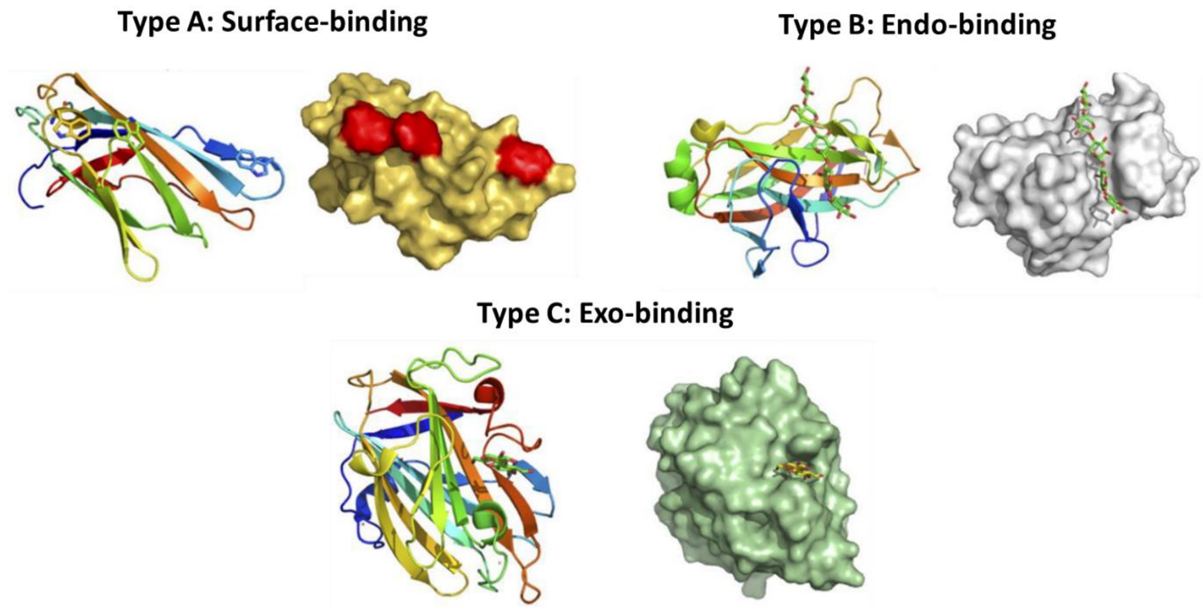


Figure 1.22: The types of carbohydrate-binding modules. Type A shows a surface-binding site (colored in red) of CBM2a from the *Cellulomonas fimi* xylanase (PDB 1XG); **Type B** shows an endo-binding site of a CBM15 from *C. japonicas* xylanase (PDB 1GNY); **Type C** shows an exo-type pocket-binding site of the CBM9 from a *Thermotoga maritima* xylanase (PDB 1I82). The structures are coloured from N-terminal (blue) to C-terminal (red). Ligands are displayed as green/red sticks. Figure was adapted from Gilbert *et al.,* 2013.

It is widely recognised that carbohydrate-binding modules (CBMs) are involved in enhancing the efficiency of the hydrolysis of various host and dietary-derived glycans (Herve *et al.,* 2010). In some instances, the catalytic efficiency of the enzyme is dependent on the CBM domain. For example, the CBM32 domain of the GH5-family mannanase *Ct*Man5A participates in substrate recognition and together with the catalytic domain forms a substrate-binding site required for the catalysis of host-derived mannotetraose by *Clostridium thermocellum* (Mizutani *et al.,* 2014). Furthermore, CBM51 domains of GH95-family α-fucosidase and GH98-family endo-β-galactosidase from *Clostridium perfringens* were shown to be involved in recognition and binding of O- and N-glycan structures, thus enhancing the efficiency of glycan degradation (Gregg *et al.,* 2008). And finally, a novel CBM domain of a NanH sialidase from an oral pathogen *Tannerella forsythia* was shown to bind sialoglycans, where it was observed that the association with the catalytic domain enhanced the CBM binding activity (Frey *et al*., 2018).

1.4. Glycan Degradation by Human Gut Microbiota

Complex glycans are a major nutrient source for the human gut microbiota. The glycosylation patterns of these structures can be extremely complex and varied, composed of different glycosidic linkages and thus, require a cooperative activity of multiple specialised carbohydrate-active enzymes to degrade. It is estimated that the glycobiome of human gut microbiota contains over 9000 genes involved in carbohydrate metabolism, majority of which are encoded by the members of *Firmicutes*, *Bifidobacterium* and *Bacteroides* species (Kaoutari *et al.,* 2013). In these microorganisms, the carbohydrate-active enzymes are usually organised into polysaccharide utilization loci (PULs), each specialized to degrade different polysaccharides, oligosaccharides and glycoconjugates into readily metabolisable monosaccharides. The products of anaerobic sugar fermentation are short-chain fatty acids (SCFAs) such as acetate, butyrate and propionate that play many roles in maintaining human health (Carding *et al.,* 2015).

1.4.1. Glycan utilisation by human gut *Bacteroides*

The ability of commensal gut microbes to persist and adapt to constantly changing environmental conditions in mammalian GI tract largely depends on a vast collection of genes they possess that encode catabolic machinery required to bind and degrade a wide spectrum of dietary and host-derived glycans (Bjursell *et al.,* 2006). The Gram-negative *Bacteroides* are one of the dominant species in the human gut primarily because of their ability to utilize a vast number of glycans. Based on the number and variety of CAZymes encoded by the members of *Bacteroide*s species, they are classified as generalists that drive microbial evolution and dispersion (Sriswasdi *et al.,* 2017). Furthermore, genetic manipulability of *B. thetaiotaomicron* and *B. ovatus* makes them ideal models for studying glycan utilization pathways of *Bacteroides*. It was found that all members of Bacteroidetes encode homologous glycan-specific outer membrane transporter systems (SusC/Ds) so it was suggested that glycan utilisation model is similar in all members of this phylum (Bolam *et al.,* 2012).

Combined comprehensive metagenomic analyses and *in vivo* experiments show that *B. thetaiotaomicron* is a prominent human gut symbiont that possesses hundreds of genes required for dietary and host-derived glycan utilization (Koropatkin *et al.,* 2012). The sequencing of the *B. thetaiotaomicron* genome has allowed for the identification of 260 predicted glycoside hydrolases (GHs), an incredibly huge number for a single organism (Xu *et al.,* 2003). Since then, only *B. ovatus* and *A. cellulolyticus* were found to encode a higher number of GHs (Wegmann *et al.,* 2016; Barabote *et al.,* 2009). In overall, members of *Bacteroides* species encode a much larger number of CAZymes compared to the other residents of human gut (Kaoutari *et al.,* 2013).

In *Bacteroides,* these enzymes are usually found in Sus-like complexes specialized to target and degrade specific dietary polysaccharides and host-derived glycans, such as mucins (Martens *et al.,* 2008). The starch-utilization system (Sus) was the first multi-protein carbohydrate degradation system described in *B. thetaiotaomicron*. It encodes eight cell-envelope associated proteins SusABCDEFGR: starch degrading enzymes Sus A, B, and G, TonB-dependent transporter SusC, three starch-binding proteins SusD, E and F, and sensor/regulator SusR. Together these proteins are working to bind, degrade and import starch molecules into the cell (**Figure 1.23**) (Anderson and Salyers, 1989; Shipman *et al.*, 1999, Koropatkin *et al.,* 2012).

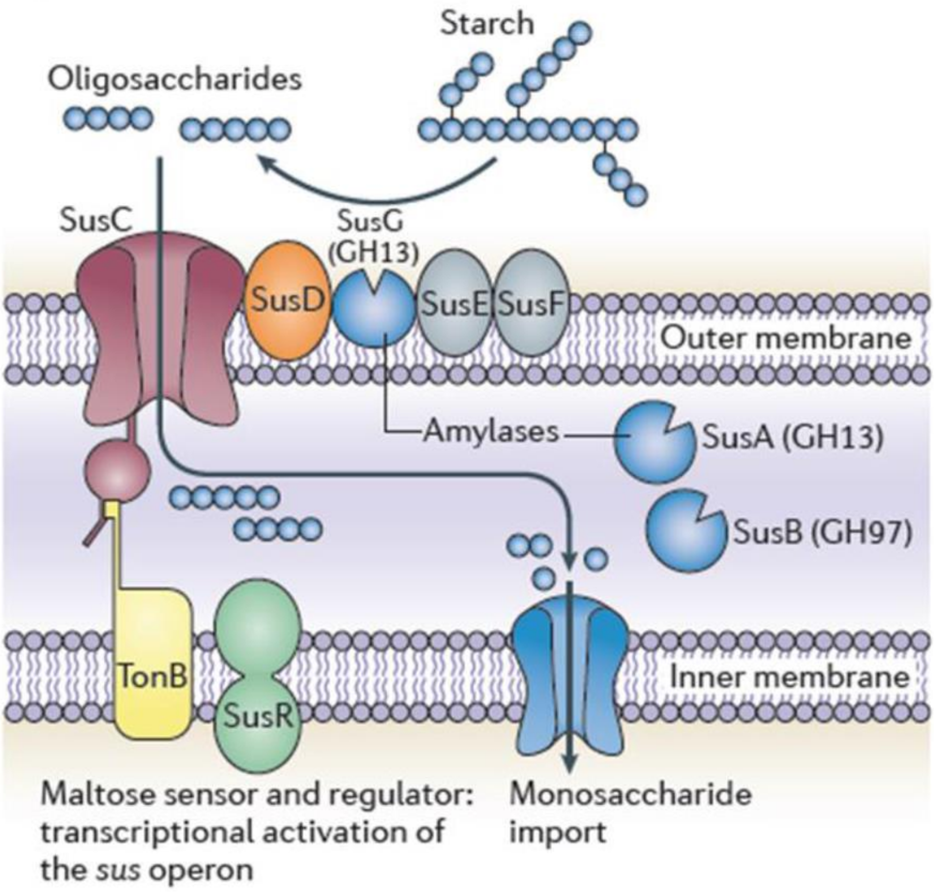


Figure 1.23: Schematic representation of a starch utilization system (Sus) in *B. thetaiotaomicron.* Starch utilization is initiated by binding to extracellular E and F proteins and initial degradation by SusG α-amylase (GH13). The oligosaccharides are then imported into the periplasm by TonB-dependent SusC/D complex where they are further degraded by SusA neopullulanase (GH13) and SusB α-glucosidase (GH97). Maltose is used as transcription regulator for the periplasmic sensor SusR whereas glucose is imported into the cell for fermentation. GH – glycosyl hydrolase. Figure is adapted from Koropatkin *et al.,* 2012.

Clusters of genes that encode Sus-like systems are termed polysaccharide utilization loci (PULs) and are characteristically defined by the presence of homologous TonB-dependent outer membrane transporter (SusC) and a glycan-binding protein (SusD) pair along with numerous enzymes targeting various dietary and host-derived glycan structures (Koropatkin *et al.,* 2009). A ‘pedal-bin’ mechanism of substrate binding and transport has been recently proposed for the SusC/D complexes where the SusD acts like a ‘lid’, moving away from the SusC in the absence of a substrate which leaves the substrate-binding site open and exposed to the extracellular environment (Glenwright *et al.,* 2017). So far, 88 polysaccharide utilization loci (PULs), comprised of 866 genes, have been identified in *B. thetaiotaomicron* (Martens *et al.,* 2011). These PULs are listed on PUL Database (PULDB), an automated prediction tool designed by CAZy team that lists over 4000 predicted PULs from approximately 70 Bacteroidetes, such as *Bacteroides, Alistipes, Prevotella, Tannerella* and *Parabecteroides*. Interestingly, some of the PULs listed on the PULDB lack SusC/D pair but encode them elsewhere in the genome (Grondin *et al.,* 2017). The overall key feature of these diverse cell envelope-associated systems is the coordinated, step-wise activity of multiple proteins that bind and digest complex extracellular carbohydrates into simple metabolites that can be easily absorbed by the host (Martens *et al.,* 2009). The PUL activity is regulated by either metal-dependent regulators, extra-cytoplasmic function (ECF) σ-factor/anti-σ-factor or most commonly hybrid two-component system (HTCS). These regulators are PUL-specific and allow bacteria to sense and target carbohydrates they are designed to degrade. SusE/F-like proteins in typical PULs are named surface glycan-binding proteins (SGBPs). Most of the characterised SGBPs are specialised to recognise and bind to specific glycans (Bolam *et al.,* 2012). These TonB-dependent transporter systems are widespread in Gram-negative microorganisms and can transport not only carbohydrates, but also vitamin B_{12} and ferric chelates known as siderophores, such as haemoglobin, serum transferrin and hemin (Noinaj *et al*., 2011).

Transcriptional profiling of *Bacteroides* grown on different plant or host-derived mucosal glycan sources has shown that each Sus-like system targets a unique carbohydrate substrate, sometimes working in collaboration with other PULs (Sonnenburg *et al.,* 2005; Martens *et al.,* 2008; Martens *et al.,* 2011). A highly extensive xylan utilisation apparatus has been described in *B. ovatus* where degradation depends on two discrete PULs – PUL-XylL and PUL-XylS. Combined, these PULs encode three SusC/D-like pairs of proteins, two surface xylan-binding proteins, two HTCS sensors/regulators and numerous xylan-debranching enzymes (Rogowski *et al.,* 2015). A complex degrading apparatus was also identified in *B. thetaiotaomicron* where plant rhamnogalacturonan-II (RG-II) is degraded by a cooperative activity of RGII-binding proteins and degrading enzymes encoded by three discrete PULs – RG-II PUL 1, 2 and 3 (Ndeh *et al.,* 2017). Similar apparatus is employed by *B. thetaiotaomicron* to utilise yeast mannan where degradation is dependent on the activity of enzymes encoded by three discrete PULs – PUL-Man 1, 2 and 3 (Cuskin *et al.,* 2015). *B. thetaiotaomicron* is also capable of utilising host glycosaminoglycans heparan sulfate (HS) and heparin (Hep) (Cartmell *et al.,* 2017). The abundance and diversity of these highly-organized systems of extracellular and periplasmic proteins involved in sensing, regulation, degradation and transport of carbohydrates is what makes *Bacteroides* such highly-competitive members of the mammalian gut microbiota. The comparative analysis of these PULs is an efficient way to understand the community dynamics and nutrient niche colonization of the gut microbiota. For example, *B. ovatus* and *B. thetaiotaomicron* contain over 100 PULs but surprisingly very few of these PULs are homologous, suggesting these dominant members of human gut microbiota have distinct glycan niches (Martens *et al.,* 2011). These findings suggest that these Bacteroidetes have adapted to forage on different nutrient sources.

B. thetaiotaomicron is a generalist capable of degrading numerous distinct types of glycans, including complex O- and N-glycan structures found on mucins. Generalists, specialising in degradation of broad-range of glycans have an evolutionary advantage over specialists, which focus on one or few glycan types and may become extinct if host diet was to change (Pudlo *et al.,* 2015). Plant polysaccharides can be incredibly complex and the task of degrading such structures is further complicated by the presence of plant cell wall, which not many species have a capacity to access. *B. thetaiotaomicron* possess numerous polysaccharide utilization loci (PULs) encoding Sus-like metabolic systems targeting a diverse range of plant glycans, including pectin, hemicellulose, xylan with the notable exception of cellulose (Koropatkin *et al.,* 2012). Previous studies have shown that *B. thetaiotaomicron* is also one of the few endogenous gut microbes that is capable of growing on host-derived mucins in the absence of dietary carbohydrate input (Martens *et al.,* 2008). It was shown that in *B. thetaiotaomicron* 15 different polysaccharide utilization loci (PULs) are upregulated in the presence of different glycan fractions derived from porcine gastric mucin (PGM) compared to the glucose control (Martens *et al.,* 2008; Koropatkin *et al.,* 2014). It was suggested that host glycan foraging by *B. thetaiotaomicron* helps to stabilize the microbial community population affected by the alterations in dietary glycan intake, in turn avoiding dysbiosis of the gut microbiota and potential disease (Tailford *et al.,* 2015).

1.4.2. N-glycan degradation pathways

The last two decades of human microbiome research has produced a tremendous amount of data that collectively reveals the significant impact the gut microbiota has on human metabolism, physiology and overall health (Sender *et al.,* 2016). Although we now possess an extensive knowledge about the degradation pathways of major classes of glycans by human gut microbiota, such as plant polysaccharides, fungal cell walls and host glycosaminoglycans (GAGs), there is a paucity of data describing the underlying processes of N-glycan degradation by the gut microbiota, particularly dietary and host-derived complex N-glycoproteins. As previously discussed, N-glycans are prominent and may be an important nutrient source for the members of gut microbiota. Because glycosylation patterns of N-glycoproteins are incredibly complex and diverse, these structures would require an extensive metabolic machinery to degrade.

The gut microbiota possesses numerous enzymes that allow degradation of various macromolecular structures, but little information is available on mechanisms gut microbes employ to degrade N-glycan structures. Up to date, the N-glycan degradation pathway research has been mainly focused on pathogenic microbes. Efficient N-glycan utilisation has been observed by *Bacteroides fragilis in vitro*. This opportunistic pathogen was shown to employ an outer member protein complex termed Don to deglycosylate human serum glycoprotein transferrin. The Don PUL is composed of seven genes, encoding a SusC/D-like protein pair, a GH18 endo-N-acetyl-β-d-glucosaminidase and a α-N-acetylglucosaminidase. The sialylated N-glycan degradation by *B. fragilis* also requires the cooperative activity of a neuraminidase NanH to release terminal sialic acid residues (Cao *et al.,* 2014). Similar PUL was identified in *Capnocytophaga canimorsus*, a Gram-negative bacterium commonly found in the dog oral cavity that can cause severe infections in humans. It was observed that *C. canimorsus* is capable of deglycosylating N-glycan structures found on human IgG antibody and fetal calf serum fetuin *in vitro* by employing a large Sus-like system consisting of the GpdCDGEF proteins and a sialidase SiaC. The N-glycans are cleaved off the protein backbone by a GH18 family endoglycosidase GpdG before being imported into the cell. This ability of *C. canimorsus* to harvest N-glycans was found to be important for bacterial growth *in vivo* (Renzi *et al.,* 2011). However, the rest of the degradation pathway is not known.

Complex N-glycan degradation was also observed by the exoglycosidases encoded by the opportunistic pathogen *Streptococcus pneumoniae* (King *et al.,* 2006). In *S. pneumoniae*, the complex N-glycan structures are first depolymerised down to the $Man_3GlcNAc_2$ core before it is released from the protein backbone by the endo-β-N-acetylglucosaminidase (EndoD) and imported into the cell for further processing. *S. pneumoniae* was shown to be capable of deglycosylating lactoferrin, IgA-1 and α_1-acid glycoprotein, where the utilisation of α_1-acid glycoprotein N-glycans as a sole carbon source was enough to sustain bacterial growth (Burnaugh *et al.,* 2008).

A common human pathogen, *Streptococcus pyogenes*, secretes a large GH18 family endoglycosidase termed EndoS that was shown to be capable of cleaving the N-glycan structures off the protein backbones of human IgG antibodies (Collin *et al.,* 2001). This activity was linked to the enhanced host immune response evasion and subsequent increased rate of survival of the bacterium in human blood *ex vivo* (Collin *et al.,* 2002).

Furthermore, a human oral pathogen *Tannerella forsythia* secretes a sialidase NanH that was shown to be capable of cleaving sialic acid residues off bovine fetal serum fetuin and bovine submaxillary mucin (BSM) *in vitro*. It was suggested that the utilisation of glycoprotein-associated sialic acid promotes the growth of *T. forsythia* biofilm *in vivo* (Roy *et al.,* 2011).

In addition, humans can only synthesize N-acetylneuraminic acid (Neu5Ac) whereas non-human mammals can synthesize both Neu5Ac and N-Glycolylneuraminic acid (Neu5Gc). Interestingly, it was observed that humans are capable of incorporating the non-human sialic acid variant Neu5Gc into their cells which in turn stimulates the xeno-autoimmunity (Varki *et al*., 2017). Considering that the sialic acids are the common attachment sites for many pathogens, it is not surprising that Shiga toxin (Stx) produced by *Shigella dysenteriae* and Typhoid toxin produced by *Salmonella enterica* were observed to be capable of targeting both the human Neu5Ac and non-human Neu5Gc that was incorporated into human glycoproteins (Varki *et al.,* 2009).

Overall these data suggest that the N-glycan structures are utilised by the bacterial pathogens not only as a nutrient source in the gut but also as a mechanism to evade host immune responses extra-intestinally (e.g. in the gut wound, blood). Although the N-glycan degradation was observed by various pathogens discussed above, the full utilisation pathway has not been described yet. It is also not clear whether similar degradation model would be employed by mutualistic gut microbiota species that can access N-glycans. *B. thetaiotaomicron* was previously shown to be capable of breakdown of high-mannose N-glycans but it was not known if it can access complex and hybrid N-glycan structures (Cuskin *et al.,* 2015).

1.5. Research objectives

It was previously shown that a number of polysaccharide utilization loci (PULs) are upregulated by *Bacteroides thetaiotaomicron* during the growth in the presence of complex mucin-derived glycan structures (Martens et al., 2008). Despite the extensive characterization of PULs involved in foraging of host-derived glycans by various pathogenic bacteria, the knowledge of N-glycan degradation pathways by mutualists residing in the human gut is lacking. Understanding the precise mechanisms employed by commensal gut microbiota to degrade these complex glycoproteins can have major implications not only in promotion of human health but also in an industrial context.

To achieve this aim, the project was divided into two main goals:

i. Investigating complex N-glycan utilization by human gut *Bacteroides*

This study aimed to 1) investigate whether human gut *Bacteroides* can utilize complex N-glycans (**Chapter 3**), 2) identify potential N-glycan utilization loci in *B. thetaiotaomicron* using transcriptomic analyses (**Chapter 3**) and 2) biochemically and structurally characterize the function of enzymes and glycan binding proteins predicted to be involved in N-glycan utilization (**Chapter 4**).

ii. To propose an N-glycan degradation model

N-glycosylation patterns are exceptionally intricate and diverse. To fully degrade such complex structures, bacteria require a cooperative activity of numerous glycan binding proteins, enzymes, transporters and regulators. The identification of N-glycan structure linkages these enzymes target allows for a rational understanding of the degradative mechanism and provides insight whether it requires a cooperative activity of multiple PULs. Combined, this information would allow for an N-glycan degradation model to be proposed.

This study aimed to 1) investigate potential cooperative activity of multiple N-glycan binding proteins and degrading enzymes, 2) determine their cellular localization, 3) investigate the importance of key enzymes to N-glycan utilization and 4) to propose a model of N-glycan degradation by *B. thetaiotaomicron* (**Chapter 5**).

Chapter 2: Materials and Methods

2.1. Molecular biology

2.1.1. Bacterial strains and vectors

The bacterial strains and plasmids that were used during this research project are listed in Table 2.1 and Table 2.2, respectively.

Table 2.1: Bacterial strains

Strain	**Genotype features**	**Description**
BL21 (DE3)	F- *dcm omp*T *hsdS*(rB- mB-) *gal* (DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])	*E. coli* strain optimised for over-expression of recombinant proteins using a T7 promoter. (Studier & Moffatt, 1986)
Tuner (DE3)	F– *omp*T *hsd*SB (rB– mB–) *gal dcm lac*Y1 (DE3)	This *E.coli* strain carries a *lacY* mutation optimised for overexpression of recombinant proteins using IPTG.(Novagen)
Top10	F- *mcrA* Δ(*mrr*-*hsd*RMS-*mcr*BC) ф80*lacZ*ΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K *rps*L (STRR) *end*A1 *nup*G λ-	*E. coli* strain used for plasmid propagation and cloning. (Invitrogen)
CC118 **λ-pir**	Δ(*ara-leu*) *ara*D Δ*lac*X74 *gal*E *gal*K *pho*A20 *thi*-1 *rps*E *rpo*B *arg*E (Am) *rec*A1 λ*pir*	*E. coli* strain used in gene deletion plasmid propagation. (Herrero et al. 1990)
S17 **λ-pir**	A *pro hsd*R RP4-2 (Tc::Mu; Km::Tn7) (λ pir)	*E. coli* strain used in conjugation of pExchange plasmid carrying gene deletion. (Skorupski & Taylor, 1996)
B. thetaiotaomicron Δtdk	*Δtdk*	*B. thetaiotaomicron* VPI-5482 strain lacking a thimadine kinase. Used to generate genomic mutants through FuDR selection. Provided by Eric Martens.
B. thetaiotaomicron	VPI-5482	Distaso (1912), Castellani and Chalmers (1919) strain. Provided by Eric Martens.
B. fragilis	NCTC-9343	Veillon and Zuber (1898), Castellani and Chalmers (1919) strain. (DSMZ.de)
B. caccae	ATCC-43185	Johnson et al. (1986) strain. (DSMZ.de)
B. ovatus	ATCC-8483	Eggerth and Gagnon (1933) strain. (DSMZ.de)
B. massiliensis	CCUG-48901	Fenner et al. (2005) strain. (DSMZ.de)
B. finegoldii	JCM-13345	Bakir et al. (2006) strain. (DSMZ.de)
B. vulgatus	ATCC-29327	Eggerth and Gagnon (1933) strain. (DSMZ.de)
B. cellulosilyticus	CCUG-44979	Robert et al. (2007) strain. (DSMZ.de)
B. nordii	JCM-12987	Song et al. (2005) strain. (DSMZ.de)
B. intestinalis	JCM-13265	Bakir et al. (2006) strain. (DSMZ.de)
B. uniformis	ATCC-8492	Eggerth and Gagnon (1933) strain. (DSMZ.de)
B. fluxus	JCM-16101	Watanabe et al. (2010) strain. (DSMZ.de)
B. salyeriae	JCM-12988	Song et al. (2005) strain. (DSMZ.de)
P. merdae	ATCC-43184	Johnson et al. (1986) strain. (DSMZ.de)
B. longum	NCTC-11818	Reuter (1963) strain. (DSMZ.de)

Table 2.2: Plasmid list

Plasmid	Features	Supplier/Reference
pET28a-b	Kan^r, T7 promotor, *lac, laciq,* integrated His tag	Novagen
pExchange-*tdk*	Amp^r, erm^r, *tdk* modified suicide vector	Provided by Nicole Koropatkin (Koropatkin *et al*, 2008)

2.1.2. Bacterial growth and selective media

The list of growth media used in this study is shown in Table 2.3. The media was used alone as inoculum broth or poured into plates following the addition of of 2% (w/v) of Bacteriological agar. *Bacteroide*s spp. were grown in sterile TYG, MM+carbon source or BHI media whereas *E. coli* strains were grown in sterile Luria-Bertani Broth (LB). When necessary, the selective antibiotics were added to the media after sterilisation (Table 2.4).

Table 2.3: List of media

Medium	Composition	Quantity (per litre)	Method
Luria-Bertani (LB)	Granules, as provided by the manufacturer (Sigma-Aldrich)	25 g	Dissolved in MiliQ water and sterilised before use
Tryptone-Yeast Extract-Glucose (TYG)	Tryptone/Peptone Yeast Extract Glucose Cysteine, free base 1 M KPO_4 pH 7.2 0.4 mg/ml $FeSO_4$ 1 mg/ml Vitamin K 0.8 % $CaCl_2$ 0.25 mg/ml Resazurin TYG Salt Solution ($MgSO_4$ 0.5 g/l, $NaHCO_3$ 10 g/l, NaCl 2 g/l)	10 g 5 g 2 g 0.5 g 100 ml 1 ml 1 ml 1 ml 4 ml 40 ml	Dissolved in MiliQ water and sterilised before use
Minimal Media Bacteroides (MM+0.5 % target Glycan)	NH_4SO_4 Na_2CO_3 Cysteine, free base 1 M KPO_4 pH 7.2 0.4 mg/ml $FeSO_4$ 1 mg/ml Vitamin K 0.01 mg/ml Vitamin B_{12} 0.25 mg/ml Resazurin MM Salt Solution (NaCl 18 g/l,$CaCl_2$ 0.53 g/l, $MgCl_2$ 0.4 g/l, $MnCl_2$ 0.2 g/l, $CoCl_2$ 0.2 g/l)	1 g 1 g 0.5 g 100 ml 10 ml 1 ml 0.5 ml 4 ml 50 ml	Dissolved in MiliQ water and sterilised before use
Brain-Heart Infusion (BHI)	Powder, as provided by the manufacturer (Sigma Aldrich)	37.5 g	Dissolved in MiliQ water and sterilised before use
His-Heme	Hematin 0.42 g/l Histidine-HCl pH 8.0	1.2 g 1 l	Used to as a supplement to TYG, MM and BHI media. 1:1000 dilution after media has been sterilised

Table 2.4: List of antibiotics

Antibiotics	Working concentration	Storage
Ampicillin	50 µg/ml	-20 °C
Kanamycin	50 µg/ml	-20 °C
Gentamycin	200 µg/ml	Prepared when necessary
Erythromycin	25 µg/ml	Prepared when necessary
5-fluoro-2'-deoxyuridine (FUdR)	200 µg/ml	Prepared when necessary

2.1.3. Sterilization

Astell Hearson 2000 Series Autoclave or a Prestige© Medical Series 2100 Clinical Autoclave were used at 121 °C for 1 h to sterilise all solutions and glassware before use. In addition, where applicable, solutions were filtered using sterile 0.22 µm Milipore filter discs (Stupor® Acrodisc® 3.2 Gelman Sciences) and sterile syringes (Plastipak®, Becton Dickinson).

2.1.4. Chemicals, carbohydrates and commercial kits

All experiments were done using double distilled water as a solvent unless noted otherwise. Millipore Milli-RO 10 Plus Water Purification System was used to produce purified MQ H2O (18.2Ω). The chemicals, media, commercial kits and substrates were mainly purchased from Sigma Aldrich, Megazyme, Carbosynth, Dextra or BDH Laboratories Ltd., with special exceptions listed in the text.

2.1.5. Storage information

Stocks of plasmid DNA, cDNA and genomic DNA were stored at -20 °C long-term and -4 °C short-term. Stocks of RNA were stored at -80 °C short-term. Bacterial strain stocks were stored in 25 % (v/v) glycerol at -80 °C.

2.1.6. Centrifugation

Beckman J2-21 centrifuge was used to harvest large-scale cultures of bacterial cells by centrifugation in 500ml Nalgene bottles at 5000 rpm for 10 minutes using JA-10 rotor (at 4 °C). Whereas JA25-50 rotor was used to harvest cell lysate at 25000 rpm for 30 minutes (at 4 °C). Hettich Mikro 220R Refrigerated benchtop centrifuge was used to spin small-scale samples (1-10 mL) at speeds up to 6000 rpm, in 25 ml universal (Sterilin) tubes. Eppendorf tubes (up to 2mL) were centrifuged using a Heraeus Pico 21 benchtop microcentrifuge at speeds up to 14,000 rpm.

2.1.7. Transformation of competent *E.coli*

Stocks of chemically competent *E.coli* strains were produced by Mr. Carl Morland and stored at -80 °C long-term (Cohen *et al.*, 1972). A 100 µl aliquot of required *E. coli* strain was thawed on ice before 1-5 µl (200 ng) of plasmid DNA or ligation reaction was added. Reaction was incubated on ice for 30 minutes before the cells were heat-shocked at 42 °C for 1 minute and incubated on ice for further 2-5 minutes. 200 µl of fresh LB media was then added and the mixture was incubated at 37 °C, with shaking, for up to 1 hour (1h for ligation reactions). Cells were then harvested by centrifugation, re-suspended in 100 µl of fresh LB media and plated onto LB plates containing an appropriate antibiotic and incubated for 16 hours lid-down at 37 °C, aerobically.

2.1.8. DNA extraction and quantification

Plasmid DNA was propagated by transforming a required plasmid into Top10 competent *E.coli* cells as described previously. A single colony was then inoculated into 5mL LB with an appropriate antibiotic and incubated at 37 °C overnight, with shaking. The cultures were harvested by centrifugation, as outlined previously, and plasmid DNA was purified using a QIAspin Prep kit (QIAGEN) as per manufacturer's instructions.

Genomic DNA was extracted using GenElute ™ Bacterial Genomic DNA kit (Sigma-Aldrich) following the manufacturer's protocol.

NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Fisher Scientific) was used to determine DNA concentration by measuring the absorbance at 260nm and 280nm. The 260/280 absorbance ratio was used to estimate the sample purity. Pure DNA samples have a ratio of ~ 1.8.

2.1.9. Polymerase chain reaction

The polymerase chain reaction (PCR) technique was used to amplify the target DNA following a variant of a protocol first developed by Mullis and Faloona in 1987. KOD Hot-Start Polymerase Kit (Novagen) was used to catalyse the synthesis of new complementary DNA strands based on either plasmid DNA or genomic DNA template using a BioRad PHC-3 ThermoCycler.

For a standard PCR, primers were constructed to target specific DNA regions with the forward primer complimentary to the 5' end of the forward strand and the reverse primer complimentary to the 5' end of the reverse strand. The primers were constructed 15-30 bases in length, with ~50% GC content and with a melting temperature (Tm) of >45 °C. Primer parameters were calculated using an OligoCalc online tool (Kibbe, 2007). Where required, restriction sites were added to the 5' ends of the primers and additional sequence of 6 nucleotides upstream of the restriction site for efficient cutting of the PCR products by the restriction endonucleases. Primers were synthesized by Eurofins Genomics (MWG) and resuspended in PCR-grade water to a 100pmol/µl working concentration.

2.1.9.1. Standard PCR conditions

Standard PCR reaction mix was prepared as shown in Table 2.5 using the KOD Hot-Start DNA Polymerase Kit (Novagen). Master mixture was kept on ice until use.

Table 2.5: Standard PCR reaction mix composition

Reagent	Volume (µl)
KOD Buffer (10x)	5
dNTPs (2mM)	5
MgSO4 (25mM)	2.5
Forward Primer (100 µM)	2.5
Reverse Primer (100 µM)	2.5
Template DNA	1
KOD DNA polymerase (1 U/µl)	0.5
PCR-grade dH2O	Up to 50

The standard PCR program used on the thermocycler is shown in Table 2.6.

Table 2.6: Standard PCR program

	Initial denaturation	95 °C	60 seconds
35 cycles	Denaturation	95 °C	30 seconds
	Annealing	55 °C	30 seconds
	Elongation	68 °C	60 seconds
	Final extension	68 °C	600 seconds
	Storage	4 °C	∞

After the amplification, DNA product was analysed using agarose gel electrophoresis. If the product yield was low or no product of right size was detected, PCR reaction parameters, such as annealing and elongation time, were adjusted. In some cases, dimethyl sulfoxide (DMSO) was added to the reaction mix (3 µl).

2.1.9.2. Agarose gel electrophoresis

Agarose gel electrophoresis was used to analyse the DNA/RNA nucleic acid samples. 1 % (w/v) agarose gel was used routinely throughout the study. Solution was prepared by mixing 0.5g of agarose powder (Sigma-Aldrich) in 50 ml of 1 x TBE buffer (2mM EDTA, 89mM Tris Base, 89mM Boring Acid) and dissolving it completely by boiling. The solution was cooled to around <50 °C and 0.5 µg/ml of ethidium bromide was added before being poured to the pre-assembled mini-gel system mould tank (Applied Biosystems). The gel was let to set for approximately 30 minutes before being used.

For DNA sample loading, agarose gel was submerged in 1 x TBE buffer. DNA samples were mixed with 5 µl of 10 x of loading dye (10 x TBE, 50% glycerol, 0.25% Bromophenol blue) and loaded into the gel wells. Hyperladder I (Bioline) was used as DNA standard marker. The size of DNA fragments was estimated by comparing their electrophoretic migration with that of known standards. The samples in the gel were separated by size by running them at 70 V for 45-60 minutes (LKB Bromma 2197 Power Supply).

Bio-Rad Gel Doc EZ Imager was used to visualise the DNA bands on the gel following the electrophoresis. Images were printed using Mitsubishi Video Copy Processor and thermal paper.

2.1.10. Site-directed mutagenesis (SDM)

Site-directed mutagenesis was used to generate single amino acid mutants in this study. Forward and Reverse synthetic primers designed with a desired amino acid mutation were used to mutate a double-stranded recombinant plasmid carrying the wild-type protein. The primers were made 12bp in length and complementary to the DNA template, apart from one mutated amino acid. Same PCR mix was used for site-directed mutagenesis as described previously with a specially-designed SDM thermocycler program (Table 2.7).

Table 2.7: Standard PCR program for SDM

	Initial denaturation	95 °C	60 seconds
20 cycles	Denaturation	95 °C	30 seconds
	Annealing	55 °C	60 seconds
	Elongation	68 °C	60 seconds per 1kb
	Final extension	68 °C	600 seconds
	Storage	4 °C	∞

Following the PCR amplification, the reaction mix was allowed to cool before 1µl of *DpnI* restriction endonuclease (Thermo Fisher Scientific 10 U/µl) and 3µl of 10 x Tango Buffer were added. *DpnI* recognises the GA_^TC sequence of methylated DNA, leaving the un-methylated PCR product undigested. The *DpnI* digestion reaction was incubated at 37 °C for 2 hours and 5 µl of digested PCR product was transformed into chemically competent *E.coli* cells.

2.1.11. Restriction digestion and purification of DNA

Restriction digestion of up to 1000ng of DNA was done following a standard protocol provided by the manufacturer. Briefly, appropriate 10 x reaction buffer was mixed with the DNA and 1 µl of required restriction endonuclease was added to the reaction (Thermo Fisher Scientific). The final reaction volume was made up to 40 µl using Ultra-Pure water (Sigma-Aldrich) and incubated at 37 °C for 1 hour.

QIAquick PCR Purification kit (QIAGEN) was used following the manufacturer’s instructions to purify the digested DNA and all PCR products.

Vector DNA was purified using QIAquick Gel Extraction kit (QIAGEN) following the manufacturer’s protocol. Briefly, following the vector digestion by restriction endonucleases, the DNA product was run on a 1 % (w/v) high purity Seachem Gold Agarose gel using the agarose gel electrophoresis as described previously. The required DNA band was cut from the gel using a scalpel and UV transilluminator before being purified using the gel extraction kit.

2.1.12. Ligation of insert and vector DNA

Digested vector DNA and inserted DNA were ligated using Rapid DNA ligation kit (Thermo Fisher Scientific) following protocol outlined in Table 2.8. In each reaction, insert to vector molar ratio of 3:1 was calculated and used.

Table 2.8: Ligation reaction mix

Reagent	Amount per reaction
Digested Insert DNA	20 ng of DNA
Digested Vector DNA	Calculated to 3:1 ratio
5 x Ligation Buffer	4 µl
T4 DNA ligase	1 µl
PCR-grade H2O	Up to 20 µl

2.1.13. Sewing PCR

'Sewing' PCR, otherwise known as PCR overlap extensions, was used generate plasmid DNA for homologous recombination used to mutate the genome of *Bacteroides thetaiotaomicron*. This method was used to generate knock-out mutants. pExchange *tdk* plasmid and *E. coli* CC118 λ pir strain was used for cloning. Primers were routinely designed with either BamHI and Xbal sites or SalI and SpeI sites.

To remove a gene, four primers flanking 1,000 bp upstream (Flank 1) and downstream (Flank 2) of the gene were designed (Figure 2.1). In order to produce a mutant lacking the gene, 20 bp of mutation extension of flank 1 primer was designed to be complementary to one of the opposite flank.

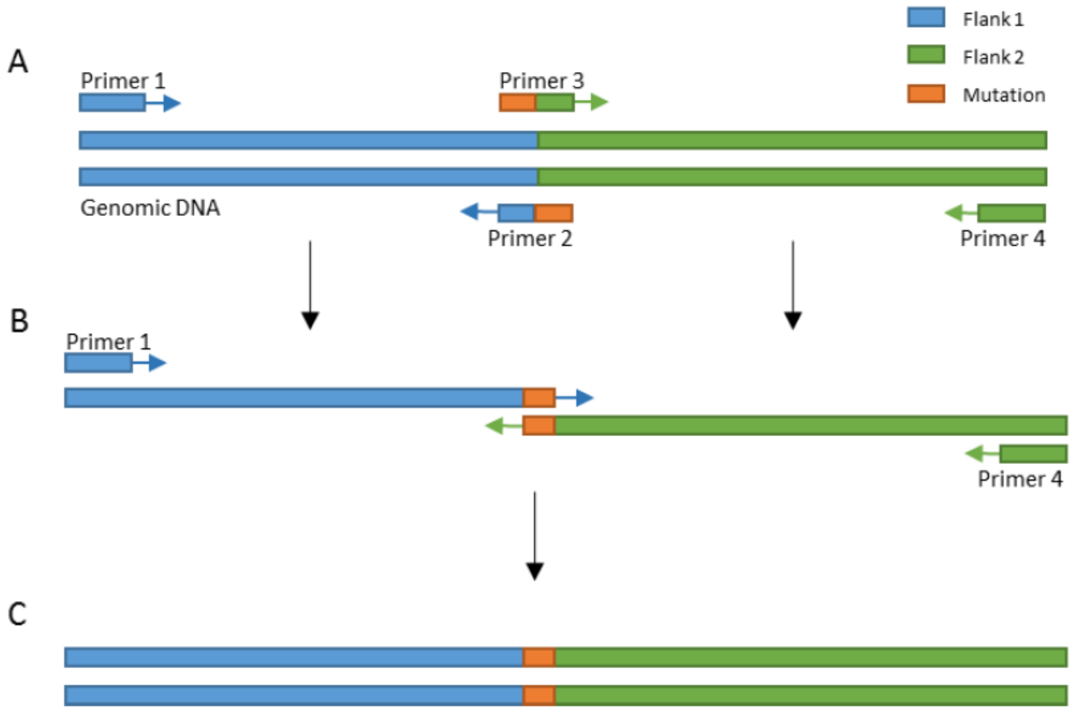


Figure 2.1: Graphical representation of the sewing PCR concept.
A) 1000 bp of upstream Flank 1 and 1000 bp of downstream Flank 2 were amplified using two separate PCR reactions. The primer 2 and primer 3 carried a complementary mutation extension sequence. B) The flank 1 and flank 2 are 'sewn' together and amplified using primer 1 and primer 4 during a third PCR reaction. C) The final PCR product carried a desired mutation and 1000bp flanking regions. Figure was adapted from Amy Glenwright thesis.

The first step was to produce 1000 bp of upstream and 1000 bp of downstream flank by amplifying them in two separate PCR reactions with 2 sets of primers: primer 1 and 2 for flank 1 and primer 3 and 4 for flank 2. The second step was to 'sew' the two flanks together at the complementary mutation extension and amplify the product in a third PCR reaction using primer 1 and primer 4. The final step was to purify the 2000 bp PCR product carrying a desired mutation and insert it into pExchange-*tdk* vector using the ligation protocol described previously. The plasmid was then transformed into CC118 λ pir *E. coli* strain and purified for further use.

2.1.14. Automated DNA sequencing

Value Read service by MWG-Eurofins Genomics was used to sequence DNA. MWG-Eurofins Genomics uses ABI sequencers. Pre-paid single value-read labels were provided by the company. For each sample, two labelled tubes containing 7 µl of 50-100 ng of DNA were sent for sequencing in both forward and reverse direction using either specifically designed primers or primers provided by the company (T7 prom and T7 term). T7 promoter sequence – TAATACGACTCACTATAGGG, T7 terminator sequence – CTAGTTATTGCTCAGCGGT.

2.1.15. RNA extraction and sequencing

Wild type *B. thetaiotaomicron* was grown anaerobically overnight at 37°C in 5ml of TyG media. The overnight cultures were then inoculated into 5ml of Minimal Media containing either 50mg of N-glycan or 25 mg of glucose as a carbon sources. The bacterial cultures were then grown to mid-exponential stage (O.D. of ~0.4 for N-glycan and ~0.6 for glucose). The cells were harvested by centrifugation using the RNAProtect Bacterial Reagent (QIAGEN) to stabilize and protect the bacterial cultures from the degradation of RNA transcripts and induction of genes. RNA was then extracted using the commercially available RNeasy Mini kit (QIAGEN), following the manufacturer's instructions. The concentration and purity of RNA was estimated by measuring the absorbance at 260nm and 280 nm respectively using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific Inc, USA). To estimate RNA purity, the ratio of the absorbance contributed by the nucleic acid to the absorbance of the contaminants is determined. Typically, a requirement for A260/A280 ratios are between 1.8 and 2.2. RNA quality was also analysed using the agarose gel electrophoresis. Intact RNA is represented by 16S rRNA band and example is shown in Figure 2.2. The extracted RNA was then frozen and stored at -80°C until required.

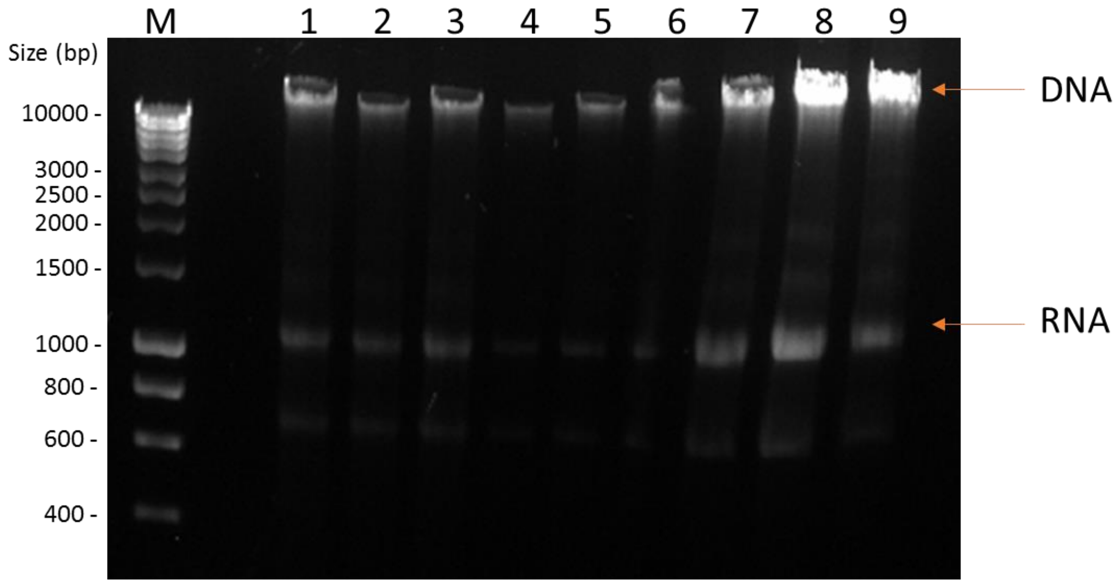


Figure 2.2. Agarose gel visualisation of RNA samples. This agarose gel shows an example of purified RNA samples. Hyperladder^{TM} I (HLI) DNA ladder was run to compare the band sizes. Bands at the top of the gel in wells 7, 8 and 9 labelled DNA represent DNA contamination present in the RNA samples and requirement for DNAse treatment.

The extracted RNA was sent to the Oxford Genomics Centre or The Earlham Institute for sequencing. The sequencing was done using the High-performance Illumina HiSeq4000 75bp system, 1 PE lane (240 million reads per lane). The data was aligned to the reference and quality-checked before being returned to us.

2.1.16. cDNA synthesis

RNA was converted to the cDNA using the QuantiTect Reverse Transcriptase Kit (QIAGEN) following the manufacturer's instructions. The kit is optimised for use in real-time PCR, removing integrated genomic DNA and allowing for reliable quantification of gene targets from mRNA transcripts.

2.1.17. Quantitative Polymerase Chain Reaction (qPCR)

Quantitative Polymerase Chain Reaction (qPCR) was used to investigate the upregulation of specific genes during the mid-exponential growth phase of *B. thetaiotaomicron* using specially designed primers and cDNA templates.

cDNA was obtained using a protocol described above. The qPCR primers were designed by taking the nucleic acid sequence of the gene of interest and using the GenScript Real-time PCR (TaqMan) Primer Design software available on Genscript website to design the primers. 3 sets of primers were designed with the amplicon size range between 100-200 bp, and Primer Tm temperatures between 55-60 °C. The designed primers were synthesized by MWG-Eurofins Genomics and resuspended in PCR-grade water to a 100 pmol/µl working concentration. The primers were checked using the wild type Bt DNA and standard PCR reaction. Results were visualised on an agarose gel and 100-200 bp PCR products were expected (Figure 2.3).

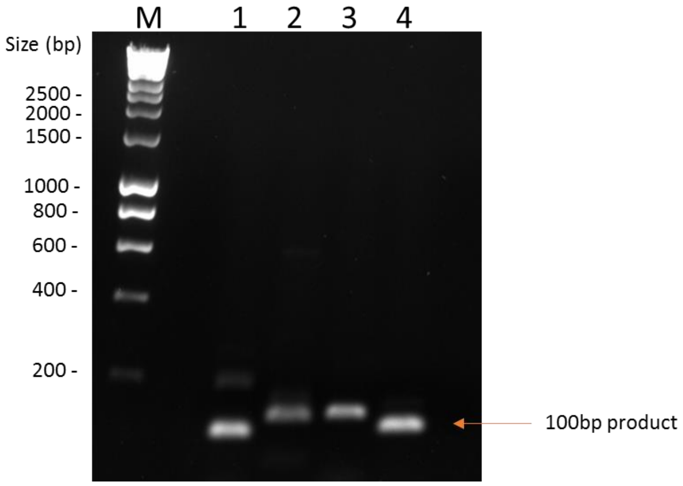


Figure 2.3: An example of agarose gel showing the products amplified by the qPCR primers. Single bands of 100-200 bp in size showed that primers are working and can be used in qPCR. 2 bands of DNA are visible on lane 1 and suggesting that this primer could not be used.

SYBR Green I kit (Roche) was used in this study. It contains a dye that absorbs light at 497 nm and emits light at 520 nm when inserted into double stranded DNA. During each amplification step of qPCR, emitted light at 520 nm is measured. The more amplification occurs; the more intensity of light is released.

For a standard qPCR, a 10 µl of a reaction mix was prepared (Table 2.9)

Table 2.9: qPCR reaction mix

Reagent	Volume (µl)
SYBR Green I Master Mix (Roche)	5
Forward primer (5 µM)	1
Reverse primer (5 µM)	1
Template cDNA (50 ng)	2
PCR-grade H2O	40

Quantitative Polymerase Chain Reaction (qPCR) was done using a program given in Table 2.10 on a Roche LightCycler® 96.

Table 2.10: Standard qPCR program

	Initial denaturation	95 °C	600 seconds
45 cycles	Denaturation	95 °C	10 seconds
	Annealing	57 °C	10 seconds
	Elongation	72 °C	10 seconds
	Light measurement	72 °C	-

Upon completion, data set was normalised using the Roche LightCycler® 96 analysis software.

2.1.18. Counter-selectable mutagenesis of the *B. thetaiotaomicron* genome

A pExchange-*tdk* plasmid carrying a desired mutation produced during a Sewing PCR described previously was transformed into S17 λ *pir E. coli* cells. From this point, S17 λ pir *E. coli* strain will be referred as a 'donor' whereas *B. theta Δtdk* strain will be referred as the 'recipient' (Koropatkin *et al.,* 2008).

The donor and recipient strains were grown overnight in 5 ml of LB and 5ml of TYG broth, respectively. Bacterial cells were harvested by centrifugation, donor and recipient cell pellets were re-suspended and equally mixed together in 1 ml of TYG (Figure 2.4). The mixture was evenly spread on BHI-Heme agar plate and incubated lid side up aerobically for 24 hours at 37°C. Because of such a thick layer, a lawn of *E. coli* cells grew first providing anaerobic conditions underneath for *B.*

thetaiotaomicron cells to grow, subsequently providing required conditions for conjugation between the donor *E. coli* cells and recipient *B. thetaiotaomicron* cells to occur. The thick lawn of cells was then scraped and re-suspended in fresh 5 ml of TYG, with no antibiotic. The 100 µl of this solution along with 1:10 and 1:100 of dilutions were plated onto BHI-Heme plates containing erythromycin (25 µg/ml) and gentamycin (200 µg/ml). These antibiotics are selective for the recipient *B. thetaΔtdk* strain and pExchange-*tdk* plasmid, therefore only cells which have gone a single recombination should grow. These plated cells were incubated at 37°C anaerobically for up to 3 days, before 10 colonies were re-streaked onto fresh plates to ensure minimal chance of the contamination by *wt B. theta*. The plates were incubated at same conditions for up to 3 more days before 10 colonies were picked and inoculated in 5 ml of TYG each and grown overnight anaerobically at 37°C.

The cells were harvested by centrifugation and pooled together. The 100 µl of this solution along with 1:10 and 1:100 of dilutions were plated onto BHI-Heme agar plates containing FUdR (200 µg/ml) antibiotic. These plates were then incubated for further 3 days anaerobically at 37°C. The FUdR allows for selection for cells which have undergone a second recombination event, eliminating pExchange-*tdk* sequence from the genome. The resistant colonies were re-streaked onto fresh plates to reduce contamination with *wt B. theta* cells.

10 colonies were then inoculated in 5 ml of TYG each for genomic DNA extraction and glycerol stock creation. The purified DNA was then screened for successful mutations using standard PCR and primer 1 and primer 4. Agarose gel electrophoresis was used to visualise the products and determine if it's a mutant. Genomic DNA of strains that showed a band of right size following the PCR were sent to sequencing.

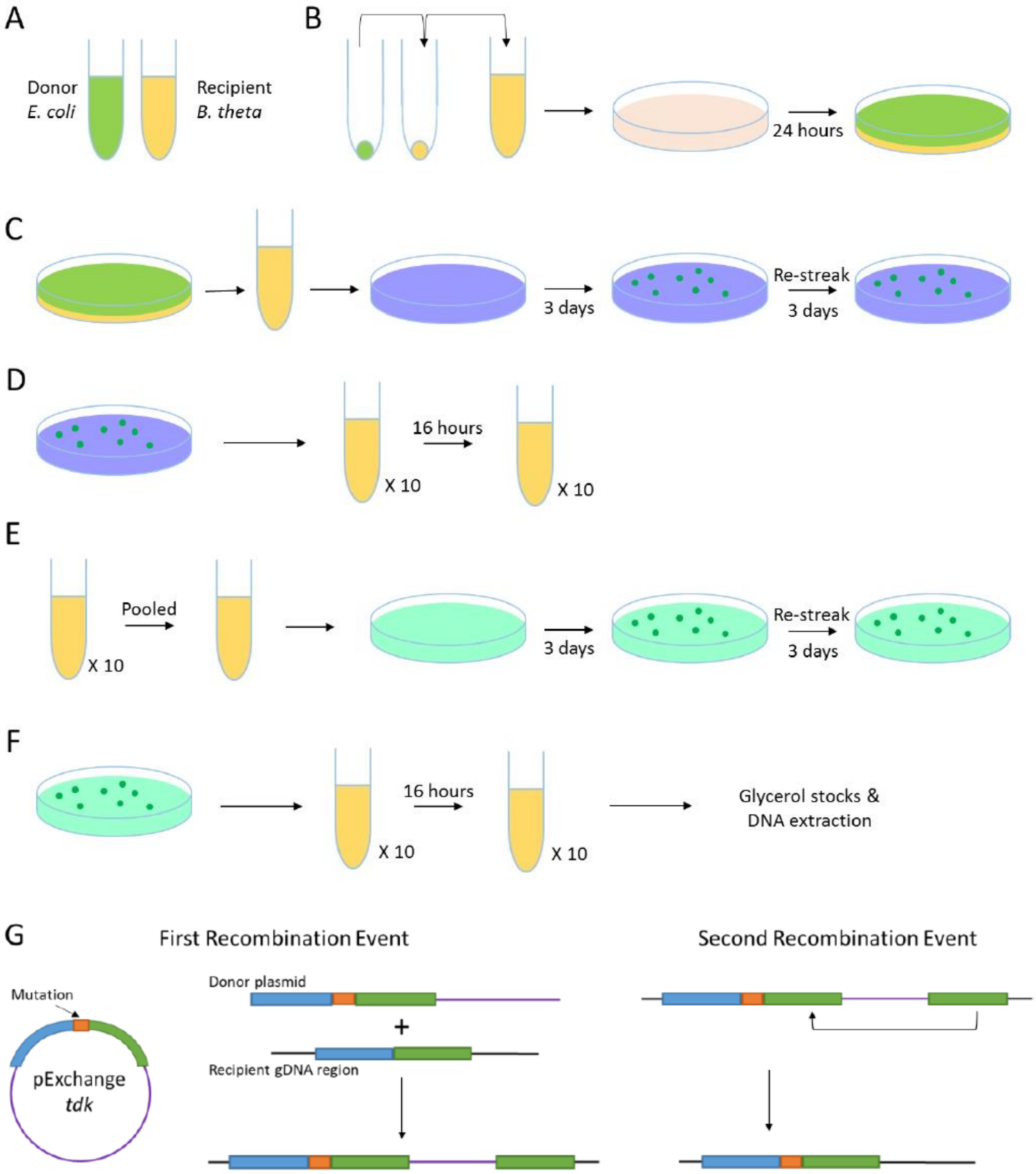


Figure 2.4: Graphical representation of mutagenesis protocol.

A) The donor *E.coli* S17 λ *pir* cells and the recipient *B. thetaiotaomicron Δtdk* cells were grown overnight in 5ml of LB and TYG media, respectively. **B)** Pellets of both donor and recipient cells were harvested by centrifugation and combined equally in 5 ml of TYG media. The mixture was plated onto BHI-heme plates and grown aerobically for 24h. **C)** The thick layer of cells was scrapped off the plate and re-suspended in 5ml of TYG media. The mixture was plated onto BHI-heme plates + gentamycin (200 µg/ml) and erythromycin (25 µg/ml) and grown for 3 days anaerobically. Resistant colonies were then re-streaked onto new plates and grown for further 3 days. **D)** 10 colonies were picked and inoculated in TYG media. The cells were grown for 16 hours, anaerobically (First recombination event). **E)** The 10 cultures were pooled, harvested by centrifugation and re-suspended in fresh 5ml of TYG. The mixture was then plated onto BHI + FUdR (200 µg/ml) plates and grown for 3 days, anaerobically (Second recombination event). Resistant colonies were re-streaked onto fresh plates to minimise contamination by wt Bt. **F)** 10 colonies were picked and grown overnight in TYG media. These cultures were used to make glycerol stocks and for DNA extraction. G) A graphical representation of the recombination events. *pExchange-tdk* plasmid carrying mutation enters the genome during the first recombination event. The second recombination event eliminated *pExchange-tdk* from the genome. Image is adapted from Glenwright, 2017 (PhD thesis).

2.1.19. Automated monitoring of bacterial growth curves

Bacterial growth curves were monitored using an Epoch^{TM} Microplate Spectrophotometer (BioTek Instruments). Corning Costar 96-Well Cell Culture Plates (Sigma Aldrich) were used for the growths inside an anaerobic cabinet, at 37 °C (Don Whitely Scientific). A set-up of triplicates of 200 µl of media + required bacterial culture in each condition was routinely used in this study. Wells filled with 200 µl of media without the bacteria were used as a control. Epoch^{TM} Microplate Spectrophotometer measured the optical density at 600nm of each well every 15 minutes and recorded the results. The data was then collected from Gen5 2.05 software and investigated using GraphPad Prism 7.0 software.

2.1.20. Expression of recombinant proteins in *E.coli*

DNA of the plasmid encoding a protein of choice was transformed into either *E.coli* Tuner or BL21 cells using a protocol described previously. A single colony was then inoculated into 5 ml of LB media, containing a suitable antibiotic, and incubated at 37 °C for 16 hours, with shaking. 1ml of overnight culture was then injected into a 2L baffled flask containing 1 L of sterile LB media and a suitable antibiotic. The inoculum was then incubated at 37 °C, with shaking, until the culture reached the OD_{600nm} of approximately 0.6 - 0.8. The cells were cooled and protein overexpression was induced by adding 1 mM of IPTG and incubating at 16 °C overnight. The cells were then harvested by centrifugation as described previously.

2.1.21. Immobilised Metal Affinity Chromatography (IMAC)

In this study, all expressed soluble proteins were carrying a His-tag made of 6 or more histidine residues. This allowed proteins to be purified using an immobilised metal affinity chromatography (IMAC) where histidine tag bound to a positively-charged transition metal ions (e.g. cobalt or nickel) immobilised in the column. In order to elute the bound protein, high concentrations of imidazole were used, which disrupts the protein-metal interaction by competing for the binding of the metal ions.

Gravity-flow columns were set up using 5 ml of TALON resin (Clontech Laboratories) which contains cobalt ions. The column was then equilibrated using 40 ml of TALON buffer (20 mM Tris, 300 mM NaCl, pH 8).

The cells were grown and harvested as described above. The cell pellets were re-suspended in 10 ml of TALON buffer, transferred into sterile 20 ml tubes and stored at -20 °C until required. The cell mixture was sonicated on ice for 2-3 minutes using a low intensity setting on a B. Braun Labsonic® U Ultrasonic Homogenizer (~42 watts and 0.5 second cycling). The sonicated cells were then centrifuged for 30 minutes at 25,000 rpm and supernatant (cell free extract CFE) was collected.

The supernatant was filtered using a 0.45 µM PTFE syringe filter (Sigma Aldrich) and loaded onto buffer-equilibrated TALON column. The flow-through liquid was collected and TALON column was washed twice in 20-25 ml of TALON buffer. These fractions were collected. The protein was eluted in four fractions: 2 x 5 ml of TALON buffer containing 10 mM of Imidazole followed by 2 x 5 ml of TALON buffer containing 100 mM of Imidazole. SDS-PAGE gels were used to visualise the purification fractions.

2.1.22. Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was used to visualise the proteins and determine their purity, quantity and approximate size (Laemmli, 1970). Routinely, 12.5 % polyacrylamide gels (Acrylogel 3; BDH Electran) were prepared and used with two glass plates (12 cm x 10 cm) and a special rubber seal.

The buffers needed for the SDS-PAGE gel preparation and operation are listed in Table 2.11. The resolving gel solution was prepared according to the protocol and pipetted in-between the glass plates, leaving approximately 2 cm gap from the top, which was filled with water to allow the top of gel set in a straight line. The stacking gel solution was then pipetted into the 2 cm gap and a well-comb was inserted to form the wells. The comb was fixed into plate using a binder clip, do make sure there are no air bubbles or gaps formed. The gel was allowed to set at room temperature for approximately 30 minutes, well-comb and rubber band were removed before the gel was transferred into a gel tank. The tank was filled with the 10 x of SDS-PAGE running buffer. Samples pre-mixed with 1:2 of loading dye and boiled for 3 minutes. 20 µl of appropriate molecular weight (MW) protein standard was pipetted onto one of the wells, alongside of the cooled samples. A current of 35 A per gel was applied and gels were allowed to run for 45-55 minutes until the loading buffer has run off the end of the gel.

Following the gel electrophoresis, InstantBlue stain (Expedeon) was used to stain the gel for 20-30 min in a 150 rpm orbital shaker to visualise the protein bands. Gels were de-stained using distilled water and photographed using a Bio-Rad Gel Doc EZ Imager with Bio-Rad Image Lab software.

Table 2.11: SDS-PAGE gel and buffer recipes.

Component name	Reagents used	Volume or Concentration
Resolving gel (12.5%) /per gel	0.75 M Tris/HCl buffer, pH 8.8 + 0.2 % SDS 40 % Acrylamide (BDH Electran acrylamide, 3 % (w/v) bisacrylamide) Double distilled H2O 10 % (w/v) Ammonium persulphate Tetramethylethylenediamine (TEMED)	2.35 ml 1.45 ml 0.875 ml 22.5 µl 7.5 µl
Stacking gel /per gel	0.25 M Tris/HCl buffer, pH 6.8 + 0.2 % SDS 40 % Acrylamide (BDH Electran acrylamide, 3 % (w/v) bisacrylamide) Double distilled H2O 10 % (w/v) Ammonium persulphate Tetramethylethylenediamine (TEMED)	0.938 ml 0.188 ml 0.75 ml 15 µl 5 µl
SDS-PAGE running buffer (1 litre)	32 mM Tris / 190 mM glycine, pH 8.3 SDS	350 ml 0.1 % (w/v)
Loading buffer (10 ml)	0.25M Tris/HCl, pH 8.8 Glycerol SDS β-mercaptoethanol Bromophenol Blue Dye	5 ml 25 % (w/v) 10 % (w/v) 2.5 ml 0.1 % (v/v)

2.1.23. Buffer exchange

Following the SDS-PAGE gel visualisation, the TALON + Imidazole fractions containing the required soluble protein were buffer-exchanged in small volumes using dialysis tubing with a molecular weight cut-off of 12kDa and a required buffer. Sodium Phosphate 20mM pH 7 buffer was routinely used in this study. The sample was loaded into the dialysis tubing, sealed with clips and submerged into 4 L of dialysis buffer. The sample was dialysed overnight at 4°C, with stirring, to allow for buffer exchange to take place.

2.1.24. Concentration of proteins

Proteins were concentrated using 2 mL of 20 mL Vivaspin (Sartorius) or Amicon Ultra (Merck) centrifugal concentrators with 5, 10, 30, 50 or 100 kDa molecular weight cut-off filters as required. Heraeus Megafuge 16R (Thermo Fisher Scientific) centrifuge with a swing-out rotor was used to concentrate the samples for 20-40min at 10 °C.

2.1.25. Size exclusion chromatography

Proteins that required high purity, such as those used in crystallography, were further purified using a size exclusion chromatography. ÄKTA Pure liquid chromatography system (GE Healthcare) was used with a HiLoad 16/600 Superdex 200pg 120 ml gel filtration column (GE Healthcare). Prior the use, the gel filtration column was equilibrated in the required buffer (usually buffer A) 10mM Tris pH 8 and B) 10mM Tris, NaCl 1M pH 8) and <5 ml of concentrated protein sample (not dialysed) was injected into the machine. The proteins were separated and purified through the column at a flow rate of 1.2 ml/min, typically eluting at 80-85 minute time point. 2 ml fractions were collected and analysed using the SDS-PAGE gels.

2.1.26. Determination of protein concentration

NanoDrop 2000 benchtop spectrophotometer (Thermo Fisher Scientific) was used to determine the protein concentration. The detector was blanked and 2 µl of purified protein sample was loaded. The absorbance at $A_{280\text{-}320nm}$ was measured and noted down. The protein concentration was calculated using the $A = \varepsilon ClD$ equation, where A= absorbance, ε = molar extinction coefficient, C = molar concentration of the sample, l = length of light path in cm and D = dilution factor. The protein extinction coefficient was calculated using the amino acid sequence and the ProtParam bioinformatics tool. Alternatively, the NanoDrop 2000 has a function of calculating the concentration of a known protein solution in mg/ml when you enter the protein molecular weight (kDa) and extinction coefficient.

2.2. Bioinformatics tools

2.2.1. Sequence alignments, homology and phylogenetic trees

Basic Local Alignment Search Tool (BLAST) by NCBI (National Centre for Biotechnology Information) was used to search for amino acid sequences (Altschul *et al*., 1997), hosted by European Bioinformatics Institute (EBI) website (https://www.ebi.ac.uk/).

Alternatively, KEGG: Kyoto Encyclopedia of Genes and Genomes was used to search for sequences of specific genes (http://www.kegg.jp/).

Amino acid sequences were aligned using the Clustal Omega Multiple Sequence Alignment tool also found on EBI website (https://www.ebi.ac.uk/Tools/msa/clustalo/).

Phylogenetic trees were constructed using the amino acid sequences obtained from Uniprot database and phylogeny.fr tool. Phylogeny.fr combines multiple bioinformatics tools, such as MUSCLE, CLustalW, T-Coffee, PhyML, TNT, TreeDyn and Drawtree, to construct a robust phylogenetic tree and analyse the relationships between a set of amino acid sequences.

PULDB and Uniprot databases were used to identify homologous PULs in other *Bacteroides* spp.

2.2.2. Prediction of protein parameters

SignalP 4.1 bioinformatics tool was used to predict the presence of signal peptides in the amino acid sequences and the location of their cleavage site. The server is hosted by DTU Bioinformatics website (http://www.cbs.dtu.dk/services/SignalP/).

ProtParam tool by ExPASy Bioinformatics Resource Portal was used to compute various physical and chemical parameters of the proteins based on their amino acid sequence. These parameters include the molecular weight and molar extinction coefficient (https://web.expasy.org/protparam/).

2.2.3. Genetic analysis tools

Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to investigate potential loci and manually search the genes of interest (http://www.kegg.jp/).

Carbohydrate-Active enZYmes Database (http://www.cazy.org/) was used to investigate predicted enzyme activities and compare them to structurally-related functional and catalytic domains already characterised and published.

Simple Modular Architecture Research Tool (SMART) hosted by European Molecular Biology Laboratory (http://smart.embl-heidelberg.de/) was used to investigate the protein domain composition.

The Integrated Microbial Genomes (IMG) website (https://img.jgi.doe.gov/) was used for analysis and comprehensive comparison of genes and their domains.

2.3. Biochemistry

2.3.1. Isothermal titration calorimetry (ITC)

Isothermal titration calorimetry (ITC) was used to investigate and quantify the thermodynamic parameters of carbohydrate binding to proteins. MicroCal^{TM} VP Isothermal Titration Calorimeter was used to quantify the binding and plot the titration curve along with the MicroCal^{TM} Origin 7.0 software to calculate the thermodynamic parameters – association constant (K_a), stoichiometry of reaction (n) and binding enthalpy (ΔH).

Proteins used for ITC were routinely dialysed in HEPES 50mM pH 7.5 buffer and carbohydrates (ligands) were dissolved in same buffer to avoid reading errors. Protein and ligand concentrations were adjusted and optimised for each reaction as required, usually 50 μM of protein and 10-50mM of ligand. Titrations were performed at 307 rpm with 28 injections of 10 μl of ligand into the cell containing the protein every 200 seconds at constant temperature of 25 °C. Protein-ligand binding results in either absorbance or release of heat (endothermic or exothermic reaction), the ITC machine measures it and plots a titration curve that is used to calculate the thermodynamic parameters using a non-linear regression model. Other parameters were calculated using:

$$-RT\ln Ka = \Delta G = \Delta H - T\Delta S$$

R = gas constant; **ΔG** = Gibbs energy change; **ΔH** = enthalpy change; **T** = absolute temperature in Kelvin; **ΔS** = entropy change

2.3.2. Thin layer chromatography (TLC)

Thin-layer chromatography (TLC) is a solvent-based chromatography technique used to separate and analyse non-volatile mixtures, such as glycans. Due to differences in charge, structure and solubility, different glycans migrate with the solvent up the silica-coated foil plate at different rates, allowing for separation. Typically, smaller sugars migrate faster than larger complexes.

Routinely, a solution of 1-butanol/ acetic acid/ water at 2:1:1 ratio was used as a running buffer/solvent. It was poured into a glass chromatography tank, sealed shut and left to equilibrate for minimum 1 hour prior to use. A silica-coated foil plates (Silicagel 60, 20 cm x 20 cm, Sigma Aldrich) were used in this study. The plates were cut into required size with 1 cm gap from the bottom for the sample loading, and labelled appropriately using a soft pencil. Samples were loaded 1 cm apart from each other, 3-9 µl of final volume per sample, along with the appropriate known sugar standards. The silica plates with loaded samples were put into the equilibrated running buffer and allowed to migrate for approximately 1 hour until solvent reached the top of the plate. For analysis of larger sugar mixtures, plates were taken out, dried and put back in for another hour to allow better separation of the glycans.

To visualise the sugars, dried plates were stained with either Orcinol stain (sulphuric acid/ ethanol/ water 3:70:20 v/v, orcinol 1 %) or Diphenylamine-aniline-phosphoric acid stain (1.7% w/v

Diphenylamine, 1.7% v/v Aniline, 85% v/v Ethyl acetate, 2% Hydrochloric Acid and 11% Phosphoric acid). Plates were dried and developed by heating to 100 °C.

2.3.3. High-performance Liquid Chromatography (HPAEC-PAD)

High-performance liquid chromatography (HPAEC-PAD) was used to assess the activity of enzymes against various carbohydrate substrates. The enzyme reactions were set up in either dH_2O or 20mM NaH2PO4 (pH 7) buffer containing 0.1-0.2% substrate and 1-2µM of enzyme in a final volume of 200µL. All of the substrates were purchased commercially from Sigma-Aldrich, Dextra, Carbosynth and OligoTech. The reactions were incubated for 1h to 16h at 37̊ C. The samples were then boiled and centrifuged at 13 000 x rpm for 10 min. The supernatants were analysed by HPLC along with the known glycan standards to allow for identification.

The Dionex DX500 ICS3000 system along with the the Dionex CARBOPAC^{TM} PA-100 HPLC column were routinely used in this study to investigate the degradation of N-glycans. The pulsed amperometric detection (PAD) was used to detect the sugars (with settings of E1 = +0.05, E2 = +0.6 and E3 = -0.6). Figure 2.5 shows the HPLC program commonly used for N-Glycan separation in this study.

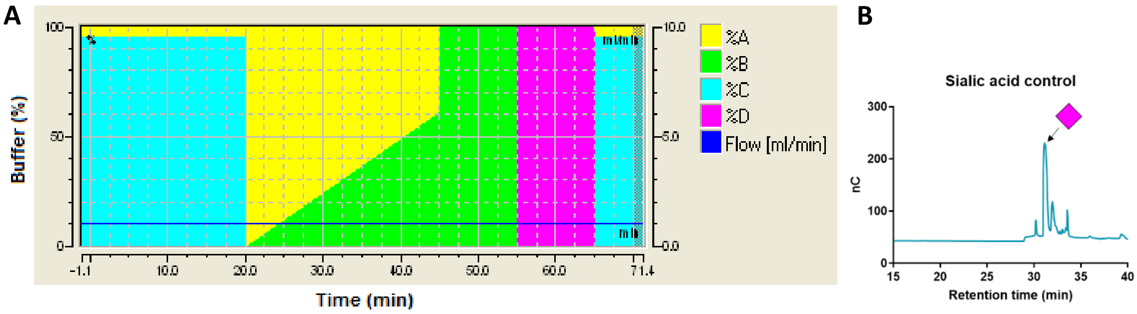


Figure 2.5: Commonly used HPLC program and buffers.
A) HPLC program. Buffer A. 100mM NaOH; B) 100mM NaOH, 1M sodium acetate; C) dH_2O; D) 500mM NaOH. Sugars were eluted with an isocratic flow followed by a gradient of sodium acetate. Following the elution, column was washed with 100% of buffer D and equilibrated with 95% of buffer C and 5% of buffer A. **B)** Example of a sialic acid (Neu5Ac) standard elution. GraphPad Prism 7.0 was used to make this graph.

2.3.4. Microscale Thermophoresis (MST)

Microscale Thermophoresis (MST) was used to investigate and quantify the bio-molecular interactions between the carbohydrate-binding modules of proteins and N-glycan substrates. The technique is based on thermophoresis – a movement of molecules induced by a variety of molecular properties, such as change in hydration shell, conformation, size or charge. It is highly sensitive technique that uses a temperature gradient induced by an infrared laser to detect and quantify the movement of fluorescently labelled molecules, thus allowing to measure the dissociation constants and determine thermodynamic parameters of protein-substrate interactions (Jerabek-Willemsen *et al.,* 2011).

In this study, microscale thermophoresis (MST) was done using Monolith NT.115 machine (NanoTemper) and Monolith NT.115 Standard Glass Capillaries. The proteins were purified using protocol described previously, dialysed in HEPES 50 mM pH 7.5 buffer for 4 hours at room temperature, concentrated using 30k cut-off concentrator, desalted using ZEBA desalting columns (Sigma Aldrich) and labelled using the Alexa Fluor 647 dye overnight at 4°C. Following the incubation, the labelled protein was centrifuged using a 5K cut-off concentrator at 3500 x rpm for 10 minutes to remove unbound dye. To ensure the purity, the protein was desalted again using ZEBA desalting columns. The purity of labelled protein was assessed by measuring the ratio of protein/dye using a spectrophotometer. Protein absorbance was measured at 280nm and dye absorbance was measured at 650nm (Molar absorbance: 250,000 M-1 cm-1). Optimal ratio is 1:1. The concentration of labelled protein was determined, protein sample was split into 10 µL aliquots and flash-frozen at – 80 °C until use.

During the MST experiment, dilution series of 16-fold of unlabelled substrate (N-glycan, 6 mM stock) was prepared. The concentration of the fluorescently labelled of proteins was kept constant at 100 nM. 10 µL of ligand was mixed with 10 µL of fluorescently-labelled protein and loaded onto the glass capillaries. The samples were then put onto a magnetic rack and loaded into the Monolith NT.115

machine. Depending on the fluorescence of the sample, 50-100% of LED power was used and fluorescence was measured. The collected data was then analysed using the MO Affinity Analysis software and plotted using GraphPad Prism 7.1 to calculate the dissociation constant Kd. Figure 2.6 shows the principle behind the microscale thermophoresis (MST).

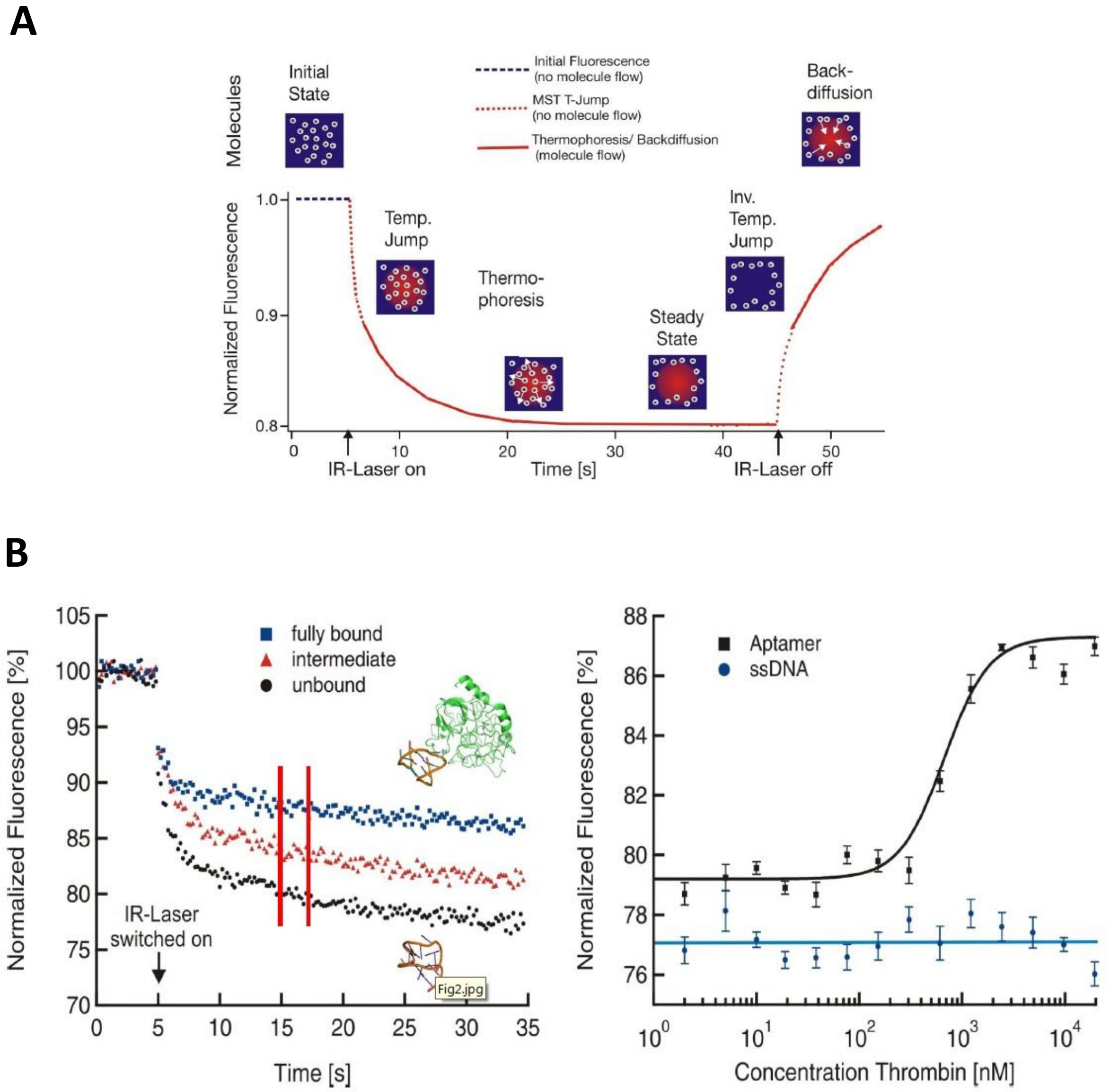


Figure 2.6: A graphical representation of the principle behind the Microscale Thermophoresis (MST). A) A figure showing a typical thermophoretic curve. During the initial state, fluorescent dye is evenly distributed throughout the capillary. Once the laser is shot through the capillary, increase in the heat decreases the heat-sensitive fluorescence levels. During the thermophoresis, fluorescently-labelled molecules move out of the laser focal point due to gradient created by the heat and reaches the steady state. When the laser is turned off, heat-sensitive fluorescent molecules move back to the laser focal point and fluorescence increases. **B)** An example of thrombin-heptamer interaction. The graph displays unbound, intermediate and fully bound fluorescently-labelled thrombin molecules (left image). The unbound and fully bound baselines, including their concentrations, are used to calculate the affinity (right image). The figure was adapted from Monolith NT. 115 online instruction manual.

2.3.5. Enzymatic assays

All enzyme assays, unless otherwise stated, were done using 1 µM of enzyme and 20 mM sodium phosphate pH 7 buffer, at 37 °C for 1 h to 16 h. The assays were usually repeated three times, to ensure the reproducibility of the result. The enzyme assay products were investigated using either TLC, HPLC (described previously) or mass spectrometry.

For catalytic activity assays, 100 mM sodium phosphate pH 7.5 buffer was routinely used. Both the substrate mixture and enzyme were pre-warmed to 37 °C prior initiation of the reaction. The assays were also performed in technical triplicates. The enzyme assay results were plotted in GraphPad Prism 7.1 and used to calculate slopes or Michaelis-Menten kinetics model was used to estimate K_M and calculate k_{cat} using the formulas:

$$\mathrm{V0} = \frac{\mathrm{Vmax}\,[\mathrm{S}]}{\mathrm{Km}+[\mathrm{S}]} \quad \text{or} \quad \mathrm{V0} = \frac{\mathrm{kcat}}{\mathrm{Km}} \times [S] \times [E]$$

V_0= initial velocity; V_{max} = rate of reaction when enzyme is saturated with substrate; K_m= substrate concentration at which reaction rate is $1/V_{max}$; [S] = substrate concentration; k_{cat}= number of substrate molecules catalysed per molecule of enzyme; [E] = enzyme concentration.

2.3.5.1. Colorimetric *p*-Nitrophenyl Phosphate (*p*NP) assays

p-Nitrophenyl (*p*NP)-linked substrates were used to analyse and measure the activity of various glycoside hydrolases. Initially, *p*NP-linked substrates were dissolved in dH_2O (to the final reaction concentration of 1mM) and mixed with the enzyme of interest (1 µM final concentration) and incubated at 37°C for 1hour. If the enzyme showed activity on the linked substrate, the colour of the reaction turned yellow.

To measure the rate of the substrate hydrolysis, a final concentration of 0.1-1 mM of the substrate was dissolved in 20mM NaH_2PO_4 pH 7.5 buffer and dH_2O. The reactions were set up in 600 µL quartz cuvettes and incubated at 37° for 2-5 minutes. Various known concentrations of enzyme were added to the cuvettes and substrate release was monitored using Pharmacia Ultrospec 4000 spectrophotometer at 420 nm. The data was fitted into Michaelis-Menten kinetics model using GraphPad Prism 7.1 and enzyme kinetic parameters were calculated using method described above. *p*-Nitrophenol extinction coefficient is 18,000 M-1 cm-1.

2.3.5.2. Commercial monosaccharide release detection kits

The commercially available kits by Megazyme were used to monitor the galactose or mannose release. D-Mannose/D-Fructose/D-Glucose kit and L-Arabinose/D-Galactose detection kits were used in this study. Both of these linked-assay kits utilise dehydrogenases to catalyse the oxidation of sugars and conversion of NAD+ into NADH in 1:1 molar ratio. The reactions were set up following the manufacturer's protocol in 600 µL quartz cuvettes. 20mM NaH_2PO_4 pH 7.5 buffer was also used in these experiments. The reactions were monitored using Pharmacia Ultrospec 4000 spectrophotometer at 340 nm, a wavelength NADH absorbs at. GraphPad Prism 7.1 and Michaelis-Menten model were also used to calculate the enzyme kinetic parameters. NADH extinction coefficient is 6230 M^{-1} cm^{-1}.

2.3.5.3. Enzyme kinetics measurement using HPLC

High-performance liquid chromatography (HPLC; HPAEC-PAD) was used to monitor enzyme catalytic activity on N-glycan substrates. Enzyme assays were set up in 20 mM sodium phosphate buffer, pH 7.5 with 1 µM of final enzyme concentration and 10 mg/ml of final N-glycan concentration. These enzyme reactions were incubated at 37 °C for up to 16 h. 30 µl of reaction were taken at required time points (e.g. 0 min; 5 min; 10 min; 20 min; 40 min; 60 min and 16h), boiled and centrifuged at 13,000 x rpm for 10 minutes before being loaded onto HPLC vials in 10-fold dilution in MiliQ water. Samples were analysed using the same method described previously (2.3.3). Data was collected using Chromeleon^{TM} Chromatography Management System (Dionex) by measuring the peak area*min. Known standards were used to determine which peaks correspond to which monosaccharides. To ensure the HPLC detection sensitivity remained constant, 0.05mM glucose standards were run in-between the samples and peak areas were measured.

In some cases, known concentrations of required monosaccharides were also assayed on HPLC, peak areas were measured and the standard curve was plotted using the GraphPad Prism 7.1. This allowed to determine and calculate the concentration of monosaccharides released off N-glycans by enzymes tested.

2.3.6. Acid Hydrolysis of Carbohydrates

Acid hydrolysis of various N-glycans was done to determine their approximate composition. Required N-glycan was hydrolysed using 100 mM HCl at 100 °C for 1 h. Following the incubation, the sample was cooled down and neutralised by addition of NaOH. Samples were analysed using TLC and HPLC against known sugar standards.

2.3.7. Preparation and purification of N-glycans

Heavily N-glycosylated proteins, such as bovine serum fetuin and bovine serum alpha$_1$-acid glycoprotein were used as sources for N-glycan purification. PNGaseF from *Elizabethkingia meningoseptica* (Sigma Aldrich) was used to cleave whole N-glycans off protein backbone. It acts by hydrolysing the bond between the reducing-end GlcNAc residue of an N-glycan structure and an asparagine residue of the protein backbone. 20mg/ml of N-glycoprotein was dissolved in 50mM NaH_2PO_4 pH7.5 buffer and 10U final concentration of PNGaseF was added. The reaction was incubated at 37 °C for 16h to ensure all of the N-glycan has been cleaved off. Sample was then run on SDS-PAGE (1:50 dilution) alongside with undigested N-glycoprotein to check deglycosylation was successful. The sample was then boiled to inactivate PNGaseF, centrifuged at 13,000 x rpm for 30 minutes and supernatant containing the N-glycans was collected. The supernatant was then further concentrated using 5K cut-off centrifugal concentrator at 3500 x rpm for 10 minutes to ensure no protein is left over. The N-glycan purity was checked on SDS-PAGE and TLC before the sample was freeze-dried using the Christ Alpha 1-2 Freeze Drier at -60 °C. N-glycans were then stored at -20 °C until required.

Trypsin from porcine pancreas (Sigma Aldrich) was used to digest glycoproteins according to the manufacturer's instructions.

2.3.8. Mass spectrometry and HPLC analysis of procainamide-labelled N-Glycans (Ludger)

Enzyme assays of 1 µM final concentration of required enzymes and 10mg/ml final of required N-glycoprotein were set up in 20mM NaH_2PO_4 pH 7 buffer and incubated at 37 °C for 1 - 16 h, as required. The samples were then boiled for 10 minutes to inactivate the enzymes and labelled with fluorescent procainamide label (Ludger) according to the manufacturer's protocol.

The samples were then sent to Paulina Urbanowicz who works for our industrial partner, Ludger Ltd., a glycotechnology company. Procainamide-labelled N-glycans were analysed using LC-ESI-MS (liquid chromatography - electrospray ionization- mass spectrometry), a highly-sensitive technique that combines physical separation capabilities of Ultra-High Performance Liquid Chromatography with the subsequent mass spectrometry analysis. Mass spectrometry data analysis software COMPASS was used to analyse the data and determine the N-glycan structures present in the samples.

2.3.9. Western Blot

Western blot was used to detect specific proteins in the *B. thetaiotaomicron* cell samples. Samples were run on SDS-PAGE gel, alongside of MagicMark XP Western Protein Standard (Thermo Fisher Scientific), using a protocol described previously. The proteins run out on the gel were then transferred onto Amersham Protran 0.45 Nitrocellulose membrane (GE Healthcare) using the Mini Trans-Blot system (Bio-Rad) and a chilled transfer buffer (Table 2.12). The blotting was done for 90 minutes at 80 V, using a freezing block and a magnetic stirrer to keep the transfer buffer cool.

Blocking buffer was prepared as listed in Table 2.12 and the membrane was blocked overnight at 4 °C, with shaking. Primary Anti-Flag antibody (Custom made, Eurogentec) was then diluted 1:2000 in 10 ml of antibody buffer and incubated with the membrane for 1 hour at room temperature, with shaking. The membrane was then washed three times with the wash buffer (Table 2.12). Secondary Anti-Flag Goat Anti-Rabbit-HRP antibody (SC-2004, Santa Cruz) was diluted 1:5000 in 10 ml of antibody buffer and incubated with the membrane for 1 hour at room temperature, with shaking. The membrane was washed three times again.

The membrane was then developed using 3 mL of 1:1 mixture of chemi-luminescence Clarity Western ECL reagents (Bio-rad) and visualised using ChemiDoc XRS+ System (Bio-Rad).

Table 2.12: Western blot buffers

Buffer name	Reagents	Amount used	Concentration
Transfer Buffer **(1 litre)**	Tris-HCl Glycine Methanol SDS dH2O	3 g 14.3 g 300 ml 0.5 g 200 ml	25 mM 192 mM 30 % 1.7 mM -
Blocking Buffer **(10 ml)**	1 x PBS tablet (Oxoid) Milk powder Tween20	10 ml 0.5 g 50 µl	- 5 % 0.5 %
Antibody Buffer **(10 ml)**	1 x PBS (Oxoid) Milk powder	10 ml 0.5 g	- 5 %
Wash Buffer **(50 ml)**	1 x PBS (Oxoid) Tween20	50 ml 250 µl	- 0.5 %

2.3.9.1 Cellular Localization

3 x 5 ml cultures of *Bacteroides thetaiotaomicron* were cultured in minimal media on 10 mg/ml of bovine serum alpha$_1$-acid glycoprotein as a sole carbon source. The bacterial cultures were allowed to grow to mid-exponential growth phase (OD$_{600nm}$ of 0.6) and were harvested by centrifugation at 3,500 x rpm for 5 minutes. The supernatant was discarded, the bacterial cells were washed twice in 5 ml of Phosphate Buffered Saline (PBS) and re-suspended in 2 ml of the same buffer. The cell mixture was then split into four 0.5 ml fractions. 2 mg/ml of Proteinase K solution was added to 3 of these fractions and 1 was left undigested as a control. Proteinase K is a protease that digests all proteins that are present on a cell surface, leaving intracellular proteins intact. The samples were incubated at 37 °C for 16 hours, with shaking. After the incubation the samples were centrifuged at 3,500 x rpm for 10 minutes and supernatant was discarded. The cell pellets were re-suspended in 0.5 ml of PBS and Proteinase K was inactivated by the addition of 200 µl of trichloroacetic acid (TCA). The solution was incubated for 30 minutes on ice. After the incubation, the precipitated proteins were centrifuged at 3,500 x rpm for 10 minutes and washed four times in 1 ml of ice-cold acetone. Following the wash, the cell pellets were re-suspended in 0.5 ml of PBS buffer and ran on SDS-page gel. Western blot protocol, as described previously, was used to detect whether proteins of interest are located on a cell surface or inside of the cell. If the protein is located outside of the cell, Proteinase K treatment would have digested it and no protein band would be visible on the Western Blot.

2.3.9.2 Protein expression levels

Similar method as described above was to investigate the specific protein expression levels. 3 x 5 ml cultures of *B. thetaiotaomicron* were grown in minimal media on 1% of bovine serum alpha1-acid glycoprotein and 3 x 5 ml cultures were grown on 0.5% of glucose as sole carbon sources. The cells were grown to mid-exponential growth phase (OD$_{600nm}$ of 0.6) and harvested by centrifugation. The cells were re-suspended in 1 ml of PBS and run on SDS-PAGE gel. Western blot was run using the protocol described previously. Protein expression levels were compared between the *B. theta* cells grown on glucose as sole carbon source to those on alpha1-acid glycoprotein, which is heavily N-glycosylated. Thicker bands visualised by the Western Blot corresponded to higher expression levels.

2.3.9.3. Determination of protein is secreted outside of the cell

Same growth protocol as described in section 2.3.9.3 was used to grow the bacterial cells, but instead of discarding the growth supernatant, it was concentrated to the same volume as whole cells and run on SDS-PAGE gel alongside of the cell pellets. Western blot was done according to the protocol described previously. If the protein is secreted outside of the cell, bands corresponding to the protein of interest would be detected in the supernatant sample.

2.3.10. Whole Cell Assays

B. thetaiotaomicron was inoculated in 4 x 5 ml of minimal media containing 2% of α_1-acid glycoprotein as a sole carbon source and grown to the mid-exponential growth phase. When the cells reached the OD_{600nm} of approximately 0.6, they were gently harvested using centrifugation at 3,500 x rpm for 5 minutes. The supernatant was collected and the cell pellet was washed in PBS.

Whole cell assays were used to investigate the catalytic activity of enzymes and their cellular localization*. B. thetaiotaomicron* is an obligate anaerobe, thus in the presence of oxygen, cells retain structural integrity but are metabolically inactive. Whereas cell-surface proteins which are not ATP-dependent, such as carbohydrate-active enzymes, remain functionally active and can be investigated.

The washed bacterial cells were re-suspended in PBS and split into four fractions. A fraction of whole cells was incubated with 2% of α_1-acid glycoprotein in PBS in a ratio of 1:1. One fraction was left untreated, one was boiled for 10 minutes as a control and one fraction was sonicated. Supernatant was incubated with 2% of α_1-acid glycoprotein. To stop these reactions, samples were centrifuged in order to remove the cells and the supernatant was boiled for 10 minutes. These samples were analysed using the thin-layer chromatography (TLC).

2.3.11. Supernatant Analysis

In order to investigate whether *B. thetaiotaomicron* can utilise all of the N-glycans present on the bovine serum α_1-acid glycoprotein, the cells were grown in minimal media containing 1 % of α_1-acid glycoprotein as a sole carbon source for 24 hours. The cells were then centrifuged at 3,500 x rpm for 5 minutes and supernatant was collected. The supernatant was analysed by TLC and HPLC for traces of undigested N-glycans.

2.4. Crystallography

2.4.1. Protein crystallisation trial setup

Soluble proteins for crystallography trials were purified using TALON IMAC columns, following protocol described in section 2.1. Proteins were then concentrated using 30K cut-off concentrator at 3,500 rpm for 20-30 minutes and further purified using size exclusion chromatography (SEC) (section 2.1). SEC fractions were collected and analysed by SDS-PAGE before pooling fractions containing purest protein and concentrating to approximately 30 mg/ml. When required, protein was mixed with a substrate of choice for co-crystallisation.

Initial protein crystallisation screens were set up using the sitting-drop vapour-diffusion method and MRC 96-well crystallisation plates (Molecular Dimensions). Pre-prepared crystallography screening plates were used. These included Index screen from Hampton Research and Structure, JCSG+, Morpheus and PACT screens from Molecular Dimensions. Mosquito Crystal Nanolitre dispensing robot (TTP Labtech) was used to set up crystal plates by dispensing 200 nl + 200 nl and 200 nl + 100 nl of protein and crystal screen solution, respectively. Once plates were set up, they were sealed and incubated at 20 °C.

2.4.2. Crystal screen optimisation

The screen conditions that crystals formed in were optimised and set up using hanging-drop vapour diffusion method (Figure 2.7) and VDX 24-well hanging drop plates (Molecular Dimensions). Modified original screen conditions with varied precipitant and salt concentrations were set up using 1 µl of protein + 1 µl of crystal screen solution along with 1 µl of protein + 2 µl of crystal screen solution. As before, the plates were sealed and incubated at 20 °C.

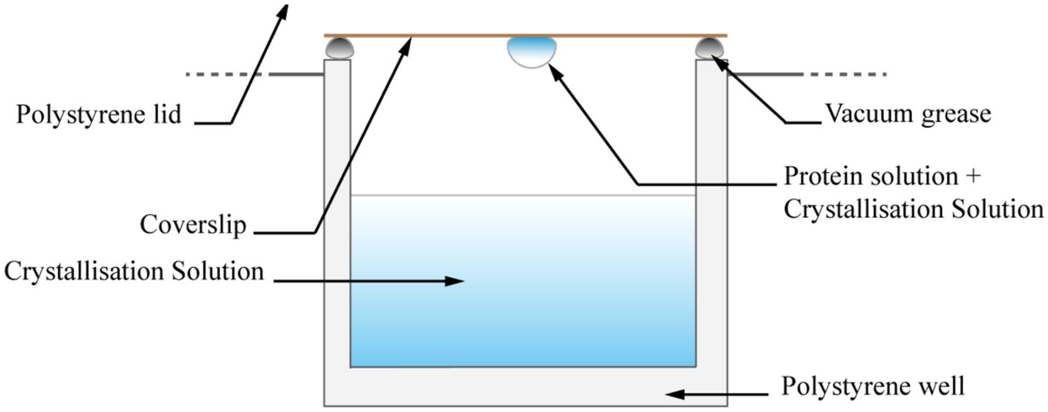


Figure 2.7: Graphical representation of hanging-drop crystallisation plate set up.
The well containing the crystallisation solution and a hanging-drop of protein + crystallisation solution mix was sealed using vacuum grease and a glass coverslip. Figure was adapted from Sarah Shapiro's thesis.

2.4.3. Crystal structure solving and visualization

With the assistance of Dr. Arnaud Beslé, the crystals were harvested using cryo-protected suitably sized loops (Hampton) and flash frozen in liquid nitrogen. In-house screening of crystal diffraction was done using Rigaku MSC microfocus generator and Raxis IV++ image plate detector. Most suitable crystals, chosen by their resolution and diffraction pattern were taken for data collection at Diamond Light Source (DLS, Didcot, Oxfordshire, UK).

The structure of BT0459 (GH20-family enzyme) was solved using a molecular replacement by Dr. Arnaud Beslé.

Protein structures were manipulated and visualised using WinCoot and PyMOL softwares.

Chapter 3: Complex N-Glycan metabolism by members of the human gut microbiota

3.1. Introduction and motivation for the research

The last two decades of human microbiome research has produced a tremendous amount of data that collectively reveals the significant impact the gut microbiota has on human metabolism, physiology and overall health (Sender *et al.,* 2016). Previous studies have shown that *B. thetaiotaomicron* is one of the few endogenous gut microbes that is capable of growing on host-derived mucins in the absence of dietary carbohydrate input. Mucins are heavily O-glycosylated proteins that provide structural and functional basis for the mucus layer (Martens *et al.,* 2008).

However, little interest has been focused towards the understanding of the underlying processes of dietary and host-derived N-glycan degradation by the gut microbiota. Glycosylation patterns of N-glycoproteins are complex and diverse, thus would require an extensive metabolic machinery to fully degrade and utilize. The gut microbiota produces numerous enzymes that allow degradation of various carbohydrate structures, but little information is available on mechanisms gut microbes employ to degrade N-glycan structures found on various dietary and host-derived glycoconjugates. Understanding the precise mechanisms employed by gut microbiota to degrade these highly complex glycans is important for boosting our comprehension of how it benefits human health.

3.2. Objectives

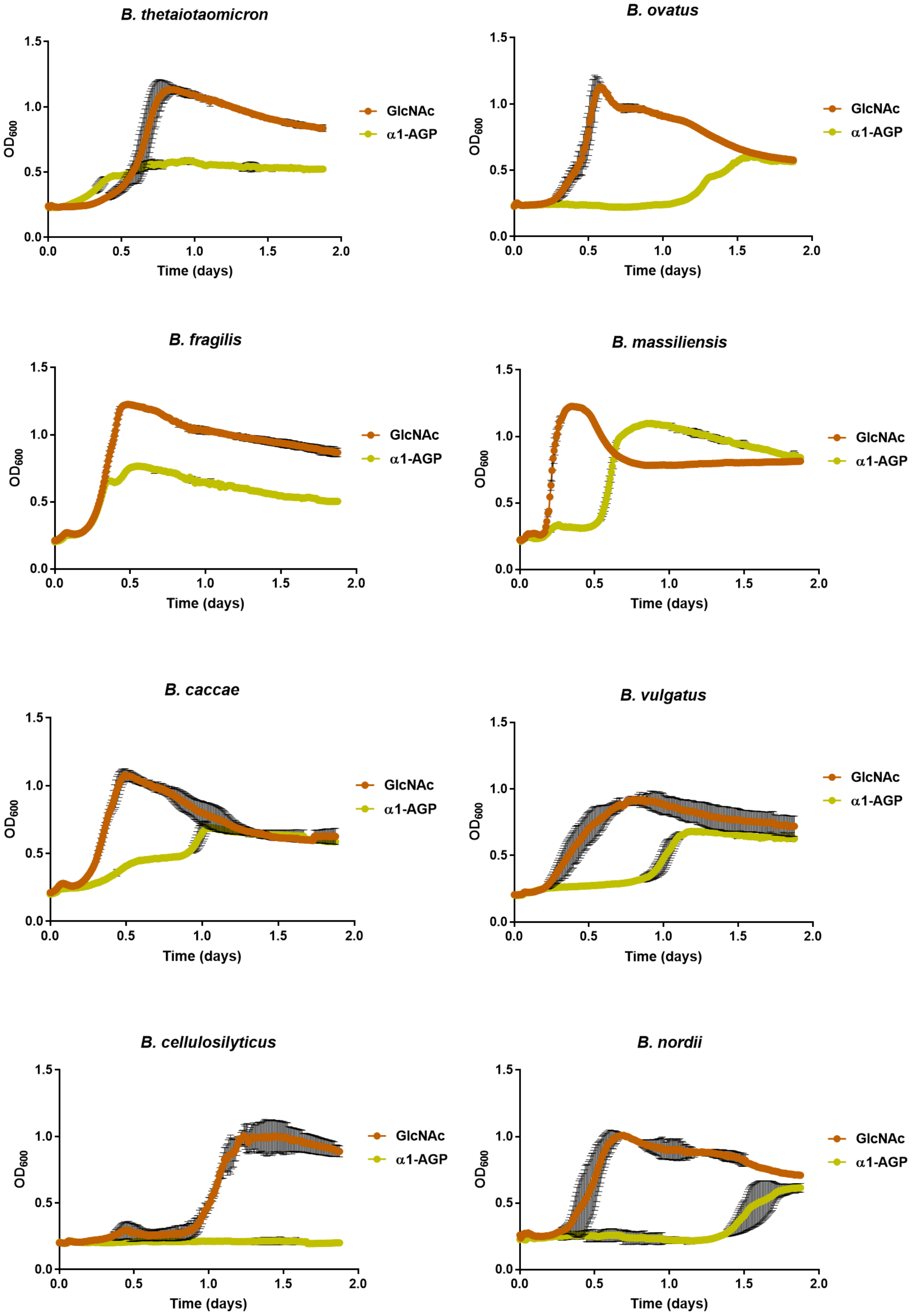
i. To investigate the capacity of gut Bacteroidetes to utilise complex N-glycans

ii. To investigate potential cooperation of multiple polysaccharide utilization loci (PULs) required for N-glycan degradation

3.3. Results

3.3.1. Prominent members of gut microbiota can utilize N-glycans as sole carbon sources

In a study done by Martens *et al.* (2009), a dominant member of human gut microbiota, *B. thetaiotaomicron* was shown to be capable of growing by utilising porcine gastric mucin III (PGMIII) as a sole carbon source, *in vitro*. PGMIII is a large, heavily O-glycosylated mucin present in the mucus layer of the gut. In this study, bacterial growth assays were set up to find out whether prominent members of human gut microbiota, including wild type *Bacteroides thetaiotaomicron,* are also capable of utilising heavily N-glycosylated proteins as sole carbon sources. α_1-Acid Glycoprotein (α_1-AGP), one of the major plasma proteins found in mammals, was chosen as an N-glycan substrate. The bacteria were grown in minimal media+hematin containing either 2% of bovine α_1-Acid Glycoprotein (α_1-AGP) or 1% of N-Acetylglucosamine (GlcNAc) as sole carbon sources following protocols described in Sections 2.1.2 and 2.1.19. The growths assays were performed twice in triplicate to minimise the risk of any errors and contamination. These data show that several members of gut microbiota can utilise α_1-Acid Glycoprotein (**Figure 3.1**). Compared to growth on N-Acetylglucosamine, some bacteria displayed a longer lag phase while growing using α_1-Acid Glycoprotein, suggesting they may not be specialised in N-glycan degradation. These include *B. ovatus, B. nordii, B. salyeriae, B. fluxus, B. langum* and *B. uniformis*. In comparison, some bacteria, such as *B. cellulosilyticus*, *B. intestinalis* and *P. merdae*, could not utilise α_1-Acid Glycoprotein at all. Finally, this data allowed us to identify bacterial species that are capable of efficiently utilizing complex N-glycan structures found on α_1-Acid Glycoprotein. These include wild type *B. thetaiotaomicron, B. fragilis, B. massiliensis and B. uniformis*.


B. thetaiotaomicron
B. ovatus
B. fragilis
B. massiliensis
B. caccae
B. vulgatus
B. cellulosilyticus
B. nordii
OD_{600}
0.0
0.5
1.0
1.5
Time (days)
0.0
0.5
1.0
1.5
2.0
GlcNAc
α1-AGP

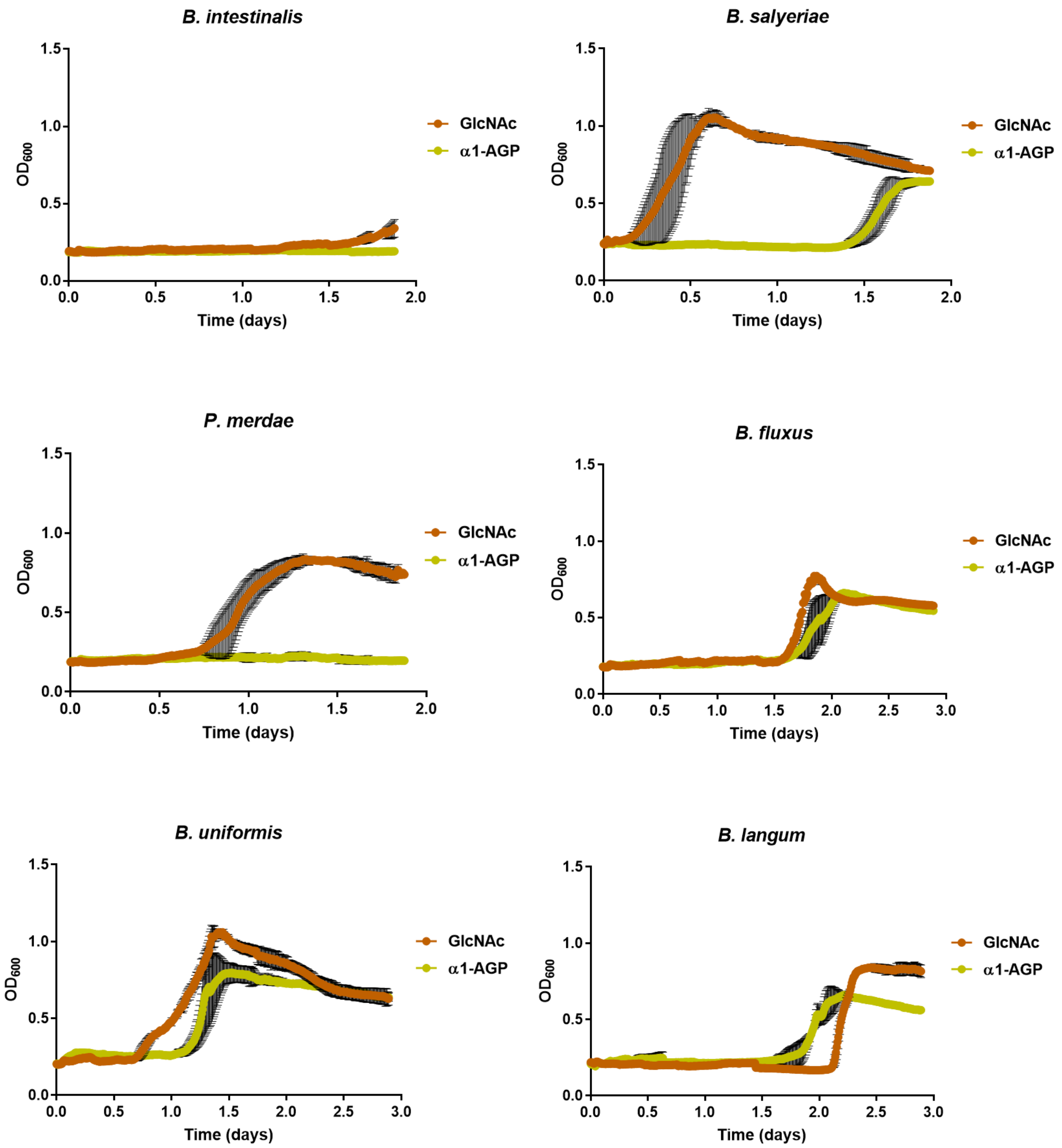


Figure 3.1: Growth curves of gut Bacteroidetes on α$_1$-Acid glycoprotein and GlcNAc. The bacterial growth assays on minimal media+hematin containing 2% bovine α$_1$-Acid Glycoprotein (α$_1$-AGP) and 1% N-Acetylglucosamine (GlcNAc) were monitored by measuring the optical density at 600nm using an automated plate reader, anaerobically at 37 °C. The growth assays were initiated by inoculating the media+substrate with 10µl of a required bacterium culture grown on 1% GlcNAc overnight, anaerobically at 37 °C. The bacterial growth assays were performed twice in triplicate and the mean data presented. The error bars indicate the standard deviation from the mean.

3.3.2. N-glycan utilization by *B. thetaiotaomicron*

Wild type *Bacteroides thetaiotaomicron,* one of the prominent members of the human gut microbiota, was chosen as a model organism to study the N-glycan degradation by commensal gut microbes further because it is easy to cultivate and manipulate. In order to investigate whether *B. thetaiotaomicron* displays the same N-glycan utilisation pattern as was seen when it was grown on bovine α_1-AGP (see Section 3.1.3.1), other commercially available heavily N-glycosylated proteins were acquired (Sigma Aldrich). These include bovine fetal serum fetuin and human blood transferrin.

Wild type *B. thetaiotaomicron* was grown in minimal media+hematin containing an appropriate concentration of N-glycoprotein substrate or 1% glucose as a control, in triplicates, following a protocol described in Section 2.1.2 (**Figure 3.2**). Glycosylation patterns of these N-glycoproteins are very different and may vary in every commercially-produced batch, therefore different N-glycoprotein concentrations ranging from 0.5% to 4% were assayed to find optimal growth conditions. Transferrin was digested using trypsin, a protease found in the digestive system, in order to mimic the state of some dietary N-glycoproteins gut microbiota would see. Fetuin was treated with PNGaseF to cleave only the N-glycans off the protein backbone. PNGaseF is unable to cleave O-glycans, so using this enzyme enabled us to investigate only the N-glycan utilisation by the bacterium. Only the N-glycan fraction was used for the growths. Growths were monitored automatically using an Epoch^{TM} Microplate Spectrophotometer, measuring optical density of the cultures at 600 nm every 15 minutes (see Section 2.1.19). The results confirmed that *B. thetaiotaomicron* can efficiently utilise N-glycans. *B. thetaiotaomicron* displayed a tri-phasic growth pattern when grown on N-glycans derived from bovine fetal serum fetuin, suggesting that multiple polysaccharide utilization loci (PULs) may be activated during the utilization.

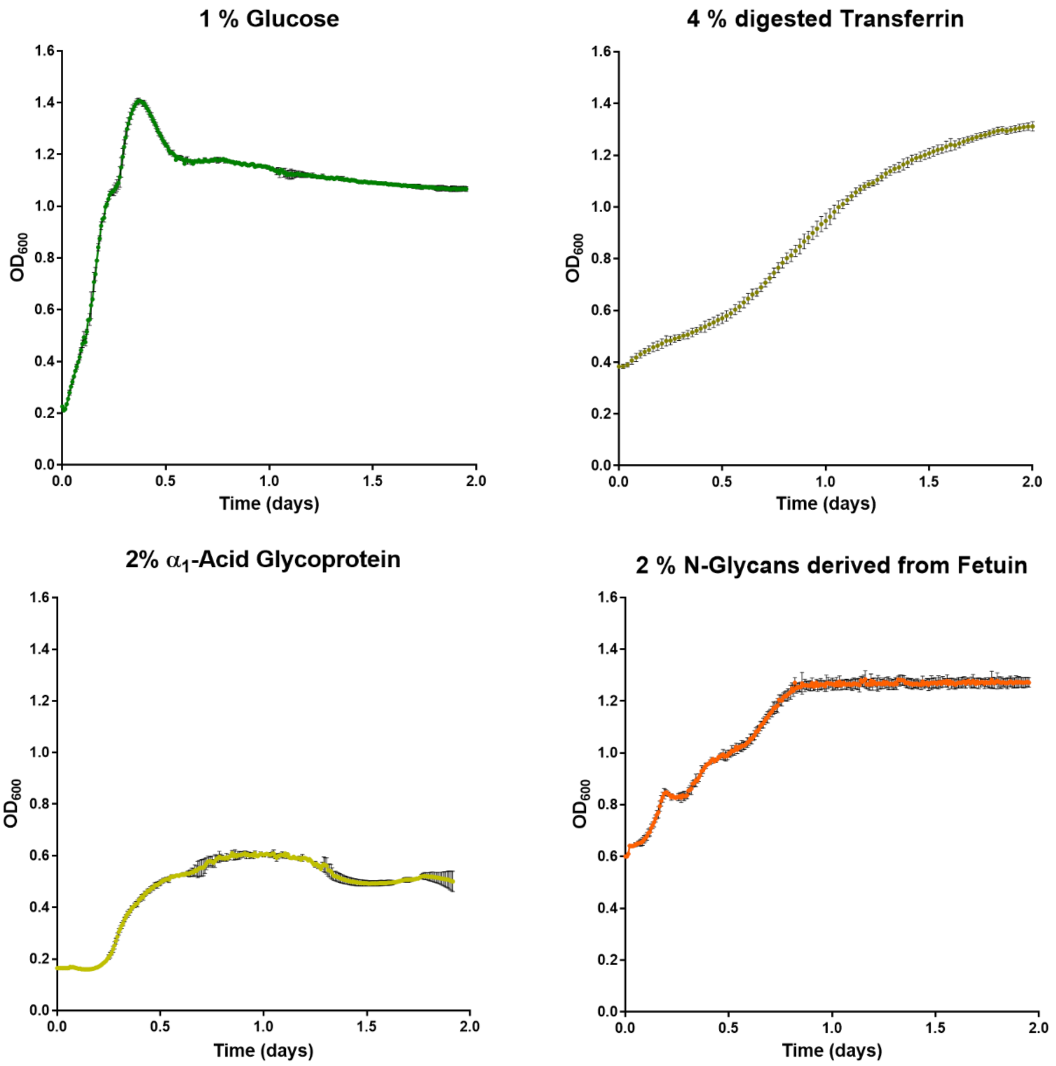


Figure 3.2: Comparing the growth curves of wild type *B. thetaiotaomicron* on selected N-glycoproteins. The bacterium was cultured in minimal media containing either 1% glucose, 4% trypsin-digested transferrin, 2% α_1-Acid Glycoprotein or 2% PNGaseF-derived N-glycans from Fetuin. Growth data for each condition was measured using an automated plate reader. The bacterial growth measurements were performed in triplicate. The error bars indicate the standard deviation from the mean.

B. thetaiotaomicron can utilise all N-glycan sources relatively well compared to the glucose control. Although it grows to a lower optical density on N-glycans compared to the glucose, it should be pointed out that N-glycans can have very complex structures that could require the activity of multiple binding proteins and degrading enzymes to fully degrade and utilize, thus N-Glycans are much more metabolically taxing to utilize than simple sugars such as glucose.

The α_1-AGP was chosen as a model N-glycoprotein substrate to further investigate the N-glycan degradation due to its commercial availability and high bi-antennary N-glycan structure content.

In order to investigate whether wild type *B. thetaiotaomicron* is utilising the N-glycans as a sole carbon source and does not require to degrade the protein fraction to sustain growth, larger-scale growth assays were set up. *B. thetaiotaomicron* was grown in minimal media+hematin containing 2% of native bovine α_1-AGP, 1 % glucose as control and 2% of PNGaseF-digested protein-only fraction of bovine α_1-AGP (see Section 2.1.2). Optical density at 600 nm was measured every hour and a sample of the culture was collected. The samples were assayed using SDS-PAGE gel to visualise protein de-glycosylation and thin-layer chromatography (TLC) to visualise the N-glycan degradation products. A separate PNGaseF-digestion assay was set up to confirm that the fully N-glycosylated bovine α_1-AGP is 41-44 kDa in size whereas fully de-glycosylated bovine α_1-AGP is 23 kDa in size (data not shown, see Section 2.3.7). The growth results show *B. thetaiotaomicron* is utilising N-glycans without digesting the protein fraction (**Figure 3.3**). De-glycosylation of the bovine α_1-AGP was initially observed on the SDS-PAGE gel (**Figure 3.3-B**). A change in size from fully N-glycosylated ~41 kDa protein to fully de-glycosylated ~23 kDa protein band was noted, suggesting it was not due to proteolytic activity. The observed degradation of the glycoprotein and the appearance of differently-sized bands could be representative of variably N-glycosylated protein. No growth was observed in protein-only fraction, confirming that *B. thetaiotaomicron* is utilising only the N-glycan fraction of α_1-AGP for the growth. N-glycan degradation was further observed on the TLC plate where based on the appearance of the sialic acid band, the N-glycan degradation was observed over time (**Figure 3.3-C**). *B. thetaiotaomicron* releases free sialic acid, but does not possess Nan operon required to process it, presumably in order to access the carbohydrate structures that the sialic acids cap (Juge *et al.,* 2016).

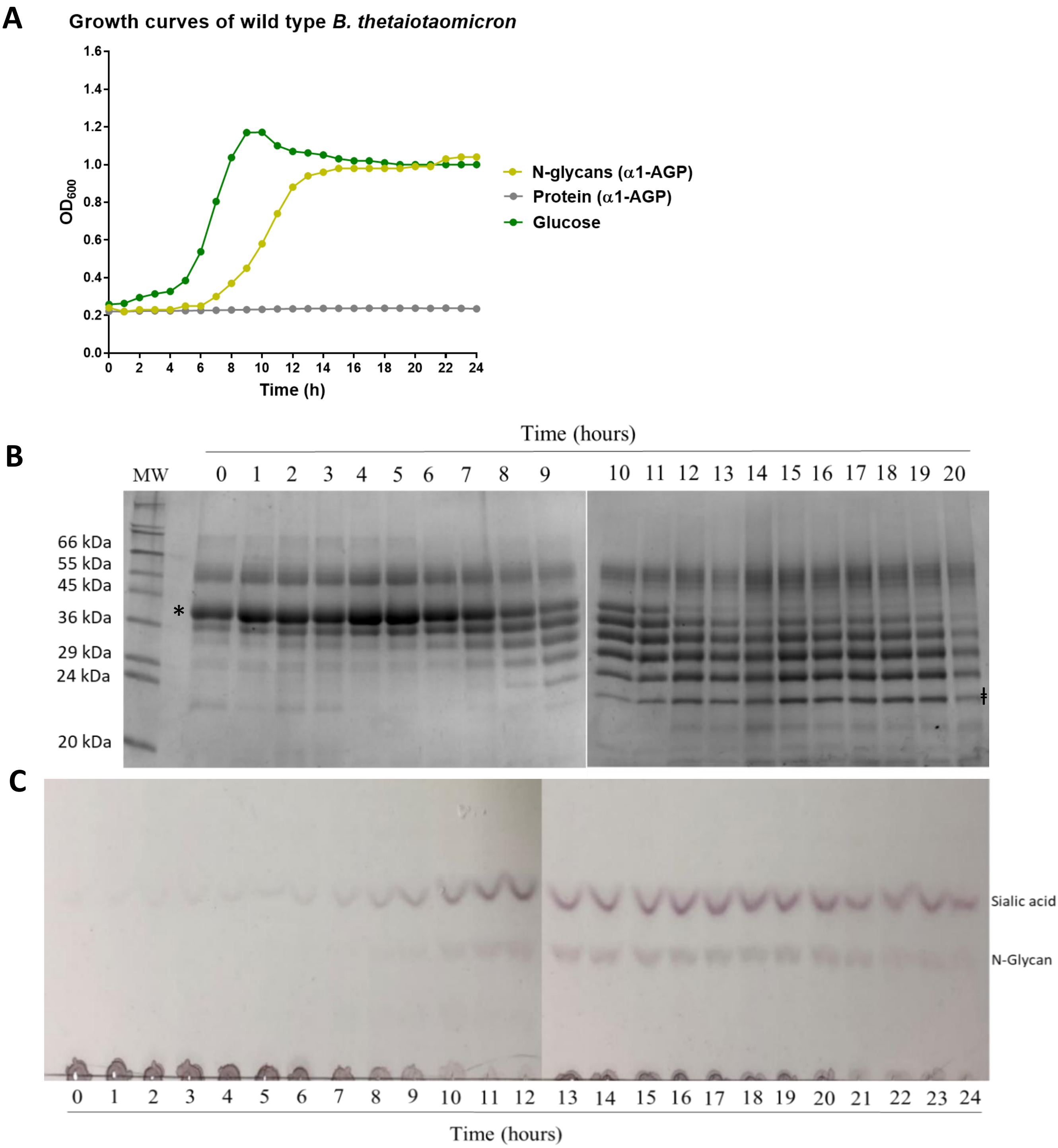


Figure 3.3: Investigating the N-Glycan degradation by the wild type *Bacteroides thetaiotaomicron*. A) Wild type *B. thetaiotaomicron* was cultured in minimal media+hematin containing 2% bovine α_1-Acid Glycoprotein, 2% bovine α_1-Acid Glycoprotein protein-only fraction or 1% Glucose as a control. Bacteria cultures were incubated for 24 hours, samples were collected and optical density (600 nm) was measured every hour. **B)** α_1-Acid Glycoprotein de-glycosylation was investigated using SDS-PAGE gel. Full-size bovine α_1-Acid Glycoprotein is 41-43kDa in size and is labelled with an asterisk (*). A band of ~23 kDa represents a fully de-glycosylated bovine α_1-Acid Glycoprotein (labelled ǂ). **C)** α_1-Acid Glycoprotein N-glycan degradation was visualised using a thin-layer chromatography (TLC). Gradual degradation of N-Glycans and sialic acid release was observed. 1mM of sialic acid (Neu5Ac) standard was used to identify the band. The band of a predicted de-sialylated N-glycan structure is labelled.

To investigate the α_1-AGP degradation by *B. thetaiotaomicron* further, the composition of the glycans in the spent media was analysed using High-Performance Liquid Chromatography (HPAEC-PAD) (see Section 2.3.3). The supernatants of the wild type *B. thetaiotaomicron* growth on MM-α_1-AGP (2% w/v) were collected at mid-exponential (after 8 h) and stationary (after 24 h) phases. The peak corresponding to the sialic acid (Neu5Ac) was identified by running a Neu5Ac standard alongside of these samples. Based on the appearance of various new peaks compared to the control MM+α_1AGP sample, the results suggest that *B. thetaiotaomicron* is utilising N-glycans found on bovine α_1-Acid Glycoprotein, leaving the sialic acid untouched (**Figure 3.4**). The precise N-glycan structures present in these samples could not be identified using this technique.

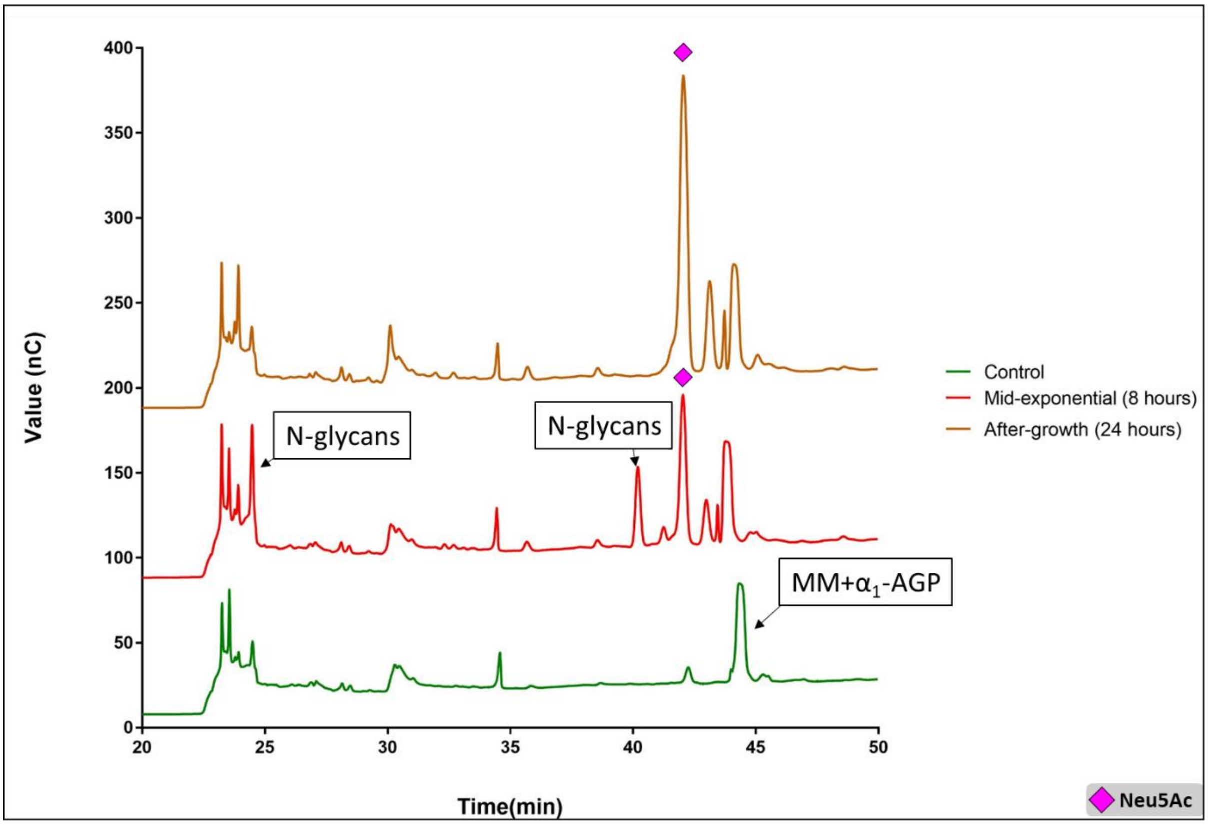


Figure 3.4: High-Performance Liquid Chromatography showing the deglycosylation of bovine α_1-Acid Glycoprotein by the wild type *Bacteroides thetaiotaomicron*. Control is a sample of MM+AAG collected before inoculation. Mid-exponential and after-growth samples show glycans present in the media during growth, products of surface degradation or secretion. Labelled peaks are suspected N-glycan structures. 1:10 dilutions of the samples were loaded. 1mM of sialic acid (Neu5Ac) standard was used.

3.3.3. RNA-seq data identifies genes upregulated by *B. thetaiotaomicron* in response to N-glycans

RNA seq was used to investigate genes upregulated during the growth of the wild type *B. thetaiotaomicron* on various N-glycoproteins as sole carbon sources compared to the glucose control. N-glycans present on the hen egg N-glycopeptide (Ludger) have been analysed and were found to contain only the complex sialylated bi-antennary N-glycan structures whereas bovine fetuin is decorated by a wide range of complex sialylated N-glycan structures (Aich *et al.,* 2013). N-glycan structures present on bovine α_1-Acid Glycoprotein were identified using mass spectrometry and contain mainly complex sialylated biantennary N-glycan structures. The graphical representation of these structures is displayed in **Figure 3.5**. Mass spectrometry data will be discussed in more detail in Section 3.1.3.6.

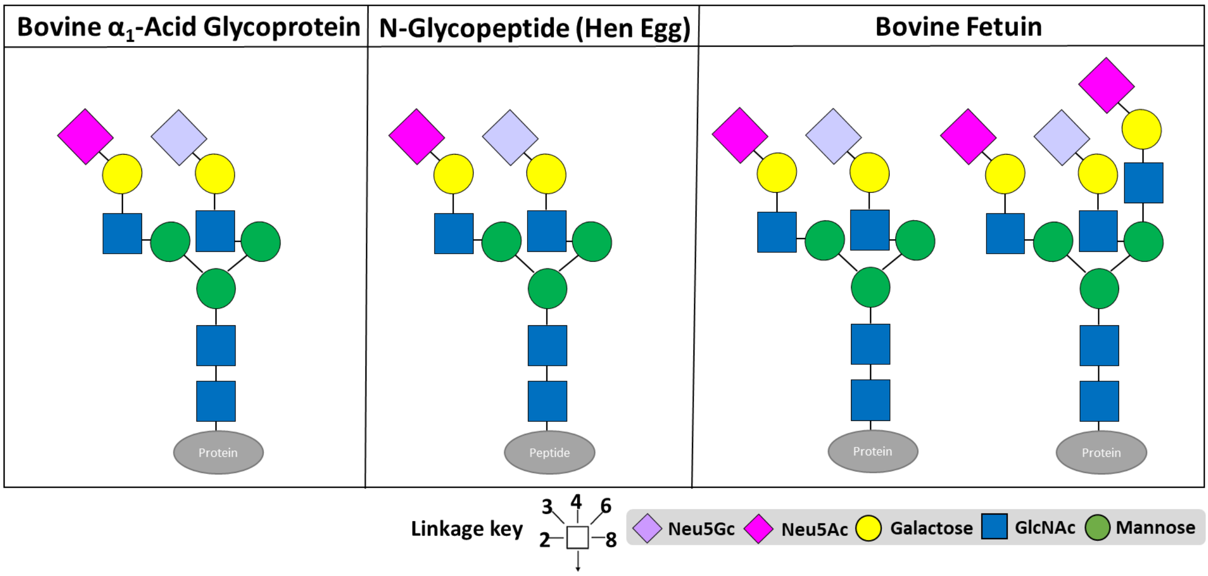


Figure 3.5: N-glycan structures found on glycoproteins used for RNAseq. Bovine α_1-Acid Glycoprotein and N-glycopeptide derived from hen egg contain mainly bi-antennary complex N-glycan structures whereas bovine fetuin has mixture of complex N-glycans. Examples of bi-antennary and tri-antennary sialylated complex N-glycan structures found on bovine fetuin are displayed. Figure was made using the information obtained from mass spectrometry analysis, Ludger Ltd. and Aich *et al*., 2013.

Wild type *B. thetaiotaomicron* was inoculated in minimal media+hematin containing either 1 % glucose, 2% of trypsin-digested bovine α_1-AGP, 2% of N-glycopeptide derived from the hen egg (Ludger) or 2% of N-glycans derived from bovine Fetuin prepared following the protocol described in Section 2.3.7. Cultures were grown to mid-exponential phase (~OD_{600nm} 0.6), the cells were harvested and RNA was extracted using QIAGEN RNeasy kit following a protocol described in Section 2.1.15. The RNA was sequenced by the Earlham Institute using the Illumina HiSeq4000 system. Unfortunately, when the results came back and were analysed, we found a significant *Serratia* species contamination present in glucose and N-glycans (Fetuin) samples. It was decided to repeat the RNA sequencing using the RNA derived from the wild type *B. thetaiotaomicron* grown on 2% bovine α_1-AGP and 1 % glucose as control. The extracted RNA was instead sent to Oxford Genomics Centre, who offered a better quote and completion time-scale. To be consistent, the RNA was also sequenced using the Illumina HiSeq4000 system.

The RNA sequencing data analysis was done with the help of John Casement (Bioinformatics Support Unit, NU). Briefly, reads were aligned against the *Bacteroides thetaiotaomicron* VPI-5482 genome using the bioinformatics tool Bowtie2. Genome annotation was obtained from the Ensembl database. Counts of reads aligning to the genomic features were obtained and differential expression analysis was done using R package DESeq2. The Benjamini-Hochberg p-values were obtained and adjusted using the same R package. The analysed RNA seq data is listed in Supplemental Table 3.1. The analysis of the results from both sequencing batches showed that *Serratia* spp. contamination present in the first batch did not show any significant influence over the gene upregulation, however for the purpose of clarity, contamination-free RNA sequencing data would be used for the publication.

RNA sequencing data analysis results were initially visualised using a Venn diagram. Venn diagram enabled identification of 66 common genes upregulated *in B. thetaiotaomicron* when it was grown on hen egg N-glycopeptide, bovine fetuin N-glycans and bovine α_1-AGP compared to the glucose control (**Figure 3.6**). Only the genes upregulated 2-fold or higher with the Benjamini-Hochberg adjusted p-values being less than 0.01 are classed as significant and thus, only these were included in the Venn diagram (Supplemental Table 3.1). Differences in upregulation patterns may be due to different N-glycan structures present on these N-glycoproteins.

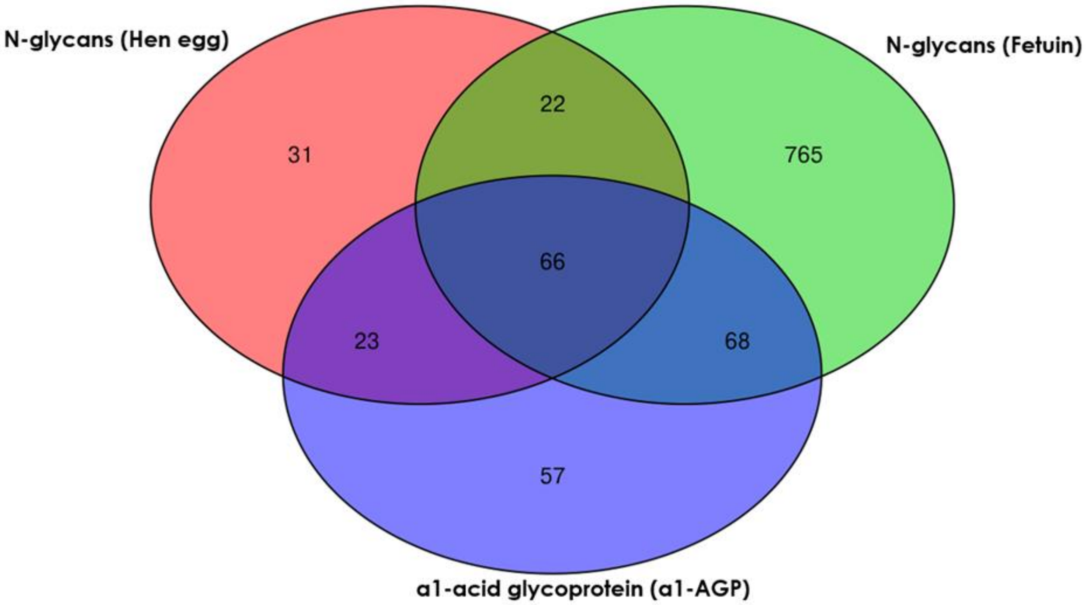
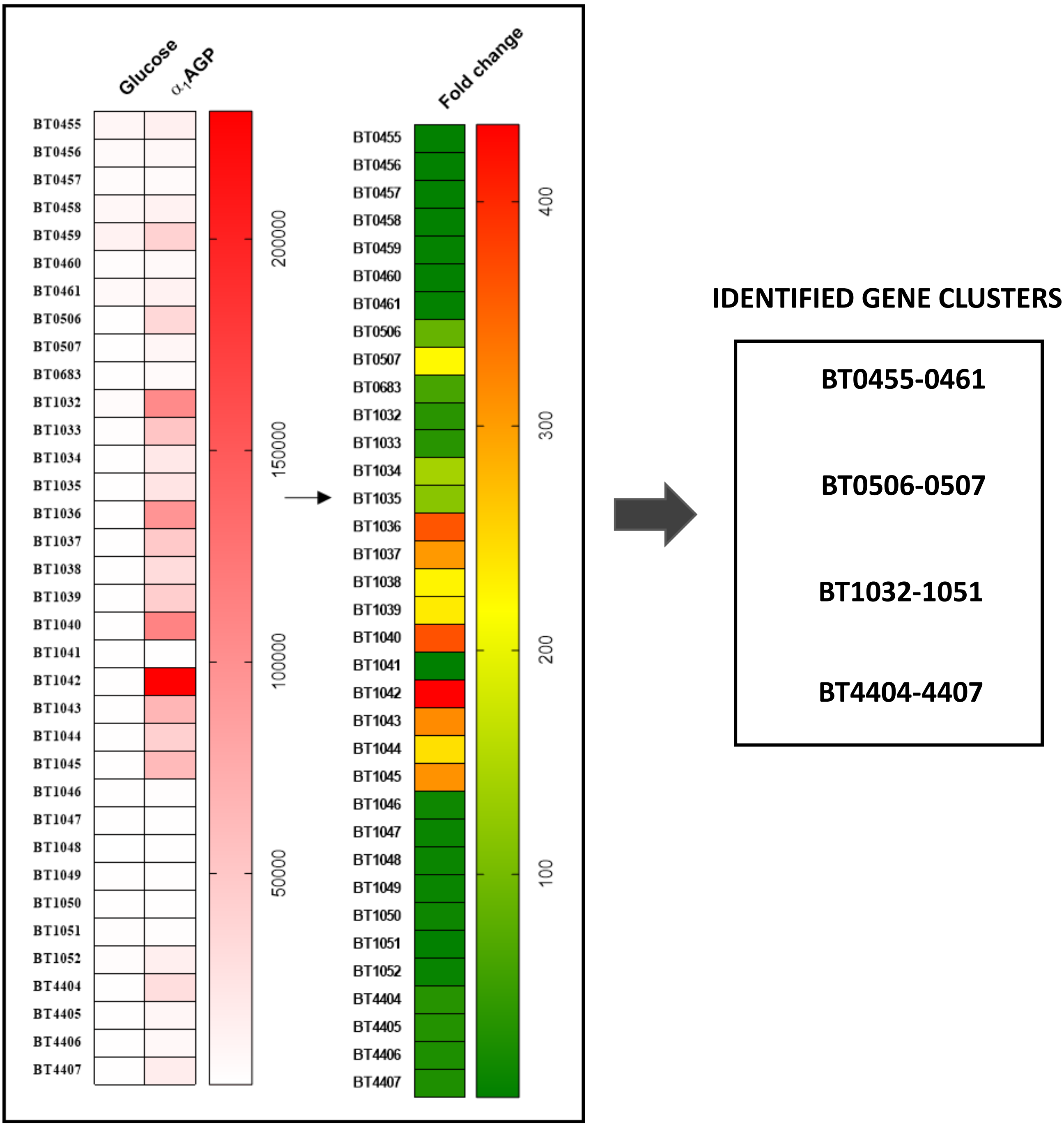


Figure 3.6: Venn diagram showing the upregulation of genes by *B. thetaiotaomicron* in response to N-glycoproteins. 66 common genes were found to be upregulated in all three conditions. 22 common genes were upregulated when *B. thetaiotaomicron* was grown on N-glycopeptide (hen egg) and N-glycans (fetuin). 23 common genes were upregulated when it was grown on N-glycans (hen egg) and α_1-Acid Glycoprotein. 68 common genes were upregulated when *B. thetaiotaomicron* was grown on α_1-Acid Glycoprotein and N-glycans (Fetuin). Venn diagram was done using a 'VennDiagram' R package and only includes genes upregulated 2-fold or higher, with p-values less than 0.01.

The genes upregulated when *B. thetaiotaomicron* was grown on α_1-Acid Glycoprotein were chosen for initial investigation because these samples contained no Serratia *spp.* contamination. The genes encoding CAZYmes and predicted glycan binding/import proteins were manually identified (**Supplemental Table 3.1**). The gene list was expanded to include genes upregulated less than 2-fold that were in the same gene clusters (predicted PULs) as the highly upregulated genes. The heatmap of the gene upregulation was made using the GraphPad Prism 7.0 software (**Figure 3.7.1**). The key gene clusters (predicted PULs) that are upregulated by N-glycans were identified. These include BT0455-0461, BT0506-0507, BT1032-1052 and BT4404-4407.



IDENTIFIED GENE CLUSTERS

BT0455-0461

BT0506-0507

BT1032-1051

BT4404-4407

Figure 3.7.1: RNA sequencing data identifying the apparatus *B. thetaiotaomicron* uses to utilise the N-glycans present on bovine α$_1$-Acid Glycoprotein. The white/red heatmap shows gene expression at the basal level whereas the green/red heatmap shows the fold change in gene upregulation compared to the glucose. The data identified key gene clusters that are upregulated by α$_1$-Acid Glycoprotein N-glycans: BT0455-0461, BT0506-0507, BT1032-1052 and BT04404-4407. Error bars and statistical information can be found in Supplemental table 3.1.

The genetic composition of BT0455-0461, BT0506-0507, BT1032-1052 and BT4407-4407 loci is displayed in **Figure 3.7.2.** BT0455-0461 locus encodes a GH33 sialidase BT0455, three GH20 β-hexosaminidases BT0456, BT0459, BT0460, putative esterase BT0457, GH2 mannosidase BT0458 and GH2 galactosidase BT0461. BT0506-0507 locus encodes predicted a GH20 β-hexosaminidase BT0506 and a regulator BT0507. BT4404-4407 locus encodes SusC/D/E and a GH18-family enzyme. BT1032-1052 locus encodes three predicted GH18-family enzymes, three SusC/D/E pairs, two surface glycan binding proteins (SGBPs), GH130 mannosyl phosphorylase, GH20 β-hexosaminidase BT1052 and a GH92 α-mannosidase BT1032 that was previously shown to be involved in high-mannose N-glycan degradation (Dr. Fiona Cuskin, unpublished data). The characterization of the enzymes encoded by BT0455-0461 and BT0506-0507 loci will be the main focus of this thesis. The BT1032-1051 and BT4404-4407 systems were explored by Dr. Lucy Crouch and will not be described in detail in this thesis.

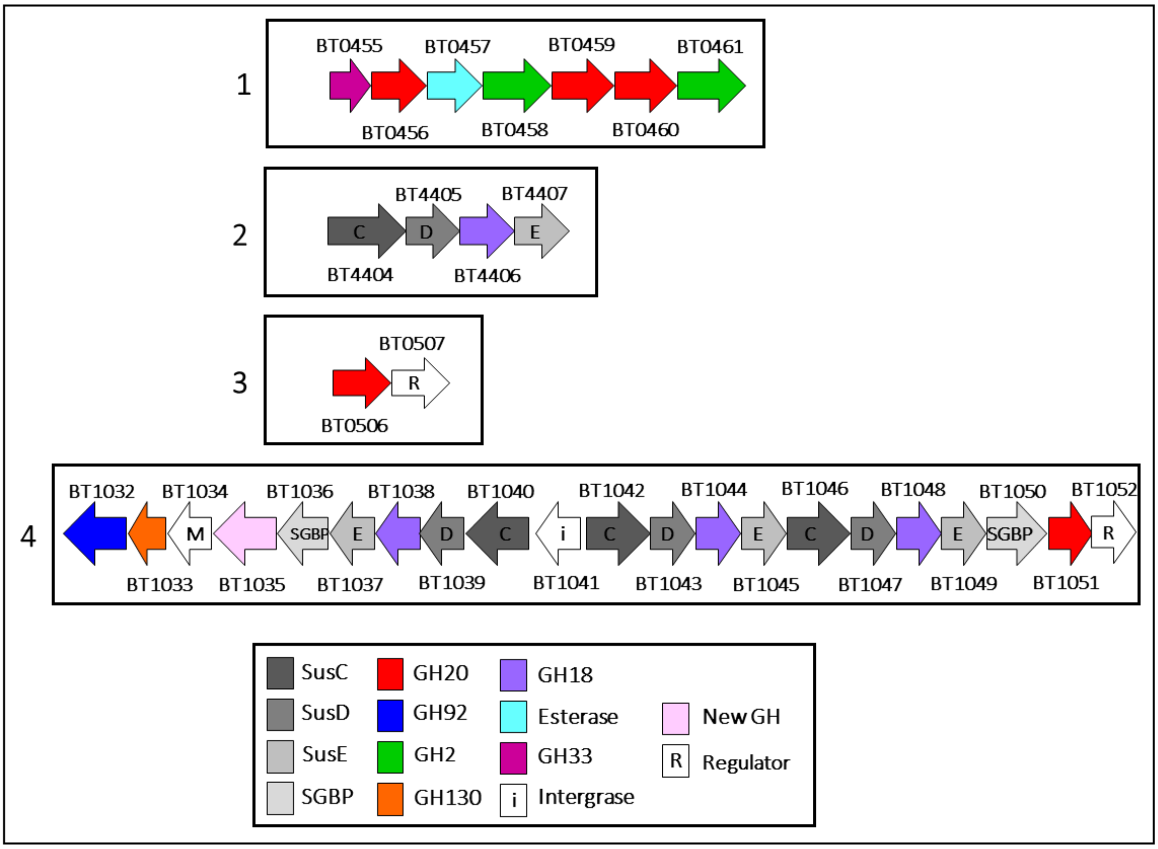


Figure 3.7.2: Genetic composition of the key PULs required for N-glycan utilisation by *B. thetaiotaomicron*. The composition BT0455-0461, BT0506-0507, BT1032-1052 loci is displayed. The information acquired from the Uniprot and Kyoto Encyclopedia of Genes and Genomes (Kegg) was used to make this figure.

To further investigate the upregulation of B.theta$^{0455-0461}$ cluster of genes, a heatmap of gene expression levels induced by bovine α_1-AGP, hen egg N-glycopeptide and bovine fetuin N-glycans was made (**Figure 3.8**).

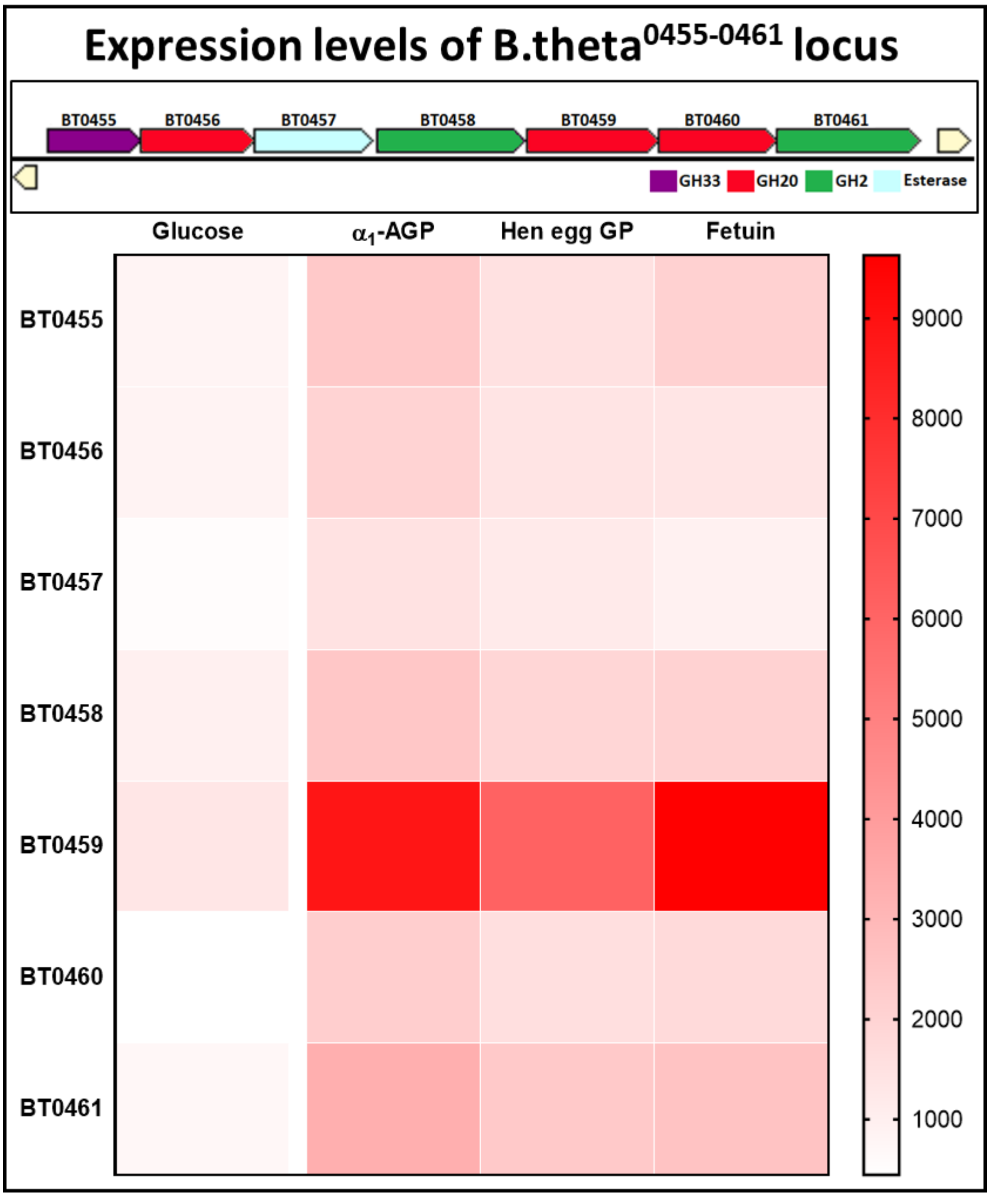


Figure 3.8: A heatmap showing basal gene expression levels of B.theta$^{0455-0461}$ locus. The heatmap was made plotting basal gene expression counts, averaged from triplicates. Error bars and further statistical information can be found in Supplemental table 3.1.

These data suggest that the genes from the B.theta$^{0455-0461}$ locus are highly expressed at the basal level. Very high basal expression and upregulation levels of BT0459^{GH20} were observed. Combined, these data indicate that this locus is important for *B. thetaiotaomicron* utilisation of N-glycans.

3.3.4. Analysis of gene expression levels by quantitative RT-PCR

To validate upregulation of the genes identified by the RNA seq data analysis when wild type *B. thetaiotaomicron* was grown on various N-glycoproteins, quantitative RT-PCR (qPCR) analysis was carried out. Messenger RNA derived from the wild type *B. thetaiotaomicron* grown on various N-glycoproteins and glucose as sole carbon sources was converted to cDNA following a protocol described in Section 2.1.16. The qPCR primers, based on the genes of interest, were designed using a GenScript tool (**Supplemental table 3.2**).

q-PCR was set up following a protocol described in Section 2.1.17 using the primers for B.theta$^{0455\text{-}0461}$ genes and B.theta$^{1032\text{-}1051}$ genes, PUL that contains an endo-β-N-acetylglucosaminidase BT1044^{GH18} required to cleave the N-glycan structures off the protein backbone (Dr. Lucy Crouch, unpublished data). The results suggest that *B. thetaiotaomicron* begins N-glycan degradation by upregulating B.theta$^{1032\text{-}1051}$ PUL, specifically BT1044^{GH18} encoding the enzyme that cleaves the core N-glycan chitobiose (GlcNAc-β1,4-GlcNAc-Asn) bond, releasing the N-glycan structures from the protein backbone (**Figure 3.9**). This gene is very highly upregulated in mid-exponential phase. Because the fetuin N-glycan structures were already cleaved off the protein backbone by the PNGaseF, neither BT1044^{GH18} not any other genes from this locus were upregulated. On the other hand, *B.theta*$^{0455\text{-}0461}$ genes are upregulated at much lower rate, however considering the RNA seq data analysis results that show high basal expression levels of these genes, the data seems to be consistent. BT0459^{GH20} is the highest upregulated gene of this locus, suggesting its importance for N-glycan degradation. It is important to note that the upregulation differences observed in BT0455, BT459 and BT0461 compared to the rest of the screened genes could be due to the presence of gene-specific regulatory factors or due to the mRNA stability/mRNA decay.

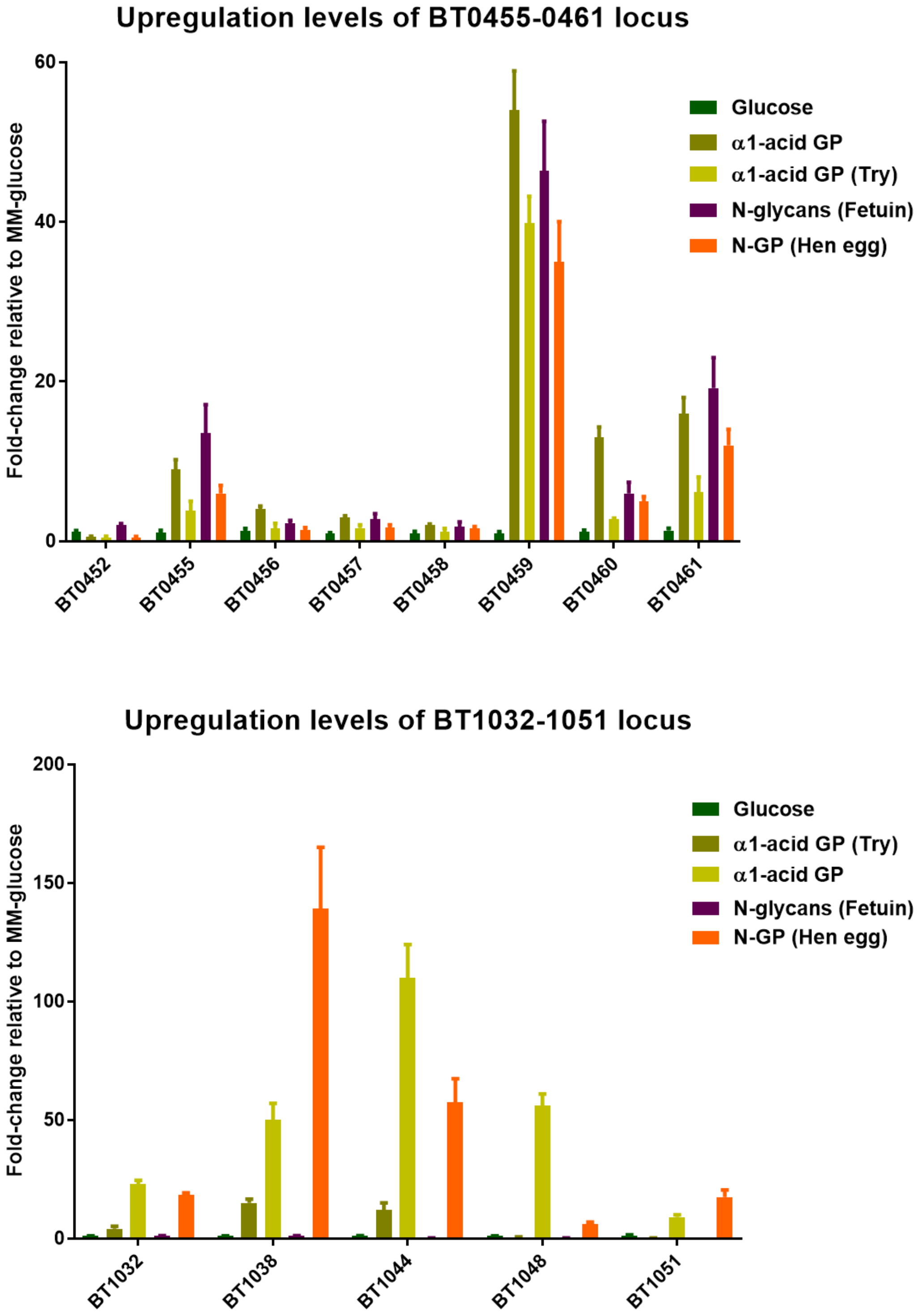


Figure 3.9: qPCR analysis of genes encoding CAZYmes upregulated during growth of *B. thetaiotaomicron* on N-glycans. Wild type *Bt* cells were grown on each of the substrates to the mid-exponential phase (OD 0.4-0.6) and harvested. RNA was extracted and quality-checked before converting it into the cDNA. N-glycans used in this experiment include native α_1-AGP, trypsinised α_1-AGP, PNGase-F-cleaved bovine fetuin N-glycans and hen egg N-glycopeptide. Upregulation levels are shown as fold-change relative to MM-Glucose. Data for each gene and growth condition was obtained in triplicates, twice. The error bars indicate the standard deviation.

3.4. Discussion

The ability to metabolise a diverse range of extremely complex glycan structures present in the human gut by numerous members of Bacteroidetes, including *B. thetaiotaomicron,* has been linked to the huge repertoire of enzymes, each specialised in degrading specific glycan structures (Ndeh *et al.,* 2018). This ability to metabolise highly complex glycan structures was observed in this chapter when several prominent members of the human gut microbiota were screened for their ability to utilise complex N-glycans as a sole carbon source. The capacity to efficiently metabolise the complex biantennary N-glycan structures present on the bovine α_1-Acid Glycoprotein was observed in wild type *B. thetaiotaomicron*, *B. fragilis*, *B. massiliensis* and *B. uniformis* (**Figure 3.1**). This is surprising considering the availability of N-glycan sources in the human gut in a form of dietary and host-derived N-glycoconjugates is relatively high (Tailford *et al.,* 2015). The ability to efficiently utilize host-derived N-glycans could be a metabolically taxing but a favourable trait in the gut that gives these microbes a metabolic advantage in times of starvation.

A more thorough investigation has revealed that a prominent member of a human gut microbiota, *B. thetaiotaomicron,* is capable of utilizing a diverse range of N-glycan structures decorating various N-glycoproteins, such as transferrin, fetuin and α_1-AGP (**Figure 3.2**). Although N-glycan degradation appears to be metabolically taxing, it was observed that this microbe only utilizes the N-glycan fraction of the bovine α_1-AGP and does not require the protein fraction for the growth (**Figure 3.3**). Considering how complex N-glycan structures can be (e.g. biantennary, triantennary, tetraantennary and bisecting), these results indicated that *Bt* must possess an extensive repertoire of specialised enzymes capable of accommodating and degrading such highly complex glycan structures. Transcriptomic data analysis helped to identify this complex N-glycan utilization apparatus of *Bt* (**Figure 3.7.1**). Based on the high gene upregulation data, the key enzymes were identified to belong to the BT0455-0461, BT0506-0507, BT1032-1051 and BT4404-4407 loci (**Figure 3.7.2**). A further analysis of the upregulation data identified seven enzymes encoded by the B.theta$^{0455\text{-}0461}$ and

BT0506-BT0507 loci that were upregulated when the wild type *Bt* was grown on three different sources of N-glycans, suggesting their importance for the complex N-glycan degradation by this bacterium (**Figure 3.8**). Based on the information obtained from CAZy database, the upregulated genes encode the sialidase BT0455^{GH33}, four β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, β-mannosidase BT0458^{GH2} and β-galactosidase BT0461^{GH2}. The observed very high basal expression levels of these genes indicates their importance to the N-glycan metabolism and proliferation of *Bt* in the mucus layer of the gut. Interestingly, the systemic analysis of the gut metatranscriptome and gut metagenome studies identified a significant number of microbial transcripts (41%) was not differentially upregulated in comparison to their genomic abundance. A huge number of gene families were found to be consistently upregulated at a low fold at the transcriptional level but highly abundant metagenomically (Franzosa et al., 2014). These findings could explain such a low upregulation fold displayed by some of the key genes in the RNA seq data.

qPCR analysis validated the RNA-seq data analysis results. The high-upregulation levels of genes from BT0455-0461 locus and key genes of BT1032-1051 were consistent with the RNA-seq data and indicated their importance for the complex N-glycan metabolism by *Bt*. The significant change observed in the upregulation data of BT0455 could be explained by the abundance and importance of the sialic acid utilization by the human gut microbiota *in vivo*. The data also suggested that the N-glycan degradation by *Bt* may be initiated by a predicted endo-glycosidase BT1044^{GH18} enzyme that could release the N-glycan structures off the protein backbone. However, full characterisation is required to confirm this hypothesis.

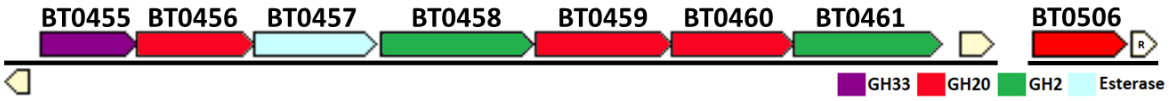
Combined, these data suggested that the complex N-glycan utilisation is a metabolically taxing and a niche trait present only in dominant members of Bacteroidetes. In *B.t*, complex N-glycan degradation appears to depend on a cooperative activity of numerous enzymes encoded by four discrete loci. However, in order to fully understand this pathway, the full biochemical characterisation of the key enzymes is required.

Chapter 4: Characterization of the key members of N-glycan utilization apparatus in *B. thetaiotaomicron*

4.1. Introduction and motivation for the research

Bacteroides thetaiotaomicron, a Gram-negative symbiotic commensal gut microbe, is capable of utilizing a variety of dietary and host-derived glycans as sole carbon and energy sources. In order to degrade complex glycan structures, *B. thetaiotaomicron* encodes a range of enzymes that are known as carbohydrate-active enzymes (CAZymes) (www.cazy.org). These include glycoside hydrolases (GHs), polysaccharide lyases (PLs), glycosyl transferases (GLs) and carbohydrate esterases (CEs).

As discussed in Chapter 3, BT0455-0461 and BT0506-0507 loci are upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources, compared to the glucose control. Interestingly, BT0455-0461 does not fit classical PUL model due to the lack of SusC/D pair, SGBP and a regulator. Numerous CAZymes are encoded by these loci, including a sialidase BT0455^{GH33}, galactosidase BT0461^{GH2}, four hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, esterase BT0457^{CE} and a mannosidase BT0458^{GH2} (**Figure 4.1**).



Protein name	Predicted size (kDa)	CAZy family	Predicted activity
BT0455	61	GH33	α-sialidase
BT0456	78	GH20	β-hexosaminidase
BT0457	79	CE	Sialic acid esterase
BT0458	99	GH2	β-mannosidase
BT0459	87	GH20	β-hexosaminidase
BT0460	78	GH20	β-hexosaminidase
BT0461	95	GH2	β-galactosidase
BT0506	86	GH20	β-hexosaminidase

Figure 4.1: Overview of enzymes encoded by the BT0455-0461 and BT0506-0507 loci. The collective information obtained from KEGG, IMG and Uniprot databases was used to make this figure (see Section 2.2).

In theory, combined cooperative activity of these enzymes would enable degradation of a typical N-glycan structure. In nature, complex N-glycans structures are commonly heavily sialylated and would require an activity of a sialidase to degrade. Sialidases are a huge group of specialised enzymes produced by both prokaryotes and eukaryotes that catalyse the cleavage of glycosidic linkages of terminal sialic acid caps of complex carbohydrates found on glycoconjugates (Park *et al.,* 2013). The majority of bacterial sialidases are grouped together based on the similarities between their amino acid sequences and are classified into the glycoside hydrolase (GH) family 33 (GH33) of the CAZy classification. The only sialidase produced by *B. thetaiotaomicron*, BT0455^{GH33}, was upregulated when it was grown on N-glycans indicating it is necessary for the degradation of these structures.

Acetylation of sialic acids, such as Neu9Ac, can hinder the activity of the sialidases. Thus, the de-acetylation of O- and N-linked acetylated glycan structures by carbohydrate esterases is a common occurrence in nature. Currently characterized carbohydrate esterases are grouped into 16 CAZy families. These include sialate-O-acetylesterases that are a broad-acting group of esterases that are classified into the SGNH superfamily of hydrolases, a distinct class of α/β hydrolases (Rangarajan et al, 2011). BT0457^{CE} is a putative sialic acid esterase that is potentially required to boost the catalytic efficiency of the sialidase BT0455^{GH33}.

Desialylation of a typical complex N-glycan structure would expose galactose linkages that would require an activity of a galactosidase, such as a predicted β-galactosidase BT0461^{GH2}, to cleave. Enzymes with β-galactosidase activity are grouped into the glycoside hydrolase A (GH-A) superfamily of glycoside hydrolases that is further subdivided into the GH1, GH2, GH35 and GH42 subfamilies (Talens-Perales *et al.,* 2016). β-galactosidases were shown to catalyse the hydrolysis of terminal, non-reducing D-galactosyl bonds found on various glycan structures, through disaccharides to oligosaccharides, including *O*- and *N*-glycans (Panesar *et al.,* 2010).

Removal of caps from a typical complex N-glycan structure would reveal numerous GlcNAc linkages that would require several specialised β-hexosaminidases to remove. Four predicted β-hexosaminidases, BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, were upregulated when *B. thetaiotaomicron* was grown on N-glycoproteins. The enzymes classified into the glycoside hydrolase (GH) family number 20 (GH20) of CAZy classification are a large group of specialised enzymes that are found in both eukaryotes and prokaryotes. The group is mainly comprised of exo-acting β-hexosaminidases such as β-N-acetylglucosaminidases and β-N-acetylgalactosamindases that cleave D-GlcNAc and D-GalNAc monosaccharides from the terminal ends of carbohydrates, respectively. Members of the GH20 family of enzymes are involved in various important biological processes, catalysing the hydrolysis of the D-GlcNAc/D-GalNAc linkages in glycoconjugates and glycosaminoglycans (Val-Cid *et al.,* 2015).

Once various GlcNAc linkages are removed from the typical complex N-glycan structure, an N-glycan core tetrassacharide (Man-α1,6(Man-α1,3)Man-β1,4-GlcNAc-) is exposed. Degradation of this structure would require a cooperative activity of several specialised mannosidases, such as a β1,4-mannosidase BT0458^{GH2} that was upregulated when *Bt* is grown on complex N-glycans. Mannosidases are a large group of enzymes that are specialised in catalysing the hydrolysis of the glycosidic mannose linkages found in complex carbohydrates. They are grouped into α-mannosidases and β-mannosidases. Majority of β-mannosidases are classified into the glycoside hydrolase (GH) family number 2 (GH2) of the CAZy classification, along with β-galactosidases and β-glucuronidases. To date, it is known that *B. thetaiotaomicron* produces 33 GH2-family enzymes, 5 of which are annotated as β-mannosidases. Like the rest of GH2 family of enzymes, β-mannosidases are retaining enzymes that follow a double-displacement Koshland mechanism (CAZypedia.org, GH2 family).

However, it is important to characterise these enzymes to determine their substrate specificities that would allow us to fully understand how these enzymes are employed by *B. thetaiotaomicron* to degrade complex N-glycan.

4.2. Research objectives

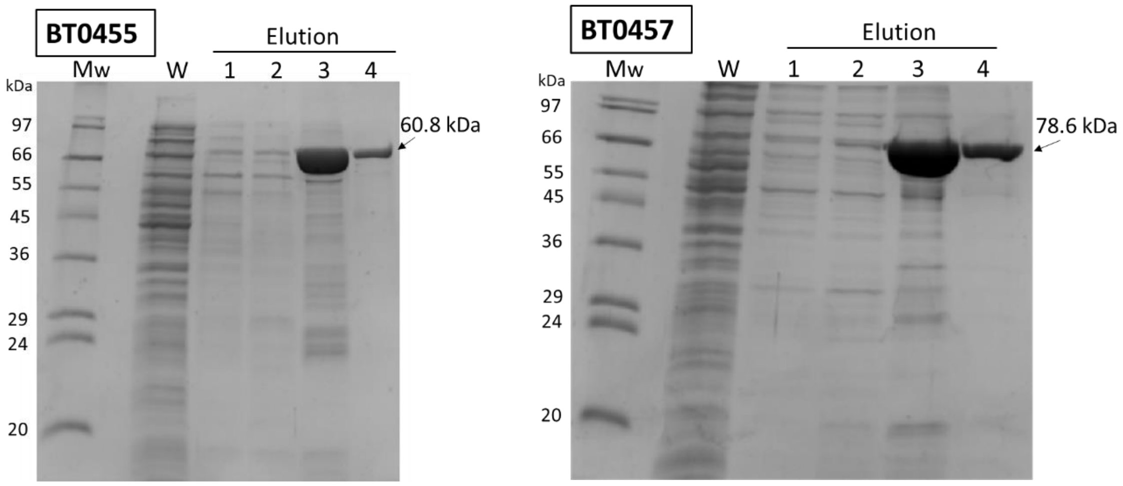
i. To functionally characterize enzymes involved in N-glycan degradation

ii. To investigate potential cooperation of multiple polysaccharide utilization loci (PULs) required for N-glycan degradation

4.3. Results

4.3.1. Protein cloning, expression and purification

Genes encoding the BT0455^{GH33}, BT0456^{GH20}, BT0457^{CE}, BT0458^{GH2}, BT0459^{GH20}, BT0460^{GH20} and BT0461^{GH2} have been amplified and cloned into pET28a derivative *E.coli* prokaryotic expression vector by the Nzytech. BT0506^{GH20} was cloned into pET28a plasmid following a protocol described in Sections 2.1.11-13. In both cases, the encoded soluble recombinant proteins contained an N-terminal His-tag sequence for easy purification (MGSSHHHHHHSSGPQQGLR). Chemically competent *E.coli* Tuner cells were routinely used for the recombinant protein expression following a protocol described in Section 2.1.20. Protein overexpression was induced using 1 mM of IPTG and overnight incubation at 16 °C. Cells were harvested by centrifugation and supernatant containing the soluble recombinant protein was collected. Purification of N-terminal His-tagged recombinant proteins was done using an immobilised metal affinity chromatography (IMAC) following a protocol described in Section 2.1.21. The purified protein concentration was determined by measuring the absorbance at 280 nm (A280nm) and calculating it using the molecular weight and extinction coefficient (see Section 2.1.26). Collected purified protein fractions were analysed by SDS-PAGE (**Figure 4.2**). Elution fractions containing the protein were combined and dialysed against the buffer using 12kDa cut-off membrane tube (see Section 2.1.23). Dialysed proteins were concentrated and used for enzyme assays. The theoretical molecular weights of the recombinant proteins are displayed in **Figure 4.2**.



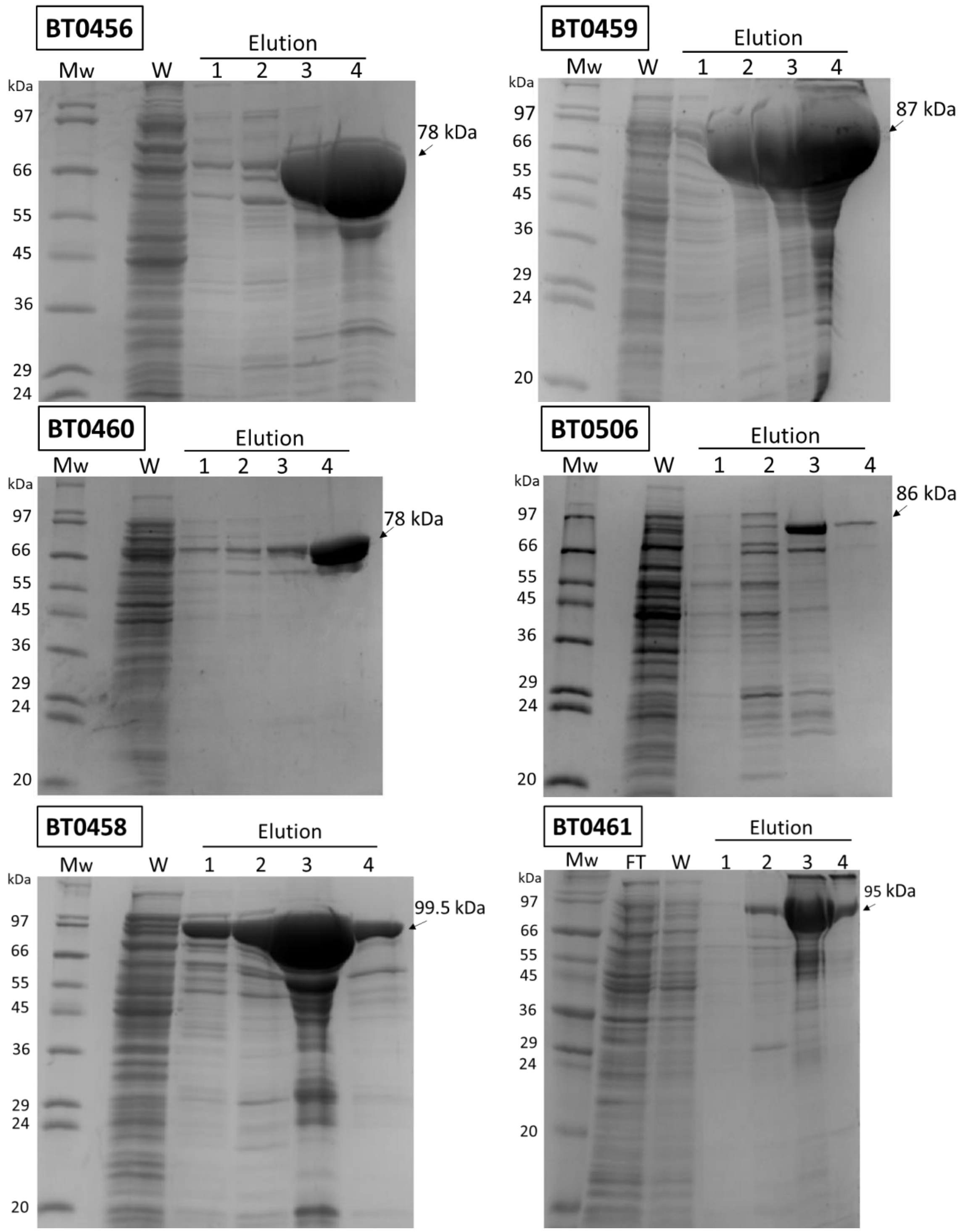


Figure 4.2: SDS-PAGE gel showing the fractions of purified BT0455-0461 and BT0506 proteins. 12.5% polyacrylamide gel was used for the SDS-PAGE analysis. Mw is a wide-range molecular marker (Sigma Aldrich). FT is a supernatant flow-through. W is fraction collected following a wash with Talon buffer. Lane 1 and 2 show protein fractions eluted using 10 mM imidazole whereas lane 3 and 4 show protein fractions eluted using 100 mM imidazole. The theoretical molecular weight of the proteins is displayed in kDa.

4.3.2. BT0455: Characterisation of the sialidase

4.3.2.1. Introduction

Sialic acid is usually found on the terminating branches of complex carbohydrates decorating mucins, O-glycoproteins and N-glycoproteins (Juge *et al.,* 2016). On N-glycoproteins, terminal sialic acid residues usually form α2-3/6 glycosidic linkages. *B. thetaiotaomicron,* a dominant member of human GI tract, possesses a single sialidase - BT0455. Because BT0455^{GH33}, also referred to as BTSA, is the only sialidase found in *B. thetaiotaomicron* to-date, the basic substrate specificity and structure of it has already been determined (Park *et al.,* 2013). It is a hydrolytic sialidase that has a wide substrate specificity and can cleave α2-3, α2-6 and α2-8-linked terminal sialic acid linkages of tested substrates such as pNP-Neu5Ac, 3'-sialyllactose, 6'-sialyllactose and colominic acid. As mentioned previously, *B. thetaiotaomicron* can release free sialic acid but it lacks the operon to utilize it. Perhaps it removes the sialic acid caps in order to access the underlying carbohydrate structures. Interestingly, BT0455^{GH33} was shown to be upregulated when *B. thetaiotaomicron* was grown on porcine gastric mucin (PGMIII) (Martens *et al.,* 2008). BT0455^{GH33} was also shown to be a key enzyme required for sialic acid release *in vivo* in a mouse gut model (Bjursell *et al.*, 2006; Ng *et al.,* 2013)

As was previously shown in Chapter 3 (**Figure 3.7.1**), BT0455^{GH33} is upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources, compared to the glucose control. Although it is the only sialidase encoded by the *B. thetaiotaomicron*, the knowledge of its activity against N-glycans is lacking. Thus, the aim of this chapter was to investigate the ability of BT0455^{GH33} to target and catalyse the cleavage of terminal sialic acids capping N-glycoproteins.

4.3.2.2. Sequence comparison of BT0455^{GH33} to other characterised sialidases

In order to investigate the sequence similarities and differences between BT0455^{GH33} and other structurally and biochemically characterized GH33 sialidases listed on CAZy database, a phylogenetic tree was constructed using the amino acid sequences obtained from Uniprot database and phylogeny.fr tool (**Figure 4.3**).

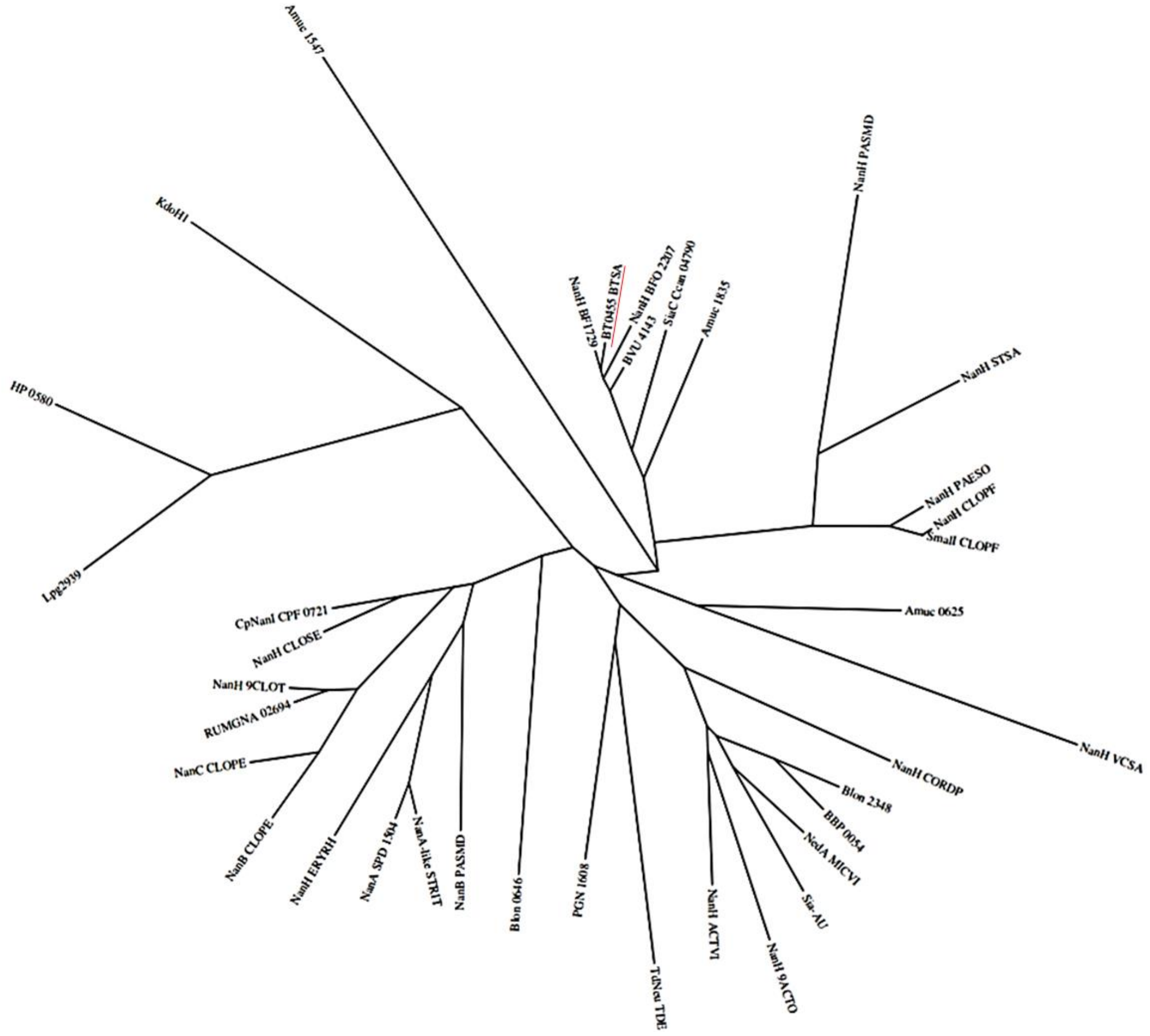


Figure 4.3: Phylogenetic tree of characterized CAZy GH33-family sialidases. Sialidases from these species are included in the tree: *B. thetaiotaomicron, B. fragilis, B. longum, C. canimorsus, C. perfringens, A. viscosus, C. septicum, C. diphteriae, F. tularensis, H. pylori, L. pneumophilia, M. viridifaciens, P. sordellii, P. gingivalis, R. gnavus, S. typhimurium, S. intermedius, S. pneumoniae, T. forsythia, T. denticola, T. pyogenes and V. cholerae*. Sialidases were selected using the CAZy list of characterized GH33-family enzymes. The phylogenetic tree was constructed using the DynTree layout.

The phylogenetic tree shows that BT0455^{GH33} is clustered on the same branch as other *Bacteroides spp*. sialidases, such as NanH from *B. fragilis* (BF_1729) and *B. vulgatus* (BVU_4143). Interestingly, it also appears to be very similar to the NanH from *Tannerella forsythia* (BFO_2207), a sialidase from *A. muciniphila* (Amuc_1835) and a sialidase from *C. canimorsus* (Ccan_045790). These similarities suggest that these sialidases that could be structurally and functionally related to BT0455^{GH33}. Interestingly, NanH from *B. fragilis* and NanH from *T. forsythia* have been previously shown to be involved in sialylated N-glycan degradation (Cao *et al.,* 2014; Roy *et al*., 2011). Combined, the phylogenetic analysis results suggest that BT0455^{GH33} along with the other homologous sialidases could share the same substrate specificity as NanH from *B. fragilis* and NanH from *T. forsythia*.

To further investigate the similarities of BT0455^{GH33} to the sialidases produced by other *Bacteroides spp*., the sequences of BT0455^{GH33}, BVU_4143^{GH33} and BF_1729^{GH33} were aligned using the Clustal Omega Multiple Sequence Alignment tool and visualised using the ESPript 3.0 tool. The results are displayed in **Figure 4.4**. Based on conserved residues highlighted in red, it shows that these sialidases are approximately 70% homologous to one another. The sequence alignment also shows that BT0455^{GH33}, BVU4143^{GH33} and BF1729^{GH33} contain the conserved features of a typical GH33-family sialidase, such as RIP/RLP motif (Arg-Ile/Leu-Phe), repetitive Asp-boxes (Ser/Thr-X-Asp-X-Gly-X-Thr-Trp/Phe) and catalytic residues, tyrosine (Y) and glutamic acid (E).

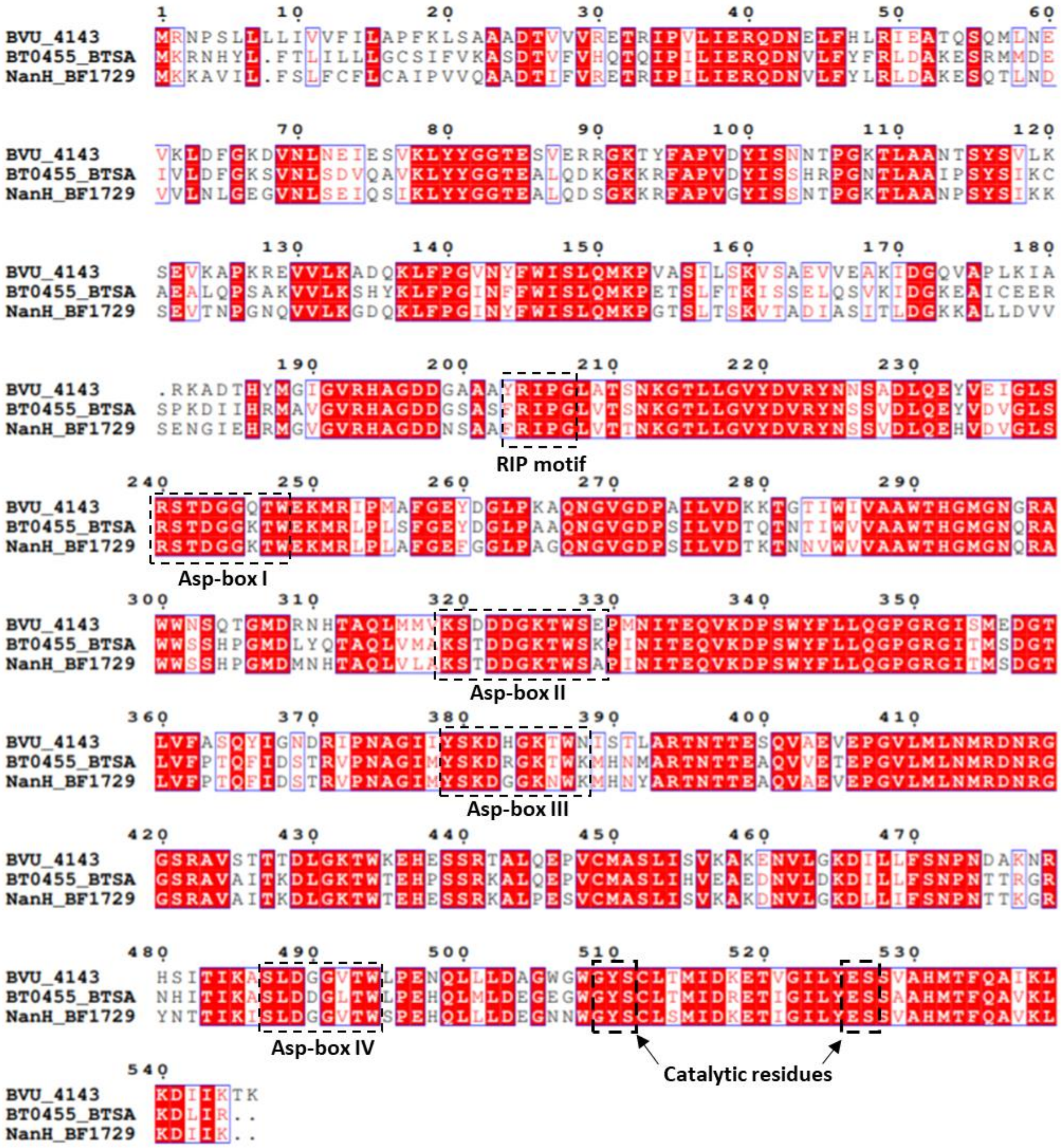


Figure 4.4: Alignment of the sialidases from *Bacteroides* species. RIP motif (Arg-Ile-Pro) contains arginine required for the catalytic activity of the GH33-family sialidase. Asp-box I, II, III and IV motifs maintain structural integrity and characteristics of a typical sialidase by folding into well-defined β-hairpins. Sequences were aligned using Clustal Omega tool and visualised using ESPript 3.0 tool. Fully (100%) conserved amino acids are displayed in red.

4.3.2.3. Defining N-glycan specificity of BT0455^{GH33}

As was previously shown in the RNA seq and qPCR data analysis, the sialidase BT0455^{GH33} is upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources, compared to the glucose control (**Figure 3.8 And Figure 3.9, respectively**). To investigate the enzymatic BT0455^{GH33} activity against sialylated N-glycan substrates, enzyme assays were set up following a protocol described in Section 2.3.5. BT0455^{GH33} was mixed with various commercially available N-glycoproteins. Because deglycosylation of an N-glycoprotein with a single enzyme may not be sufficient enough to see a noticeable change in a protein band size, BT0461^{GH2}, a GH2-family β-galactosidase that is present in the same B.theta$^{0455-0461}$ locus, was added into the mixture. The activity of BT0461^{GH2} will be discussed later in this thesis. The enzyme assays were incubated in sodium phosphate 20mM (pH 7) buffer for 1 hour at 37 °C and products visualised using SDS-PAGE. Deglycosylation of the glycoprotein can be observed by a decrease in glycoprotein band size. The results show that these enzymes are active versus bovine fetuin, but show no activity against bovine lactoferrin, bovine immunoglobulin G (IgG) and human transferrin (**Figure 4.5**). These results could be explained by the presence of differentially siallylated N-glycan structures on these glycoproteins. Representative glycan structures present on bovine fetuin are Neu5Ac-**α2,3**-Gal-β1,4-GlcNAc, Neu5Ac-**α2,3**-Gal-β1,3-(Neu5Ac-α2,6)-GalNAc and Neu5Ac-**α2,3-**Gal-β1,3-GalNAc, with N-glycans contributing to a 79% of a total glycan mass (Wu *et al.,* 2016). Whereas lactoferrin, transferrin and IgG predominantly contains Neu5Ac-**α2,6**-Gal-β1,4-GlcNAc structures (Park et al., 2013; Barboza *et al.,* 2012). However, it is important to point out that these results may also be due to the inaccessibility of sialic acid to BT0455^{GH33} (e.g. due to high levels of sialic acid acetylation) or due to insufficient amount of glycosylated structures present on these commercially-acquired N-glycoproteins.

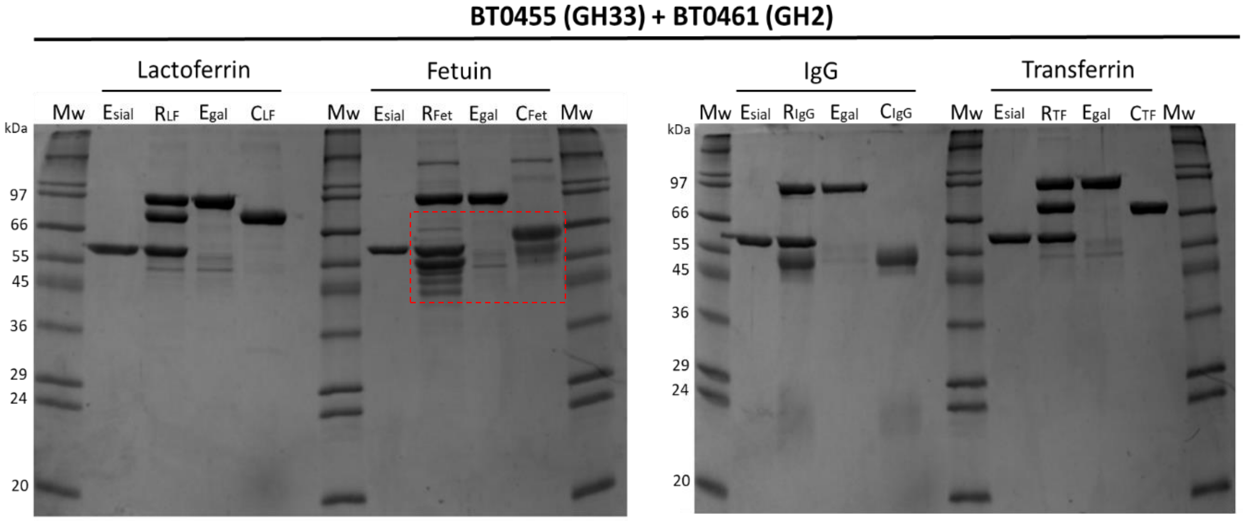

BT0455 (GH33) + BT0461 (GH2)
Lactoferrin
Fetuin
IgG
Transferrin
Mw Esial RLF Egal CLF Mw Esial RFet Egal CFet Mw
Mw Esial RIgG Egal CIgG Mw Esial RTF Egal CTF Mw
kDa
97
66
55
45
36
29
24
20
kDa
97
66
55
45
36
29
24
20

Figure 4.5: SDS-PAGE gel showing the degradation of N-glycoproteins by BT0455^{GH33} and BT0461^{GH2}. The enzyme assays were set up by mixing 1µM of required enzymes with 1mg/ml of each of the glycoproteins in sodium phosphate 20mM (pH 7) buffer and incubating for 1 hour at 37 °C. 1:20 dilutions of the enzyme assay products were loaded onto the SDS-PAGE gels. Degradation is observed by the decrease in glycoprotein band size in reaction lane compared to the control (highlighted in a red box). Mw is a wide-range molecular marker (Sigma Aldrich). E_{sial} is BT0455^{GH33}; R is the reaction mixture; E_{gal} is BT0461^{GH2} and C is N-glycoprotein control.

To further investigate the BT0455^{GH33} enzymatic activity against various sialylated substrates and to observe the sialic acid release, enzyme assays were set up and products were visualised using the thin-layer chromatography (TLC) following a protocol described in Section 2.3.2. Speculating that the release of the N-glycan structures from the protein backbone would affect the BT0455^{GH33} activity, some of the N-glycoproteins were de-glycosylated using the Peptide-N-glycosidase F (PNGaseF) treatment that hydrolyses the bond between the reducing-end GlcNAc residue of an N-glycan structure and the asparagine residue of the protein backbone. Commercially-available PNGaseF from *E. meningtoseptica* (Sigma Aldrich) that can remove oligomannose, hybrid and complex N-glycans structures was used to cleave the N-glycans off the protein backbone of bovine α_1-AGP and human transferrin following a protocol described in Section 2.3.7. BT0455^{GH33} enzyme assays were set up against various substrates following a protocol described in Section 2.3.5 and incubated for 1 hour at 37 °C. The reaction products were visualised using the TLC (**Figure 4.6**). Based on the sialic acid release, the results show that BT0455^{GH33} is active against all substrates tested: bovine submaxillary mucin (BSM), bovine α_1-AGP, PNGaseF-digested bovine α_1-AGP, PNGaseF-digested human transferrin and bovine fetuin. BSM is a heavily sialylated mucin (up to 17%) that contains different forms of sialic acid –neuraminic acid, KDO and KDN (Bhavanandan *et al.,* 1998). As evident by the multiple sialic acid bands visible on the TLC, BT0455^{GH33} is a broad-acting enzyme that can tackle different forms of sialic acid (**Figure 4.6**). No difference in BT0455^{GH33} activity against PNGaseF-digested and native bovine α$_1$-AGP was observed. Interestingly, BT0455^{GH33} was active against the PNGaseF-liberated N-glycans from human transferrin whereas it was previously shown to be inactive against a native form of the human transferrin (see **Figure 4.5**).

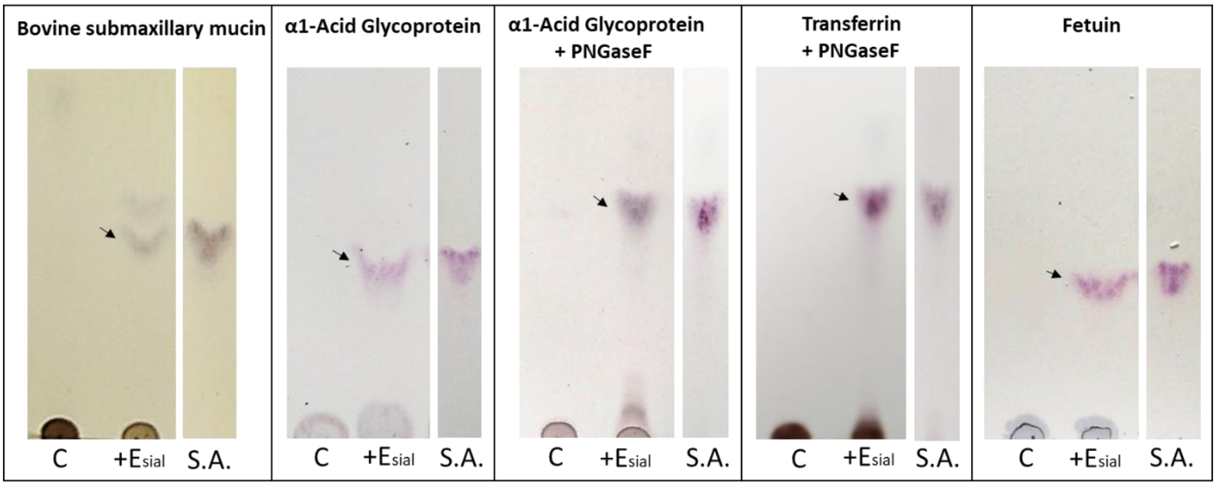


Figure 4.6: Screening for BT0455 activity against N-glycoproteins. C – N-glycoprotein control; +E$_{sial}$ – N-glycoprotein + BT0455^{GH33}. S.A. stands for Neu5Ac sialic acid control. Thin-layer chromatography (TLC) plates were stained using the DPA stain.

To visualise the sialic acid release in more detail, the enzyme assays were set up and products were visualised using the High-Performance Liquid Chromatography (HPLC/HPAEC-PAD) following a protocol described in Section 2.3.3. Known monosaccharide standards, such as sialic acid, were run alongside of the enzyme assay samples in order to accurately identify the different peaks. Due to the structural differences of the N-glycoproteins, different elution programs and HPLC columns were used, resulting in the differentiation in the peak elution pattern. The appearance of peaks representative of N-Acetyl-Neuraminic acid (Neu5Ac) was observed after incubating BT0455^{GH33} with bovine fetuin, PNGaseF-digested bovine fetuin, human transferrin and PNGase-F-digested bovine α$_1$-AGP (**Figure 4.7**). These results confirm that BT0455^{GH33} is a broad-acting sialidase that can uncap both bovine and human sialylated N-glycans. Based on the detection and elution pattern, the multiple unidentified peaks present in PNGaseF-cleaved fetuin and α_1-AGP samples could be representative of differentially de-sialylated N-glycan structures. However, without known N-glycan standards it is impossible to identify.

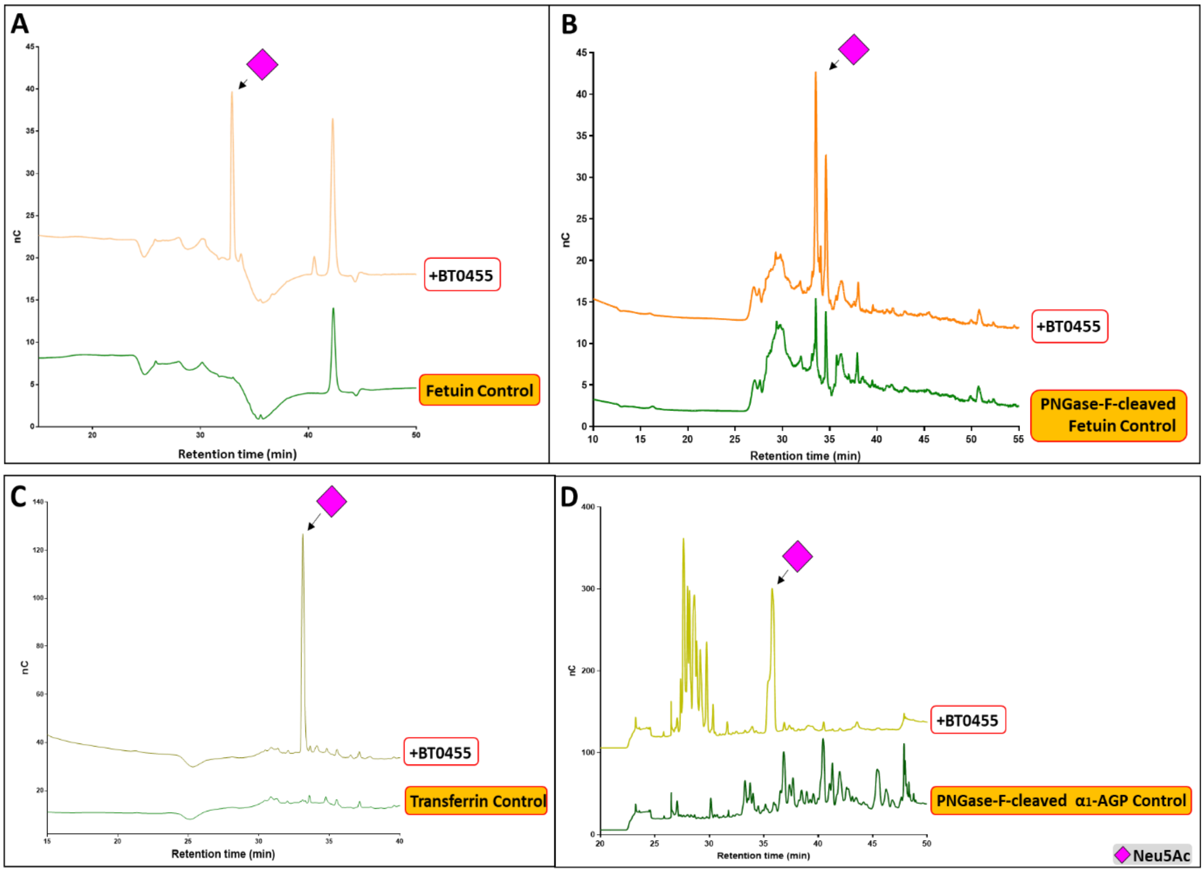
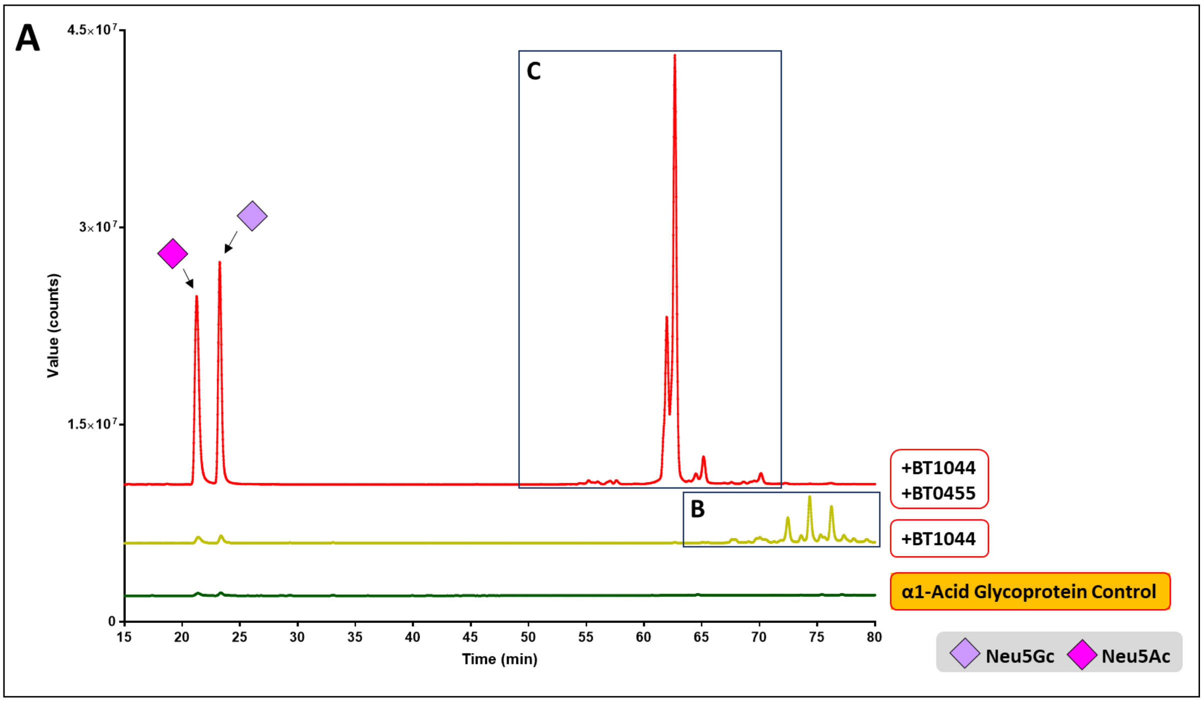


Figure 4.7: BT0455^{GH33} releases sialic acid from N-Glycoproteins. HPAEC-PAD chromatograms showing the release of sialic acid. 10mg/ml of N-glycoproteins were incubated with 1µM of BT0455^{GH33} enzyme in 20mM sodium phosphate (pH 7) for 1 hour at 37°C. 1:10 dilution was loaded onto the HPAEC-PAD. The final concentration of sialic acid standard (Neu5Ac) run is 1mM. **A.** Desialylation of Fetuin. **B.** Desialylation of PNGaseF-cleaved Fetuin. **C.** Desialylation of Transferrin. **D.** Desialylation of PNGaseF-cleaved α_1-Acid glycoprotein (α1-AGP).

To investigate the de-sialylation of N-glycan structures in more depth, the enzyme assay products were visualised using the U-HPLC combined with the mass spectrometry (LC-ESI-MS) analysis. The enzyme assay products were labelled with procainamide and sent for analyses using the equipment provided by our industrial partner Ludger following a protocol described in Section 2.3.8. Bovine α_1-AGP was chosen as a model N-glycan substrate. BT1044 is a GH18-family endo-β-N-acetylglucosaminidase that was previously shown to be highly upregulated in the RNA seq data

(**Figure 3.7.1**) (www.cazy.org). BT1044^{GH18} acts as an endo-β-*N*-acetylglucosaminidase, hydrolysing the bond between the chitobiose (GlcNAc-β1,4-GlcNAc) of the core of complex N-glycan to release the glycan from the protein (Dr. Lucy Crouch, unpublished data). Because *Bt* does not produce a PNGaseF-like enzyme and would not normally see PNGaseF-cleaved N-glycan structures, BT1044^{GH18} was used to pre-treat bovine α_1-AGP and cleave the N-glycans off the protein backbone prior the incubation with BT0455^{GH33}. The labelled enzyme assay products were analysed using combined U-HPLC and mass spectrometry, which allowed to accurately determine the mass of each peak and predict the N-glycan structure it corresponds to. The results are displayed in **Figure 4.8** and show that BT0455^{GH33} is capable of cleaving two different forms of sialic acid – N-Glycolyl-neuraminic acid (NeuGc) and N-Acetyl-Neuraminic acid (NeuAc). Human N-glycoproteins only contain NeuAc. These results suggest that BT0455^{GH33} is a broad-acting sialidase capable of removing all α2-3/6-linked sialic acid caps off N-glycan structures released from bovine α_1-AGP protein backbone by the endo-β-*N*-acetylglucosaminidase BT1044^{GH18}.



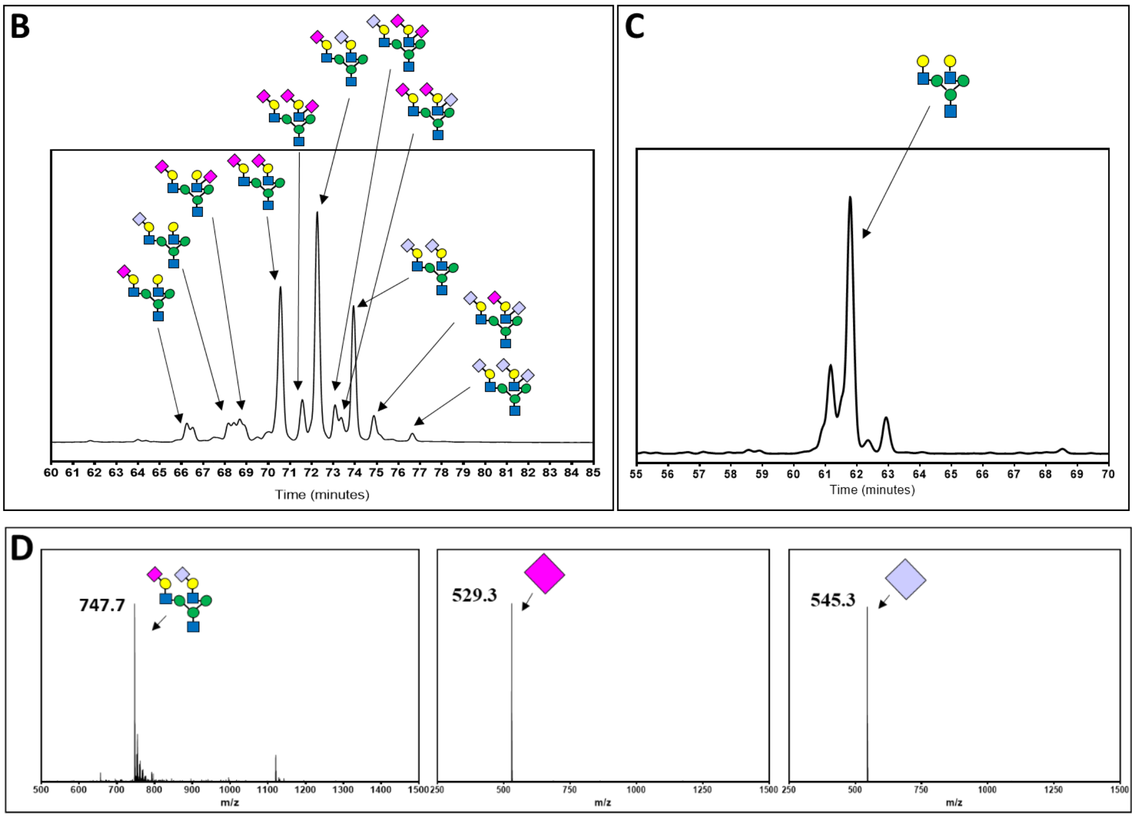
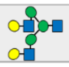
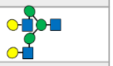
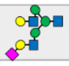
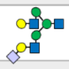
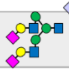
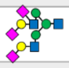
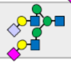
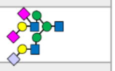
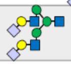


Figure 4.8: De-sialylation of bovine α$_1$-AGP N-glycan structures by BT0455^{GH33}.
A. HPLC analyses of α$_1$-AGP degradation by BT1044^{GH18} and BT0455^{GH33}. Figure **B** and Figure **C** show N-glycan structures corresponding to each peak that were constructed using the mass spectrometry data. **D.** Examples of mass spectrometry peaks with calculated masses [M/Z] and corresponding sugar structures.

MS data and predicted structure for each peak shown in **Figure 4.8** are displayed in **Table 4.1**

Table 4.1: Mass spectrometry data of GH18-cleaved bovine α$_1$-AGP degradation by BT0455^{GH33}.

Retention time (min)	Calculated mass [M/Z] (+2 H adducts)	Possible structure composition
21.40	529.30	◆
23.30	545.28	◇
61.90	829.35	
62.60	829.35	
67.80	652.27	
69.70	655.60	
70.40	747.30	
73.70	844.33	
74.40	752.6	
75.20	849.66	
76.10	757.96	

Retention time (min) corresponds to the peaks in the HPLC chromatograms.

Overall, the data presented in this section strongly supports the claim that the sialidase BT0455^{GH33} is a broad-acting enzyme that has an N-glycan specificity. BT0455^{GH33} was found to be capable of taking all α2-3/6-linked sialic acid caps off transferrin, fetuin and α_1-AGP, including Neu5Ac and Neu5Gc. BT0455^{GH33} was also capable of liberating different forms of sialic acid off bovine submaxillary mucin (BSM).

4.3.3. BT0457: Characterization of a sialic acid esterase

4.3.3.1. Introduction

Sialic acids are acetylated monosaccharides that are usually found capping the terminal branches of glycoconjugates, such as N-glycans, O-glycans and gangliosides. Since they are typically found on the terminal ends of glycan chains, they are a perfect target for the utilization by the gut microbes. However, acetylation of the sialic acids, such as Neu5Ac and Neu9Ac, can hinder the efficiency of the sialidase activity. Carbohydrate esterases (CEs) are enzymes that cleave acetyl groups off carbohydrate groups capping complex glycan structures, such as conjugated sialic acid caps.

B. thetaiotaomicron encodes BT0457 that, based on the amino acid sequence similarity and the domain composition, is predicted to be a sialic acid-specific 9-O-acetylesterase. The aim of this section was to investigate the activity of BT0457^{CE} esterase.

4.3.3.2. Bioinformatics analysis of BT0457^{CE}

In order to compare the BT0457^{CE} to other known esterases, sequence similarity database (SSDB) tool hosted by Kyoto Encyclopedia of Genes and Genomes (KEGG) database resource was used to generate a dendrogram of top 50 best orthologue matches to BT0457^{CE} (**Figure 4.9**). KEGG SSDB contains all known information about the amino acid sequence similarities between proteins in organisms that have been completely sequenced. The results displayed in **Figure 4.9** show that this type of esterase is widespread among the different species of bacteria, suggesting their importance in sialylated glycan utilization. The results also show that BT0457^{CE} is clustered together with esterases encoded by other members of *Bacteroides* species.

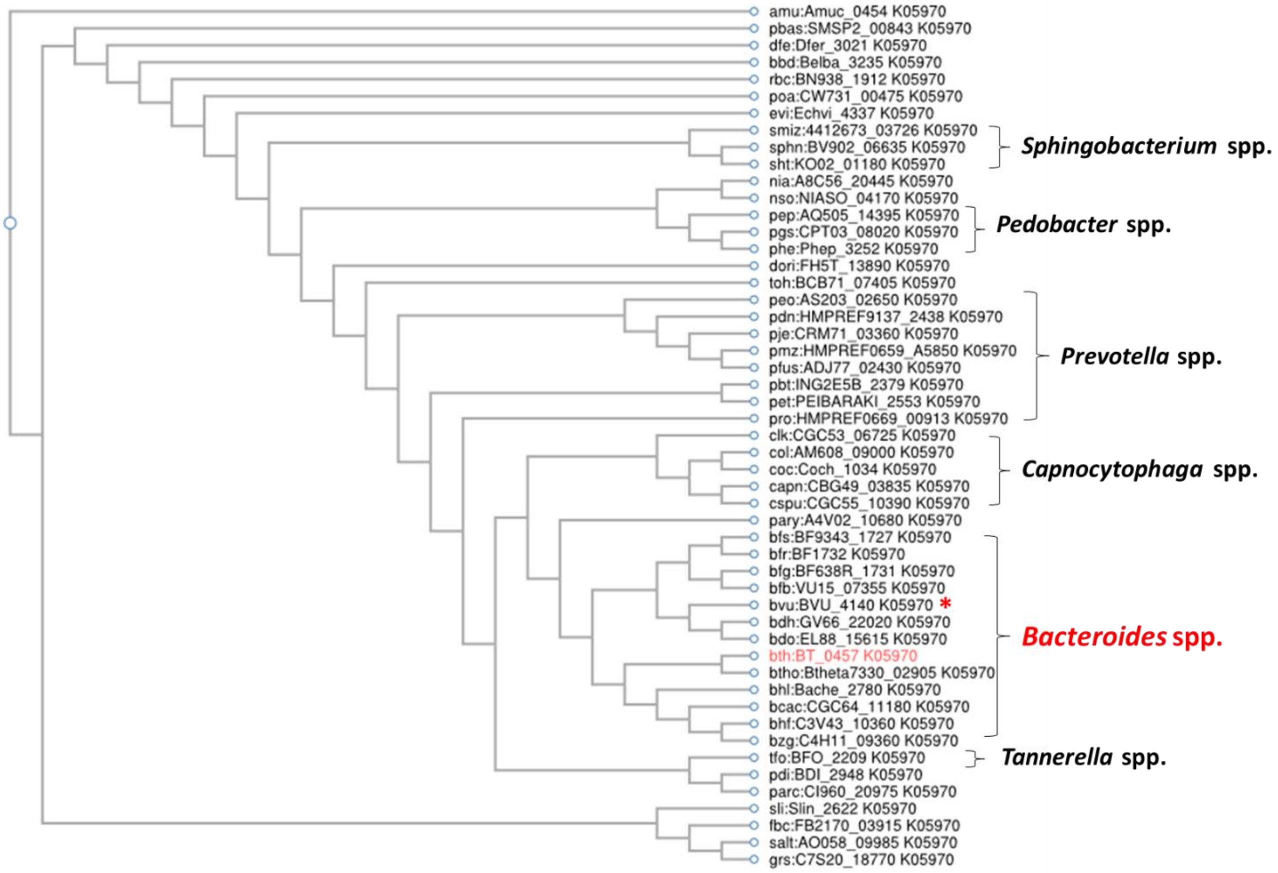


Figure 4.9: A dendrogram showing the taxonomic relationship of BT0457 to 50 orthologous esterase-coding genes. KEGG SSDB tool was used to perform a sequence similarity search on BT0457^{CE} and visualise the results by producing a dendrogram of 50 best matches. Dendogram was produced using Smith-Waterman similarity score to identify the best hits.

Based on the dendrogram displayed in **Figure 4.9**, *B. vulgatus* esterase BVU4140^{CE} appears to be closely related to BT0457^{CE}. Genome comparison showed that this esterase is predicted to be a part of *B. vulgatus* BVU4133-4145 PUL that is very similar to BT0455-0461 in *B. thetaiotaomicron*. BVU4133-4145 encodes a predicted sialic acid esterase BVU4140^{CE}, sialidase BVU4143^{GH33} and three GH20-family hexosaminidases. Furthermore, the sialidase BVU4143^{GH33} was previously shown to be homologous to BT0455^{GH33} (**Figure 4.4**). Previously characterised sialic acid-specific 9-O-acetylesterase (NanS) from *T. forsythia* was also chosen for structural comparison (Phansopa *et al.,* 2015). To investigate the similarities of BT0457^{CE} to BVU4140^{CE} and NanS from *T. forsythia*, the sequences were aligned using the Clustal Omega Multiple Sequence Alignment tool and visualised using the ESPript 3.0 tool. The results are displayed in **Figure 4.10**. Based on the fully conserved residues, these sialate-O-acetylesterases appear to be closely related and have a sequence similarity of over 60%. SGNH family of esterases are defined by the presence of four blocks of conserved amino acid sequences. Block I contains the G-D/N-S motif, which contains the nucleophilic serine residue. Block II contains the glycine (G), which is conserved among all members of SGNH family of hydrolases. Block III contains the conserved asparagine (N) residue and Block IV contains two conserved catalytic residues, aspartic acid (D) and histidine (H) (Rangarajan *et al.,* 2011).

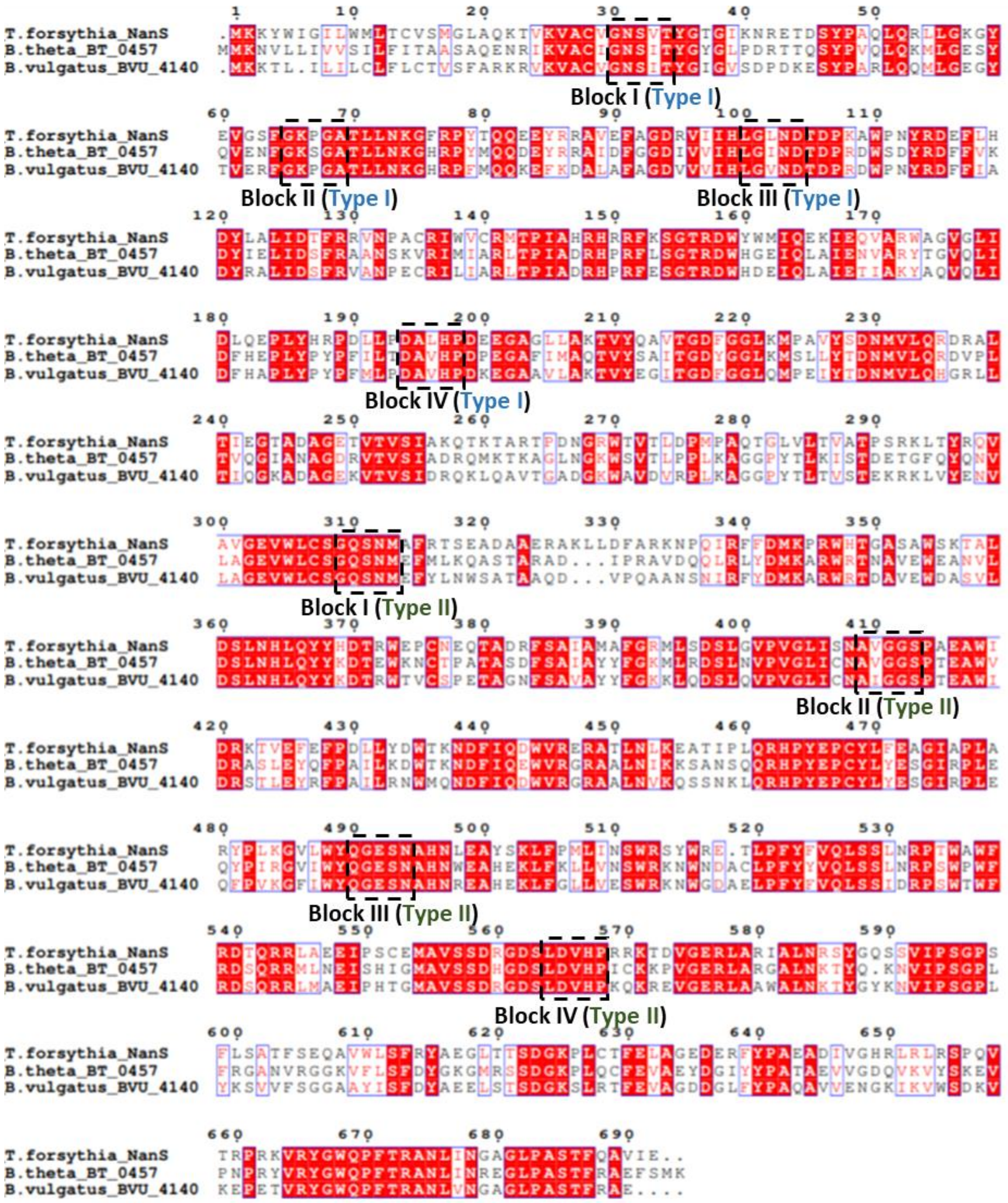


Figure 4.10: Amino acid sequence alignments of confirmed and putative sialate O-acetylesterases from *T. forsythia* (NanS), *B. thetaiotaomicron* (BT_0457) and *B. vulgatus (*BVU_4140). Conserved amino acid blocks are displayed where block I contains G-D/N-S motif, block II contains the glycine (G), block III contains asparagine (N) and block IV contains aspartic acid (D) and histidine (H). Blocks are further divided in Type I, II and III according to the conserved SGNH residues present. Residues highlighted in red boxes are 100% conserved. Amino acid residues framed in blue boxes are 70% or more identical based on their physicochemical properties.

The conserved residues of these four blocks are characteristic of the SGNH family of hydrolases. Sialate-O-acetylesterases BT0457^{CE}, BVU4140^{CE} and NanS from *T. forsythia* have type I and type II of SGNH conserved residues which indicates they possess two structural domains (**Figure 4.11**). The domain representative of the type I of SGNH family of hydrolases is displayed as Lipase_GDSL_2 by

PFAM database. Meanwhile, DUF303 domain is predicted to belong to the type II domain of the SGNH family hydrolases (Phansopa *et al.,* 2015). The results show that all three sialate-O-acetylesterases have an identical domain composition, suggesting they could have very similar enzymatic activities and target same sialylated glycan structures.

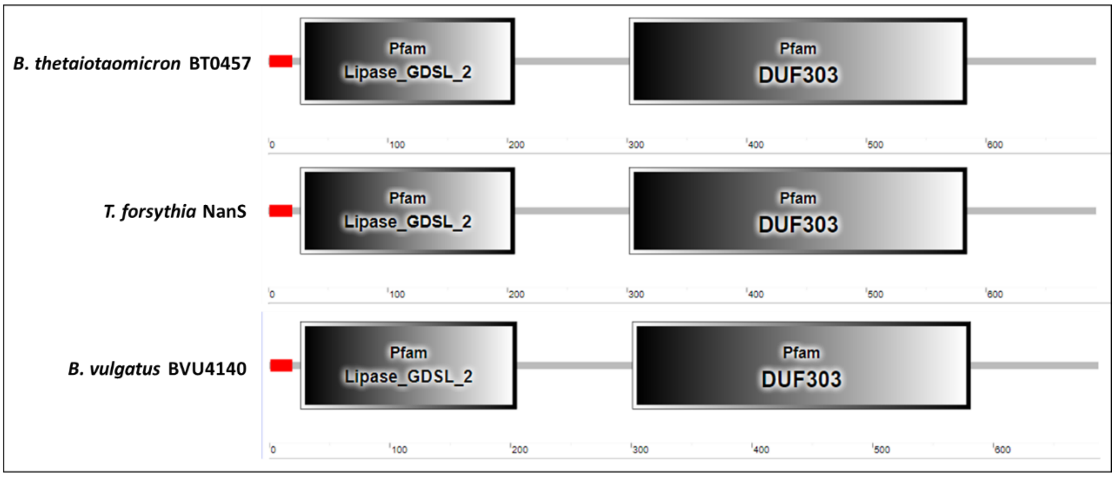


Figure 4.11: Domain architecture of BT0457, NanS and BVU4140 esterases. SMART tool was used to visualise the domain composition. Predicted signal sequence is shown in red. Pfam domains are displayed in grey boxes. Amino acid sequence length is displayed below the illustrated domains.

In order to investigate and compare the amino acid sequence and domain composition of BT0457^{CE} to other esterases in *B.thetaiotaomicron*, Kyoto Encyclopedia of Genes and Genomes (KEGG) database resource was used to identify paralogous protein-coding genes. The sequences were aligned using the Clustal Omega Multiple Sequence Alignment tool and visualised using the ESPript 3.0 tool. The results show that the amino acid sequence of BT0457^{CE} is not closely related to any other esterases produced by *B.thetaiotaomicron*, but all enzymes seem to share blocks of fully conserved residues highlighted in red in the **Figure 4.12**. One of these conserved amino acid blocks contains conserved residues of Block III (Type II) that belong to the predicted type II SGNH family of hydrolases.

It is better illustrated in the **Figure 4.13** showing the PFAM domain composition. The results show that BT0457^{CE} is the only esterase to contain the domain of the type I of SGNH family of hydrolases, named Lipase_GDSL_2 in Pfam. Interestingly, majority of them seem to contain DUF303 domain, predicted to be representative of the type II domain of the SGNH family of hydrolases.

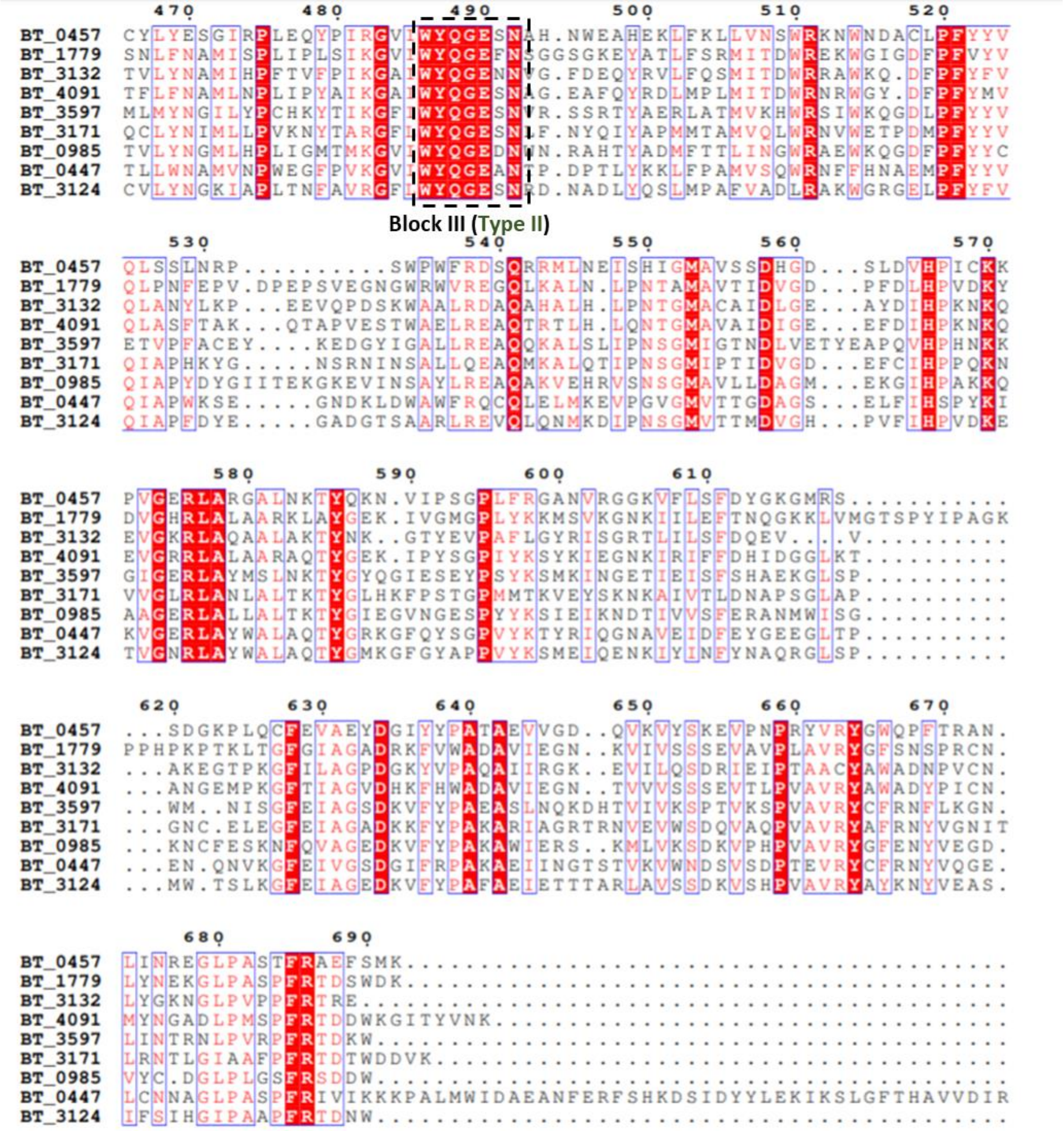

470 480 490 500 510 520
BT_0457 CYLYESGIRPLEQYPIRGVIWYQGESNAH.NWEAHEKLFKLLVNSWRKNWNDACLPFYYV
BT_1779 SNLFNAMISPLIPLSIKGVIWYQGEFNSGGSGKEYATLFSRMITDWREKWGIGDFPFVYV
BT_3132 TVLYNAMIHPFTVFPIKGAIWYQGENNVG.FDEQYRVLFQSMITDWRRAWKQ.DFPFYFV
BT_4091 TFLFNAMLNPLIPYAIKGAIWYQGESNAG.EAFQYRDLMPLMITDWRNRWGY.DFPFYMV
BT_3597 MLMYNGILYPCHKYTIKGFIWYQGESNVR.SSRTYAERLATMVKHWRSIWKQGDLPFYYV
BT_3171 QCLYNIMLLPVKNYTARGFIWYQGESNLF.NYQIYAPMMTAMVQLWRNVWETPDMPFYYV
BT_0985 TVLYNGMLHPLIGMTMKGVIWYQGEDNWN.RAHTYADMFTTLINGWRAEWKQGDFPFYYC
BT_0447 TLLWNAMVNPWEGFPVKGVIWYQGEANTP.DPTLYKKLFPAMVSQWRNFFHNAEMPFYYV
BT_3124 CVLYNGKIAPLTNFAVRGFLWYQGESNRD.NADLYQSLMPAFVADLRAKWGRGELPFYFV
Block III (Type II)
530 540 550 560 570
BT_0457 QLSSLNRP..........SWPWFRDSQRRMLNEISHIGMAVSSDHGD...SLDVHPICKK
BT_1779 QLPNFEPV.DPEPSVEGNGWRWVREGQLKALN.LPNTAMAVTIDVGD...PFDLHPVDKY
BT_3132 QLANYLKP...EEVQPDSKWAALRDAQAHALH.LPNTGMACAIDLGE...AYDIHPKNKQ
BT_4091 QLASFTAK...QTAPVESTWAELREAQTRTLH.LQNTGMAVAIDIGE...EFDIHPKNKQ
BT_3597 ETVPFACEY....KEDGYIGALLREAQQKALSLIPNSGMIGTNDLVETYEAPQVHPHNKK
BT_3171 QIAPHKYG.....NSRNINSALLQEAQMKALQTIPNSGMIPTIDVGD...EFCIHPPQKN
BT_0985 QIAPYDYGIITEKGKEVINSAYLREAQAKVEHRVSNSGMAVLLDAGM...EKGIHPAKKQ
BT_0447 QIAPWKSE.....GNDKLDWAWFRQCQLELMKEVPGVGMVTTGDAGS...ELFIHSPYKI
BT_3124 QIAPFDYE.....GADGTSAARLREVQLQNMKDIPNSGMVTTMDVGH...PVFIHPVDKE
580 590 600 610
BT_0457 PVGERLARGALNKTYQKN.VIPSGPLFRGANVRGGKVFLSFDYGKGMRS...........
BT_1779 DVGHRLALAARKLAYGEK.IVGMGPLYKKMSVKGNKIILEFTNQGKKLVMGTSPYIPAGK
BT_3132 EVGKRLAQAALAKTYNK..GTYEVPAFLGYRISGRTLILSFDQEV....V..........
BT_4091 EVGRRLALAARAQTYGEK.IPYSGPIYKSYKIEGNKIRIFFDHIDGGLKT..........
BT_3597 GIGERLAYMSLNKTYGYQGIESEYPSYKSMKINGETIEISFSHAEKGLSP..........
BT_3171 VVGLRLANLALTKTYGLHKFPSTGPMMTKVEYSKNKAIVTLDNAPSGLAP..........
BT_0985 AAGERLALLALTKTYGIEGVNGESPYYKSIEIKNDTIVVSFERANMWISG..........
BT_0447 KVGERLAYWALAQTYGRKGFQYSGPVYKTYRIQGNAVEIDFEYGEEGLTP..........
BT_3124 TVGNRLAYWALAQTYGMKGFGYAPPVYKSMEIQENKIYINFYNAQRGLSP..........
620 630 640 650 660 670
BT_0457 ...SDGKPLQCFEVAEYDGIYYPATAEVVGD..QVKVYSKEVPNPRYVRYGWQPFTRAN.
BT_1779 PPHPKPTKLTGFGIAGADRKFVWADAVIEGN..KVIVSSSEVAVPLAVRYGFSNSPRCN.
BT_3132 ...AKEGTPKGFILAGPDGKYVPAQAIIRGK..EVILQSDRIEIPTAACYAWADNPVCN.
BT_4091 ...ANGEMPKGFTIAGVDHKFHWADAVIEGN..TVVVSSSEVTLPVAVRYAWADYPICN.
BT_3597 ...WM..NISGFEIAGSDKVFYPAEASLNQKDHTVIVKSPTVKSPVAVRYCFRNFLKGN.
BT_3171 ...GNC.ELEGFEIAGADKKFYPAKARIAGRTRNVEVWSDQVAQPVAVRYAFRNYVGNIT
BT_0985 ...KNCFESKNFQVAGEDKVFYPAKAWIERS..KMLVKSDKVPHPVAVRYGFENYVEGD.
BT_0447 ...EN.QNVKGFEIVGSDGIFRPAKAEIINGTSTVKVWNDSVSDPTEVRYCFRNYVQGE.
BT_3124 ...MW.TSLKGFEIAGEDKVFYPAFAEIETTTARLAVSSDKVSHPVAVRYAYKNYVEAS.
680 690
BT_0457 LINREGLPASTFRAEFSMK...................................
BT_1779 LYNEKGLPASPFRTDSWDK...................................
BT_3132 LYGKNGLPVPPFRTRE......................................
BT_4091 MYNGADLPMSPFRTDDWKGITYVNK.............................
BT_3597 LINTRNLPVRPFRTDKW.....................................
BT_3171 LRNTLGIAAFPFRTDTWDDVK.................................
BT_0985 VYC.DGLPLGSFRSDDW.....................................
BT_0447 LCNNAGLPASPFRIVIKKKPALMWIDAEANFERFSHKDSIDYYLEKIKSLGFTHAVVDIR
BT_3124 IFSIHGIPAAPFRTDNW.....................................

Figure 4.12: Amino sequence alignments of confirmed and putative esterases from *B. thetaiotaomicron* paralogous to BT0457. KEGG database was used to identify the paralogues. The figure shows the alignment of sequences between 467 and 740 amino acids. Amino acids highlighted in red boxes are fully conserved. Amino acid residues framed in blue boxes are 70% or more identical based on their physicochemical properties

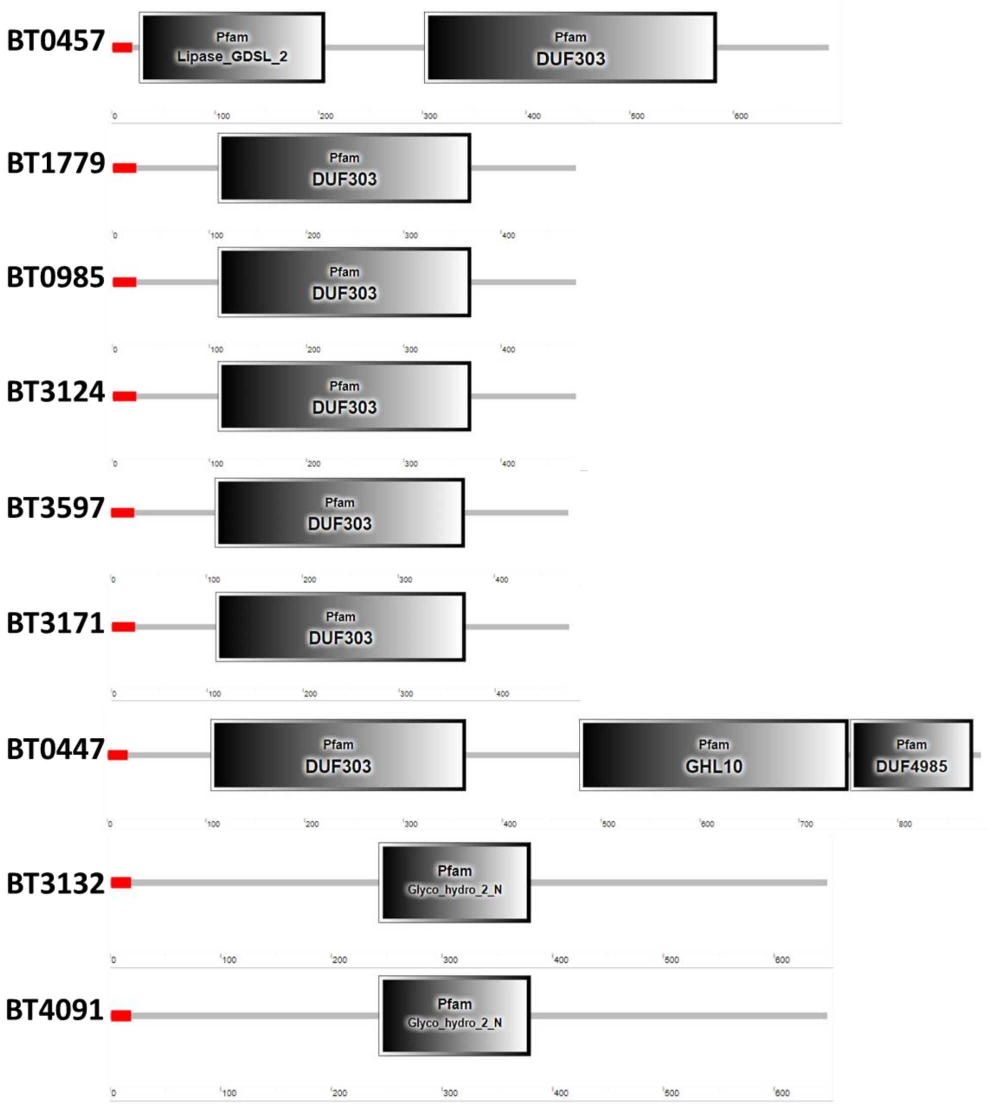


Figure 4.13: Domain composition of *B. thetaiotaomicron* esterases paralogous to BT0457. Predicted signal peptide sequence is displayed in red. Pfam domains are shown in grey boxes. Amino acid sequence length is displayed below the domains.

4.3.3.3. Screening for substrate specificity of BT0457^{CE}

RNA seq data analysis displayed in **Figure 3.7.1** shows that the esterase BT0457^{CE} is upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources, compared to the glucose control. To further investigate the activity of the predicted sialic acid-specific 9-O-acetylesterase BT0457^{CE} against sialylated carbohydrates, enzyme assays were set up following a protocol described in Section 2.3.5. To initially assess the BT0457^{CE} enzyme activity, colorimetric assays with para-Nitrophenol (pNP)-linked acetate were set up. If the enzyme showed activity on the linked substrate, the colour of the reaction turned yellow. The results showed that BT0457^{CE} is active against pNP-acetate. However, when it was attempted to obtain the kinetic parameters of BT0457^{CE} activity against the pNP-acetate following a protocol described in Section 2.3.5.1, it displayed a really low enzymatic efficiency and thus, the rate of substrate hydrolysis could not be measured. This may be due to the fact that pNP-acetate is synthetic and not a true substrate for BT0457^{CE}.

To visualise the de-acetylation of the sialic acid by BT0457^{CE} in more detail, the enzyme assays against Bovine Submaxillary Mucin (BSM) were set up in sodium phosphate 20mM (pH 7) and incubated overnight at 37 °C. The enzyme assay products were visualised using the HPAEC-PAD following a protocol described in Section 2.3.3. Bovine Submaxillary Mucin (BSM) was chosen because it is heavily sialylated by a number of different variants of sialic acids (Zauner *et al.,* 2012). A known N-Acetylneuraminic acid standard (Neu5Ac) was run alongside of the enzyme assay samples in order to accurately identify this sialic acid peak. Neu9Ac was not available to purchase commercially at the time of this experiment, thus could not be used. The results show that N-Acetylneuraminic acid (Neu5Ac), cleaved by BT0455^{GH33} from the bovine submaxillary mucin (BSM), is not a substrate for the BT0457^{CE} esterase (**Figure 4.14**). Based on the change in peak intensity, the results suggest that the treatment with BT0457^{CE} enables the sialidase BT0455^{GH33} to cleave more Neu5Ac from the BSM glycans. The peaks labelled 1, 2, 3 and 4 could not be identified but based on the intensity of the peak labelled number 2, it could represent a form of a de-acetylated sialic acid. These peaks could also represent different de-sialylated forms of *O*- or *N*-linked glycan structures present on BSM.

Due to high sequence similarity to *T. forsythia* NanS and no detected activity on Neu5Ac, the overall results indicate that BT0457^{CE} could be a sialic acid-specific 9-O-acetylesterase. However, due to time constraints, more in-depth analyses could not be done to confirm this claim.

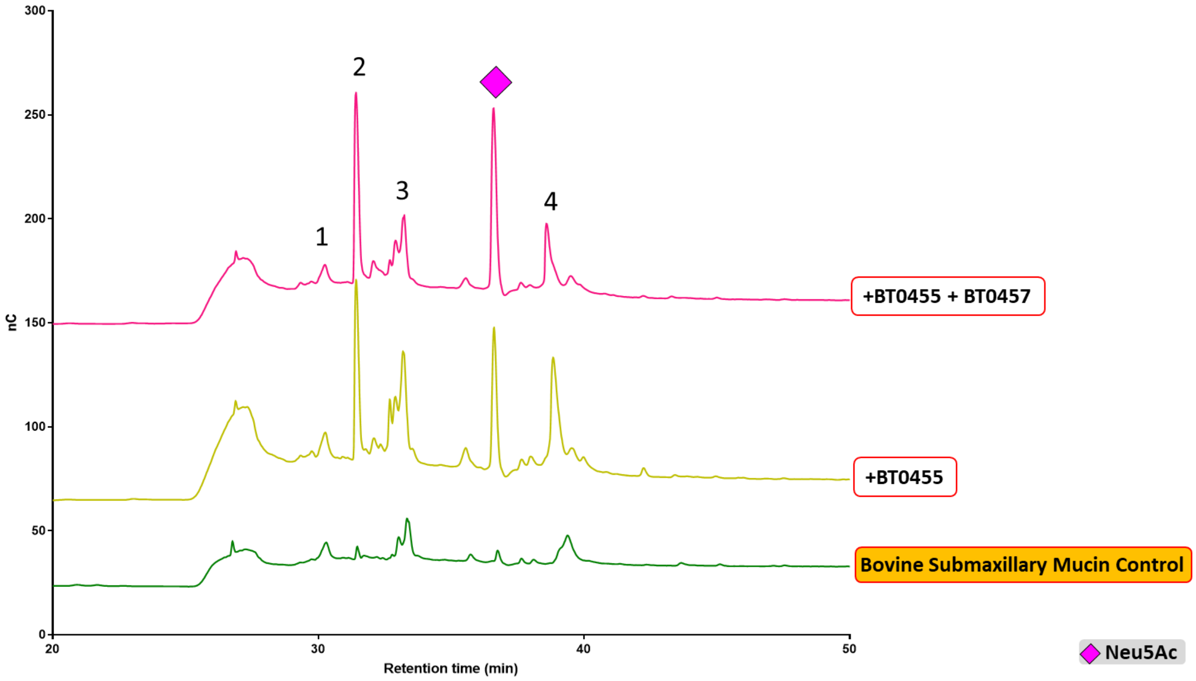


Figure 4.14: De-acetylation of sialic acid released from Bovine Submaxillary Mucin (BSM). HPLC analyses of bovine submaxillary mucin (BSM) degradation by BT0455^{GH33} and BT0457^{CE} are shown. 1mM of sialic acid standard was used (Neu5Ac). Peaks labelled 1-4 could not be identified using known carbohydrate standards.

4.3.4. BT0461: Characterization of a galactosidase activity

4.3.4.1. Introduction

The complete degradation of N-glycan structures depends a cooperative activity of multiple carbohydrate-active enzymes. One of the key enzymes required is a β-galactosidase. *B. thetaiotaomicron* possesses a number of β-galactosidases but according to the CAZy database, only five have been characterized to date. BT0461 is a GH2-family β-galactosidase that is predicted to target terminal β1-3/6- and/or β1-4- D-galactosyl bonds. BT0461^{GH2}, like the sialidase BT0455^{GH33}, was found to be upregulated when *B. thetaiotaomicron* was grown on porcine gastric mucin (PGMIII) but its precise substrate specificity remains unknown (Martens *et al.,* 2008).

As previously shown in **Figure 3.8**, BT0461^{GH2}, along with the other genes from the BT0455-0461 locus, is upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as a sole carbon source. Thus, the aim of this chapter was to investigate and characterise the substrate specificity of BT0461^{GH2}.

4.3.4.2. Bioinformatics analysis of BT0461^{GH2}

To investigate the similarities of BT0461^{GH2} to other characterised β-galactosidases, the amino acid sequence of BT0461^{GH2} was used as a query for BLASTP and KEGG search to identify the homologous proteins. The identified putative β-galactosidase orthologues and suspected paralogues were used for phylogenetic and bioinformatics analysis.

Initially, the sequence similarity database (SSDB) tool hosted on KEGG database was used to generate a dendrogram of top 50 best matches to BT0461^{GH2} (**Figure 4.15**). The results show that BT0461^{GH2} type of GH2 β-galactosidase is widespread and has homologs in a variety of species of commensal and pathogenic bacteria. The orthologues of BT0461^{GH2} β-galactosidase found in other *Bacteroides* species are highlighted in Cluster 1 and Cluster 2, suggesting they may have different substrate specificities. Because the closest biochemically and structurally related characterised GH2 β-galactosidase to BT0461^{GH2} was identified as LacZ in *E.coli*, which shares only 24% homology, substrate preferences of BT0461^{GH2} remain to be identified.

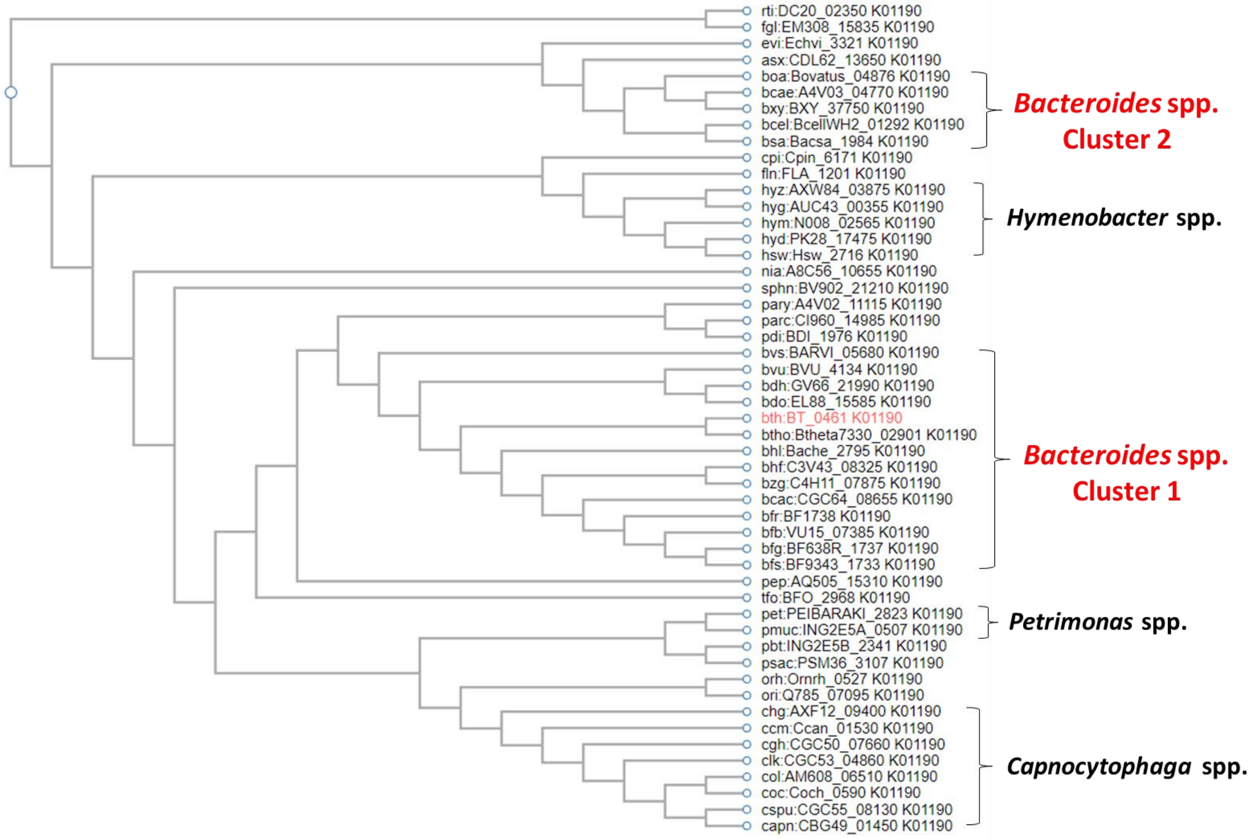


Figure 4.15: A dendrogram showing the taxonomic relationship of BT0461^{GH2} to its 50 best orthologous β-galactosidase-coding genes. KEGG SSDB tool was used to perform a sequence similarity search on BT0461^{GH2} and visualise the results by producing a dendrogram of best orthologues. Smith-Waterman similarity score was used to identify the best hits.

To investigate the sequence similarities of BT0461^{GH2} to closely related β-galactosidases further, BVU4134^{GH2} and Ccan01450^{GH2} putative β-galactosidases were chosen for the comparison and PFAM domain composition was obtained using the protein sequences and Simple Modular Architecture Tool (SMART). It was found that the amino acid sequence of BT0461^{GH2} is 60% and 78% homologous to BVU4134^{GH2} and Ccan01450^{GH2}, respectively, and these enzymes share a very similar domain composition (**Figure 4.16**). BT0461^{GH2} appears to have an identical domain composition to the BVU4134^{GH2}, further supporting the previously discussed claim that *B. vulgatus* BVU4133-4145 PUL is very similar to BT0455-0461 locus in *B. thetaiotaomicron*. Ccan01450^{GH2} shares Glyco_hydro_2, Glyco_hydro_2C and DUF4982 PFAM domains with BT0461^{GH2} and BVU4134^{GH2}. These domains are characteristic of a typical GH2-family β-galactosidase (CAZypedia.org, GH2 family). In overall, based on these results it appears to be safe to hypothesise that these enzymes could have similar substrate specificities.

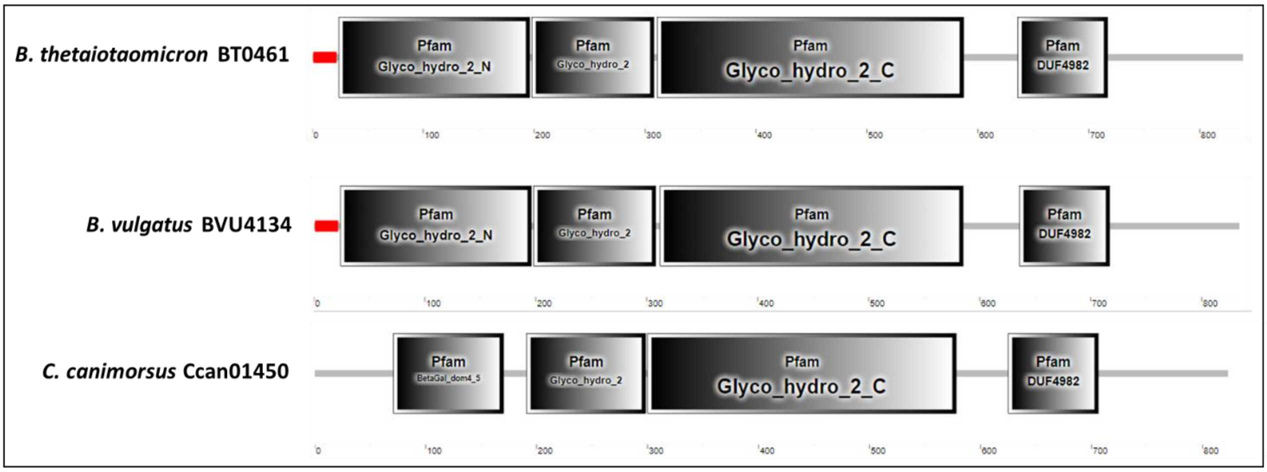


Figure 4.16: Domain composition of GH2-family β-galactosidases. PFAM domain composition of β-galactosidases from *B. thetaiotaomicron* (BT0461^{GH2}), *B. vulgatus* (BVU4134^{GH2}) and *C. canimorsus* (Ccan01450^{GH2}) are displayed in grey boxes. Predicted signal peptide sequence is displayed in red. Amino acid sequence length is displayed below the domains. DUF stands for domain of unknown function.

4.3.4.3. Defining the substrate specificity of BT0461^{GH2}

To initially assess BT0461^{GH2} enzyme activity, assays with para-Nitrophenol (pNP)-linked substrates were set up following a protocol described in Section 2.3.5.1 using 1 mM of pNP-substrates and incubation for 30 min at 37 °C . As predicted by the bioinformatics analyses (see Section 4.3.4.2), BT0461^{GH2} was only active against p-nitrophenyl-β-D-galactopyranoside (pNP-β-Gal) and showed no activity against p-nitrophenyl-α-D-galactopyranoside (pNP-α-Gal), p-nitrophenyl-β-D-mannopyranoside (pNP-β-Man) nor p-nitrophenyl-β-D-glucopyranoside (pNP-β-Glc), confirming it is a β-galactosidase. To further investigate the BT0461^{GH2} specificity, enzyme assays against various disaccharide structures were set up following a protocol described in Section 2.3.5. The assays were incubated for 16 hours at 37 °C and reaction products were visualised using the TLC.

The positive results were identified by the release of galactose monosaccharide and showed that β-galactosidase BT0461^{GH2} is active against all galactose disaccharides that have β1,4 linkages (**Figure 4.17**). These include β1,4-Galactobiose, Gal-β1,4-GlcNAc (LacNAc) and Gal-β1,4-Glucose (Lactose). BT0461^{GH2} showed no activity against β1,6-Galactobiose, suggesting it may have a β1,4-linkage specificity.

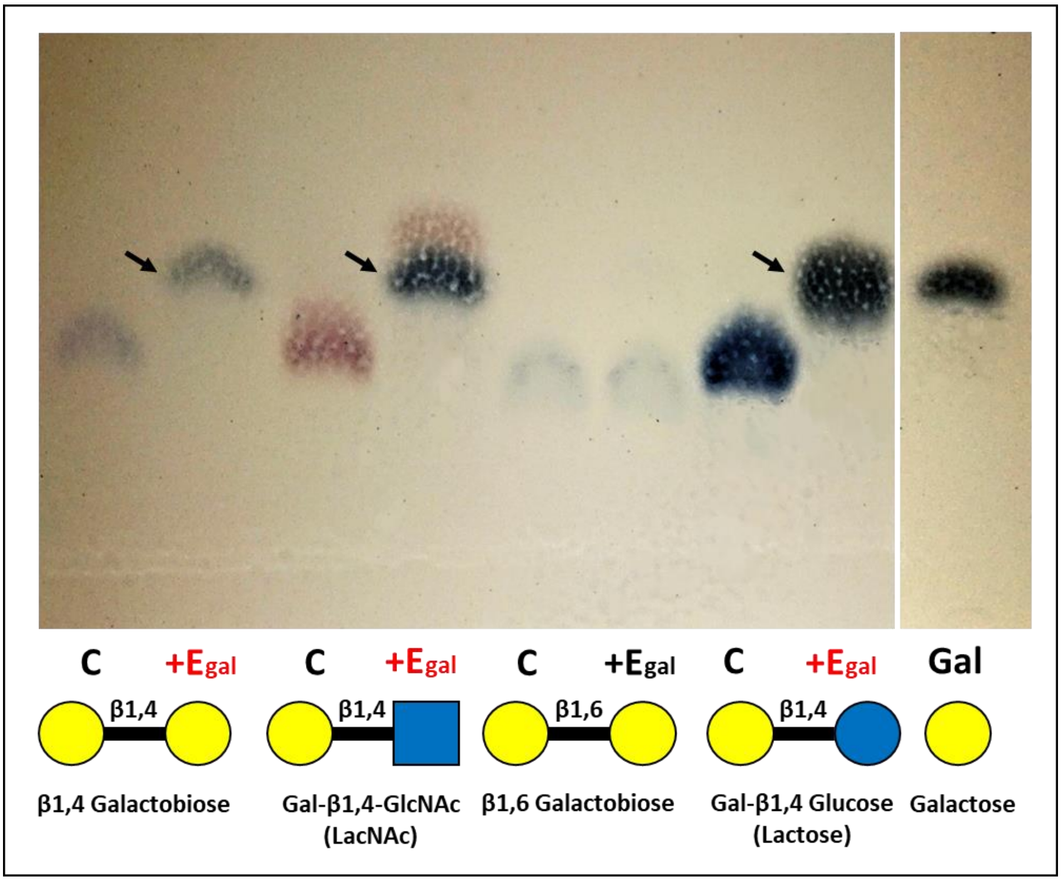
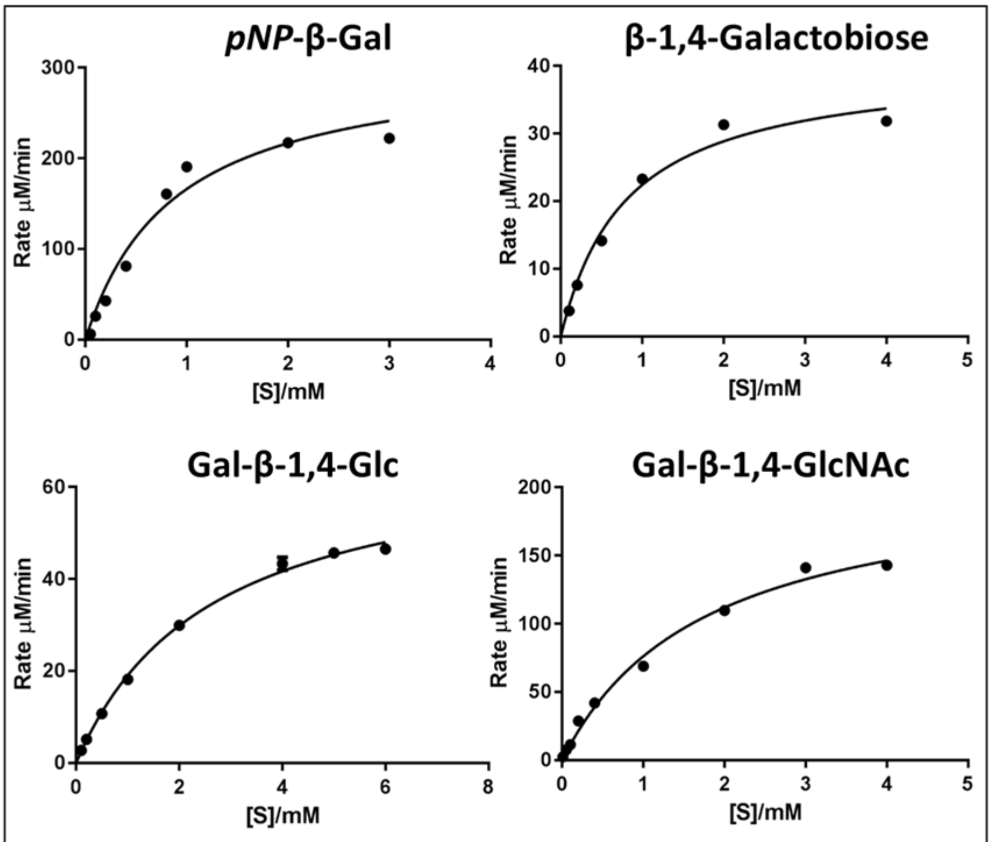


Figure 4.17: Screening for the enzyme activity of BT0461^{GH2} against simple disaccharides. 1µM of BT0461^{GH2} was incubated with 1-2mM of disaccharides. C – disaccharide control; +E$_{gal}$ – substrate incubated with BT0461^{GH2}. Gal is 1mM Galactose standard. Thin-layer chromatography (TLC) plates were stained using the DPA stain.

To determine the substrate preference of the retaining β-galactosidase BT0461^{GH2}, the catalytic activity of the enzyme against different β1,4-linked disaccharides was measured at 340nm using D-Galactose detection kit (Megazyme) following the manufacturer's instructions (see Section 2.3.5.2). BT0461^{GH2} activity against p-nitrophenyl-β-D-galactopyranoside (pNP-β-Gal) was measured at 320nm following a protocol described in Section 2.3.5.1. The kinetic parameters and respective kinetic curves for each disaccharide are shown in **Figure 4.18**. The enzyme kinetics results show that $BT0461^{GH2}$ displays the highest activity against the pNP-β-Gal, however it is a synthetic substrate and not found in nature. Based on the k_{cat}/K_M values, $BT0461^{GH2}$ shows a 2-fold preference for Gal-β1,4-GlcNAc over Gal-β1,4-Gal. Gal-β1,4-GlcNAc is a structure commonly found in N-glycan structures. Although $BT0461^{GH2}$ is active on lactose (Gal-β 1,4-Glc), based on low k_{cat}/K_M value, it does not appear to be a preferred substrate.



	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0461	pNP-β-Gal	3118	0.87±0.127	3584
	Gal-β1,4-GlcNAc	2109	1.77±0.181	1191
	Gal-β1,4-Gal	405	0.8±0.08	506
	Gal-β1,4-Glc	138	2.6±0.148	54

Figure 4.18: Activity of $BT0461^{GH2}$ vs disaccharide substrates. All reactions were performed at 37 °C using 20 mM NaH_2PO_4 (pH 7.5) buffer. Reaction rates were measured at 320 nm (for pNP-Gal) and 340 nm for disaccharides and plotted in GraphPad Prism 7.0 software using non-regression analysis (see Section 2.3.5). The standard errors were generated from biological triplicates.

To investigate whether BT0461^{GH2} can catalyse the hydrolysis of the Gal-β1,4-GlcNAc(α1,3-Fuc) trisaccharide, where fucose could potentially interfere with the BT0461^{GH2} activity, enzyme assays against Lacto-N-fucopentaose III were set up following the same protocol used above (see Section 2.3.2). An α-fucosidase BT1625^{GH29} was incubated together with the BT0461^{GH2} to investigate if it has any effect. Based on the RNA-seq data, BT1625^{GH29} is also upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources (Supplemental table 3.1). Based on the disappearance of the band representative of a released galactose monosaccharide, the results showed that β-galactosidase BT0461^{GH2} is unable to take the galactose off when the fucose is linked onto the GlcNAc (**Figure 4.19**). However, once the fucose is released by the fucosidase BT1625^{GH29}, β-galactosidase BT0461^{GH2} can hydrolyse the Gal-β1,4-GlcNAc linkage.

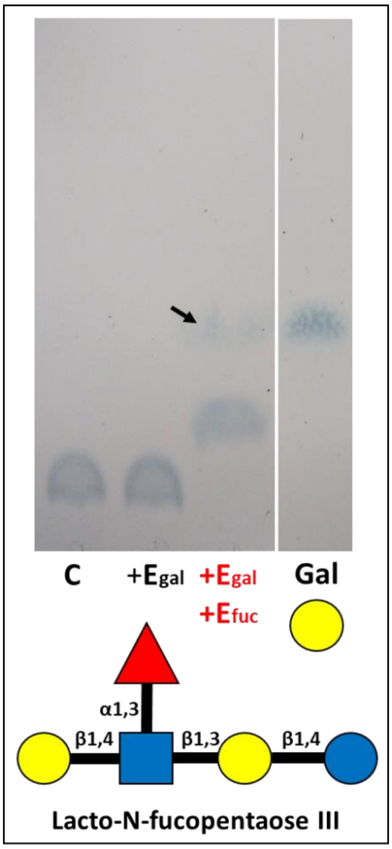


Figure 4.19: BT0461^{GH2} activity against Lacto-N-fucopentaose III. 1μM of enzymes were incubated with 1mM of pentasaccharide. C – substrate control; +E$_{gal}$ – substrate + BT0461. +E$_{fuc}$ – substrate + BT1625^{GH29}. Gal is 1mM Galactose standard. TLC plates were stained using the DPA stain. DPA could not stain fucose. Arrow indicates the galactose product.

To investigate BT0461^{GH2} activity on N-glycan substrates, enzyme assay against bovine α$_1$-Acid glycoprotein and bovine fetuin was set up. Considering the majority of N-glycans are capped by terminal sialic acids, the sialidase BT0455^{GH33} was also included in the fetuin assay. Enzyme activity assays of 1 µM of sialidase BT0455^{GH33} and 1 µM of β-galactosidase BT0461^{GH2} against 10mg/ml of N-glycoproteins were set up following a protocol described in Section 2.3.5. The enzyme assays were incubated in sodium phosphate 20mM (pH 7) buffer for 1 hour at 37 °C and visualised using the SDS-PAGE gel and TLC (see Section 2.3.2; Section 2.1.22). The results suggest that BT0461^{GH2} is an exo-glycosidase that cannot cleave sialylated galactose residues off the α$_1$-AGP N-glycan structures (**Figure 4.20**). BT0461^{GH2} requires the pre-treatment with the sialidase BT0455^{GH33} to remove the sialic acid caps off the sialylated α$_1$-AGP and fetuin N-glycans before it can cleave galactose caps off. Thus, the sequential release of sialic acid and galactose by BT0455^{GH33} and BT0461^{GH2}, respectively, was observed (**Figure 4.20-A and B**). Deglycosylation of bovine fetuin N-glycans by BT0455^{GH33} and BT0461^{GH2} was also observed by a decrease in N-glycoprotein size from ~64 kDa to ~55 kDa (**Figure 4.20-C**).

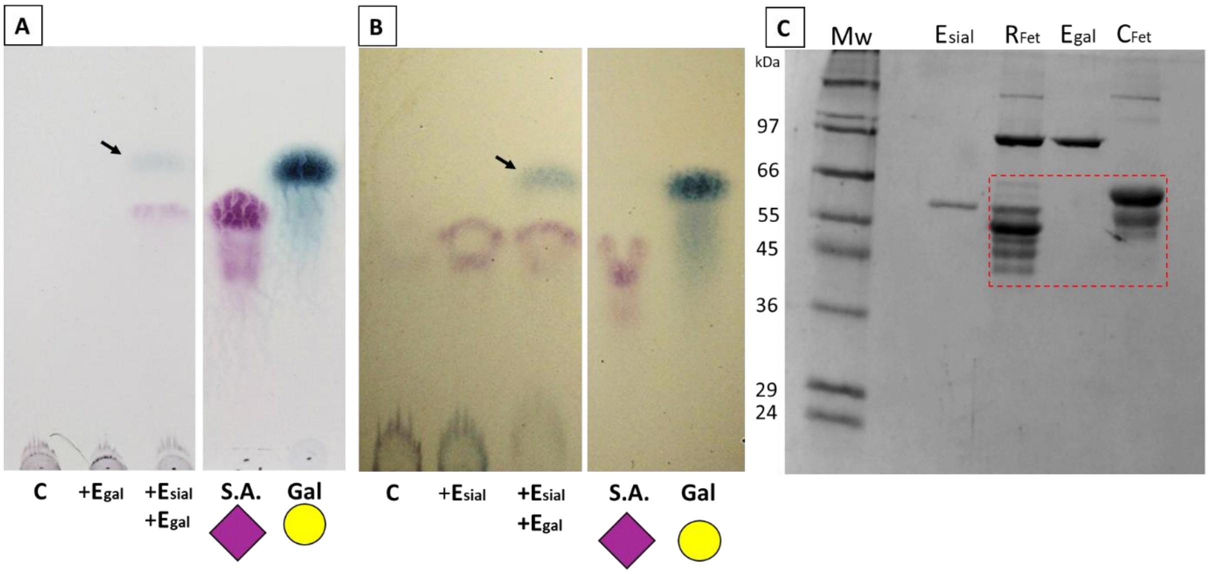


Figure 4.20: N-glycan degradation by the sialidase BT0455^{GH33} and β-galactosidase BT0461^{GH2}.
A) α$_1$-acid glycoprotein degradation. **B)** Bovine fetuin degradation. TLC plates showing sialic acid and galactose release by BT0455^{GH33} and BT0461^{GH2}, respectively. E$_{sial}$ is BT0455; E$_{gal}$ is BT0461; S.A. is 2mM sialic acid (Neu5Ac) standard; Gal is 1mM galactose standard. DPA stain was used to stain the TLC plates. Arrow indicates the galactose product. **C)** SDS-PAGE showing Fetuin N-glycoprotein degradation by BT0455^{GH33} and BT0461^{GH2} (highlighted in red box). Mw is a wide-range molecular marker (Sigma Aldrich). R$_{Fet}$ is the reaction mixture and C$_{Fet}$ is control. 1:20 dilutions of the assays were loaded. Degradation is observed by the decrease in N-glycoprotein band size in reaction lane compared to the control.

To further investigate the enzymatic BT0461^{GH2} activity against complex de-sialylated N-glycan structures, enzyme assays against native, PNGaseF- and GH18-digested bovine α$_1$-AGP were set up following a protocol described in Section 2.3.2. The products were visualised using TLC (see Section 2.3.5). The results displayed in **Figure 4.21-A** show that there is no difference in the enzymatic activity of β-galactosidase BT0461^{GH2} on N-glycans released from protein backbone by either PNGaseF or GH18 (BT1044) enzymes. The β-galactosidase BT0461^{GH2} is active on all forms of N-glycans. The PNGaseF and GH18-cleaved N-glycan structures are highlighted in red boxes. The examples of differences of these structures are displayed in **Figure 4.21-B**.

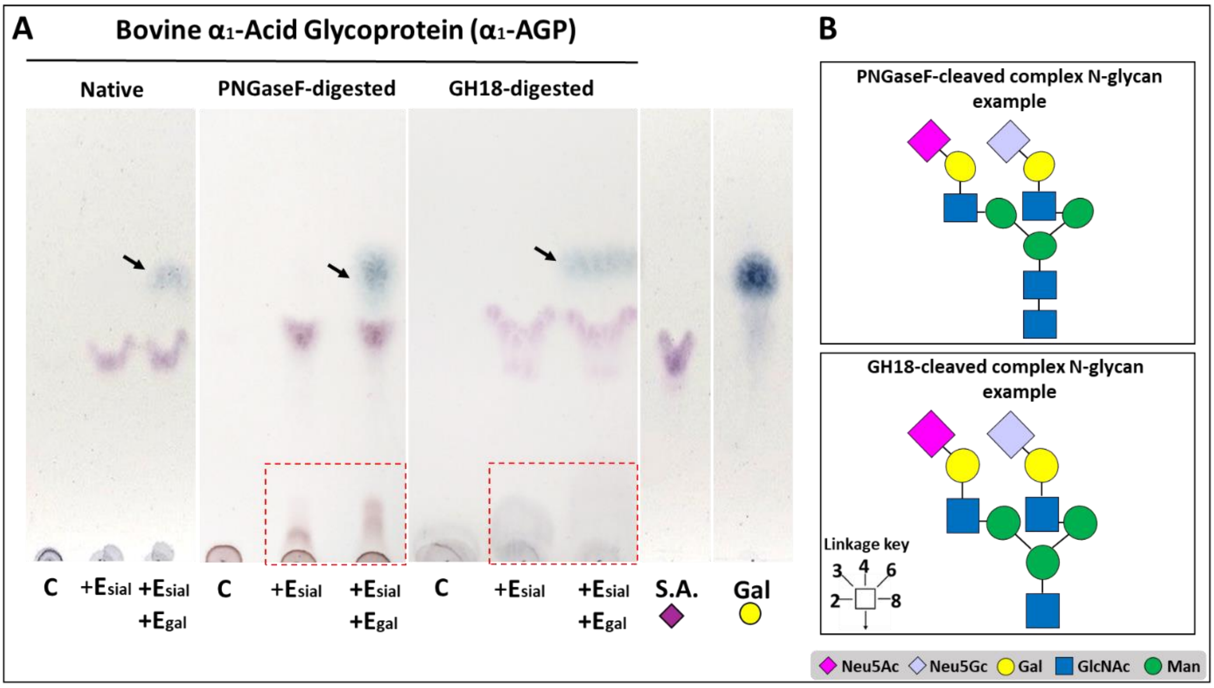
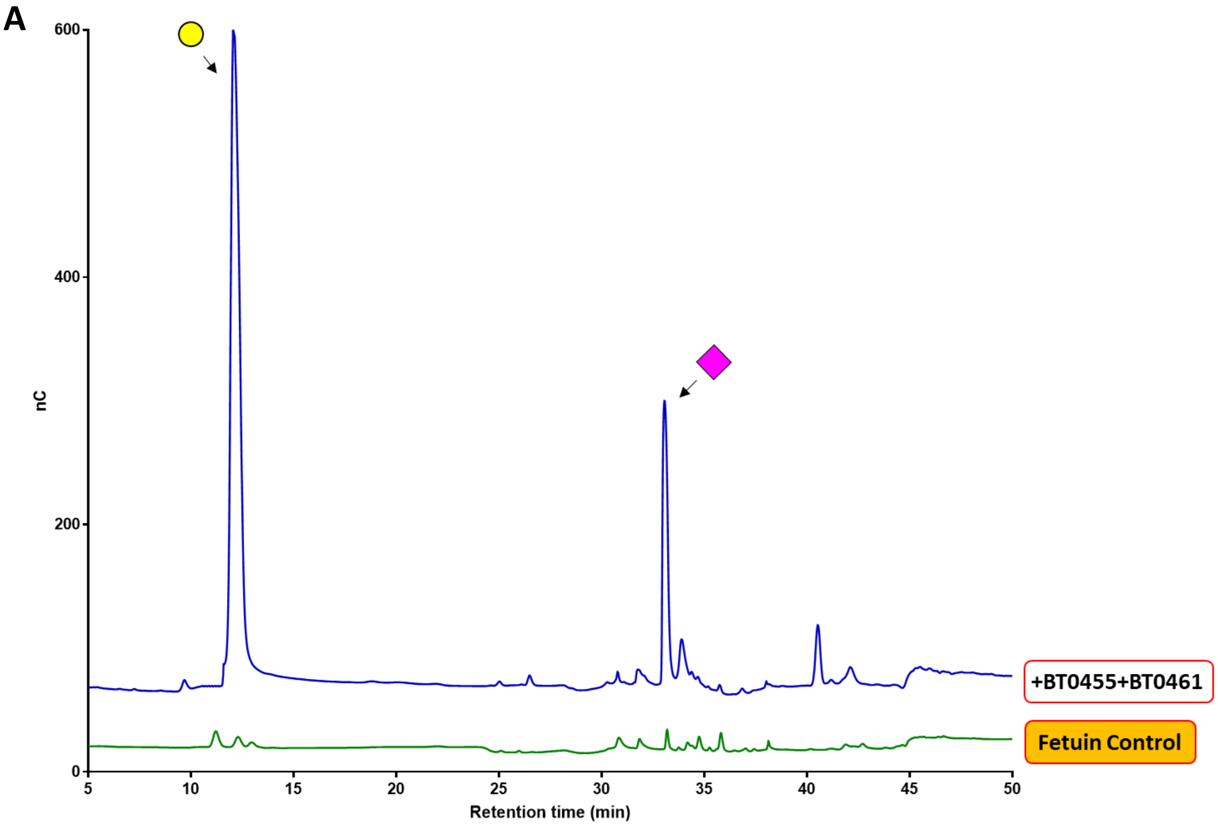


Figure 4.21: Degradation of desialylated bovine α$_1$-AGP N-glycan structures by β-galactosidase BT0461^{GH2}. A) Degradation products visualised on a TLC plate. E$_{sial}$ is BT0455^{GH33}; E$_{gal}$ is BT0461^{GH2}; C is substrate control. N-glycan structures are highlighted in red boxes. Thin-layer chromatography (TLC) plate was stained using the DPA stain. S.A. is 2mM sialic acid (Neu5Ac) standard; Gal is 1mM galactose standard. **B)** Examples of PNGaseF-cleaved and GH18-cleaved complex N-glycan structures are based on most abundant N-glycan structure found on α$_1$-AGP, identified using mass spectrometry data (see Figure 4.8).

To investigate the galactose release in more detail, the enzyme assays were set up and products were visualised using different optimised HPAEC-PAD programs (Section 2.3.3). As was previously observed, BT0461^{GH2} shows no activity on N-glycans that are capped by the sialic acids, thus the substrates used in this assay were de-sialylated using the sialidase BT0455^{GH33} (**Figure 4.20**). Based on the appearance of a peak identified using a standard as a galactose monosaccharide, the results show that β-galactosidase BT0461^{GH2} can cleave galactose off de-sialylated bovine fetuin and human transferrin N-glycoproteins (**Figure 4.22**). Combined, these results suggest that BT0461^{GH2} is an exo-glycosidase and N-glycoprotein de-glycosylation by BT0455^{GH33} and BT0461^{GH2} is a sequential process.



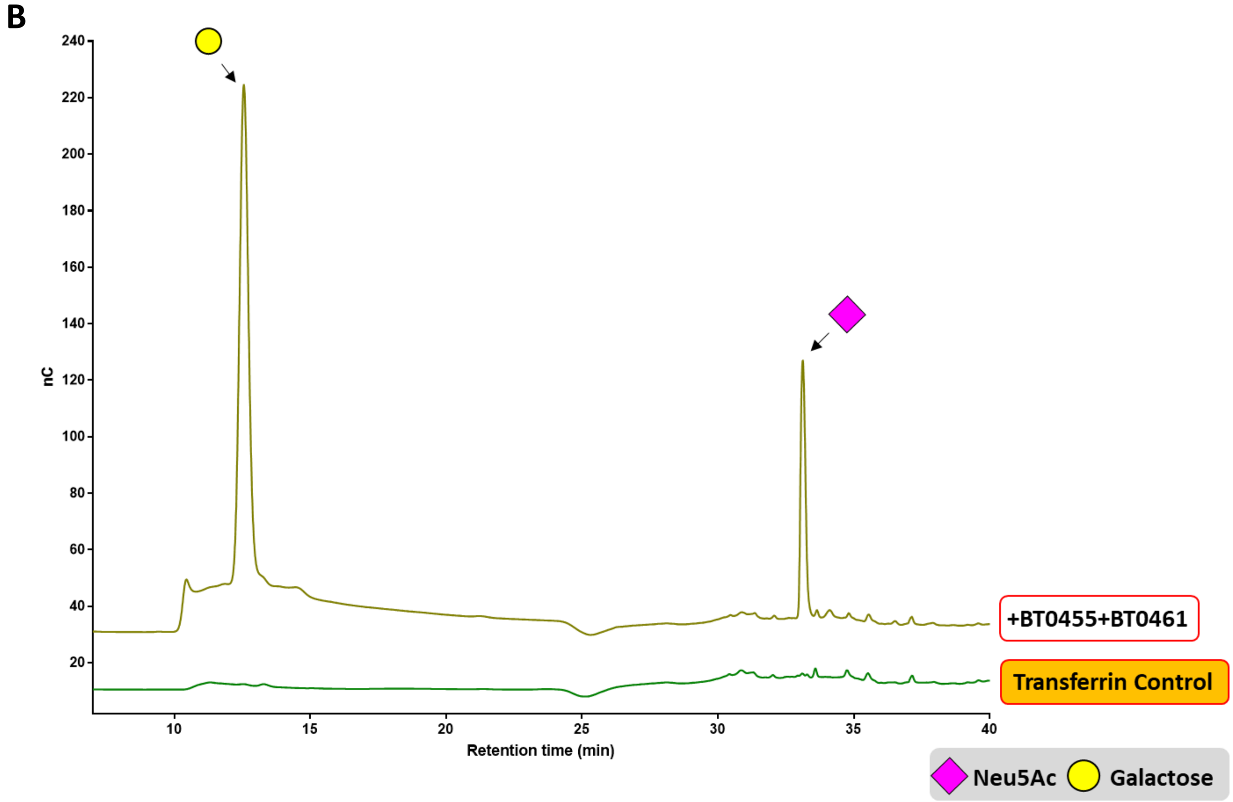


Figure 4.22: BT0461^{GH2} releases galactose from de-sialylated N-Glycoproteins. HPAEC-PAD chromatograms showing the release of galactose. **A.** Galactose release by BT0461^{GH2} from de-sialylated bovine Fetuin. **B.** Galactose release by BT0461^{GH2} from de-sialylated human Transferrin. 10mg/ml of each of N-glycoproteins was incubated with 1µM of BT0455^{GH33} and BT0461^{GH2} enzymes in 20mM sodium phosphate (pH 7) buffer. 1:10 dilution was loaded onto the HPAEC-PAD.

To investigate the de-glycosylation of N-glycan structures in more depth, the enzyme assay products were visualised using the U-HPLC combined with the mass spectrometry (LC-ESI-MS) analysis, which allows for accurate calculation of the mass of each peak and prediction of the N-glycan structure it corresponds to. The results displayed in **Figure 4.23** show that BT0461^{GH2} is an exo-glycosidase capable of taking majority of terminal β1,4-galactose caps off N-glycan structures pre-treated with endo-β-N-acetylglucosaminidase BT1044^{GH18} and the sialidase BT0455^{GH33}. It was observed that BT0461^{GH2} leaves some galactose caps in-tact on the complex bi-antennary N-glycan structure (**Figure 4.23-III**). It is important to note that these N-glycan structures were assembled based on the mass spectrometry data and thus, they may contain different galactose linkages that BT0461^{GH2} cannot target.

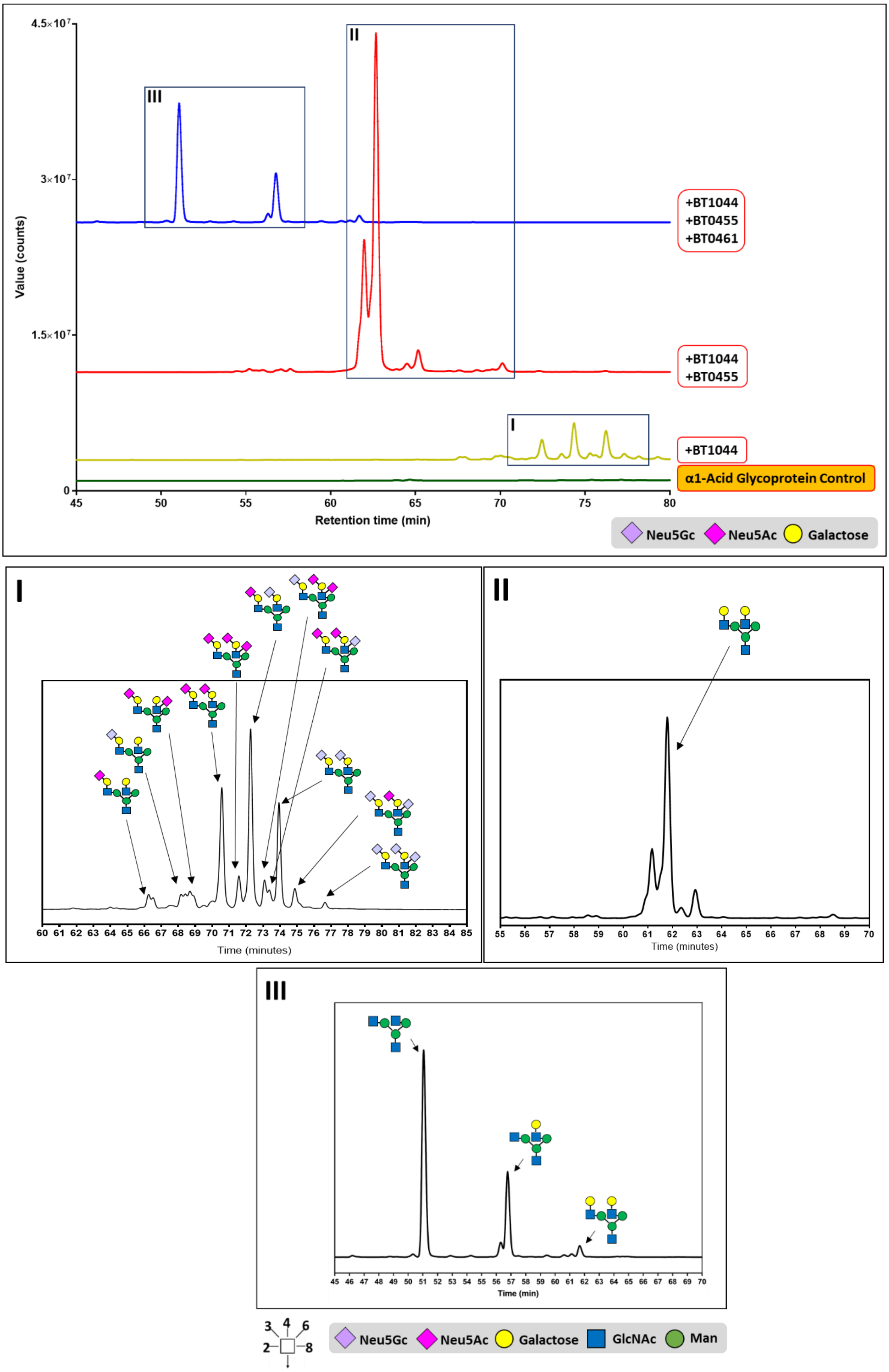


Figure 4.23: Degradation of bovine α_1-AGP N-glycan by BT0461^{GH2}.
Top panel: U-HPLC analyses of bovine α_1-AGP degradation by BT1044^{GH18}, BT0455^{GH33} and BT0461^{GH2}. Bottom panel: Figure I, II and III show N-glycan structures corresponding to each peak that were constructed using the mass spectrometry data (See Table 4.1).

4.3.4.4. Generating the BT0461^{GH2} catalytic mutant

One of the signature characteristics of the GH2-family β-galactosidases is the presence of the conserved catalytic residue – Glutamic acid (Glu; E). Bioinformatics analysis (see Section 4.3.4.2) and comparison of BT0461^{GH2} to the homologous GH2-family β-galactosidases from other *Bacteroides* species let us identify the catalytic residue of BT0461^{GH2} as Glu512 found in the sequence LGS**E**TAS (**Figure 4.24-A**). The catalytic mutant E512Q of BT0461^{GH2} was generated using site-direct mutagenesis following a protocol described in Section 2.1.10 (**Figure 4.24-B**). The primers used to generate the mutant are displayed in Supplemental table 3.2. Mutating negatively-charged glutamic acid (E) into a polar glutamine (Q) allowed us to retain the structural integrity of the protein but rendered it catalytically inactive.

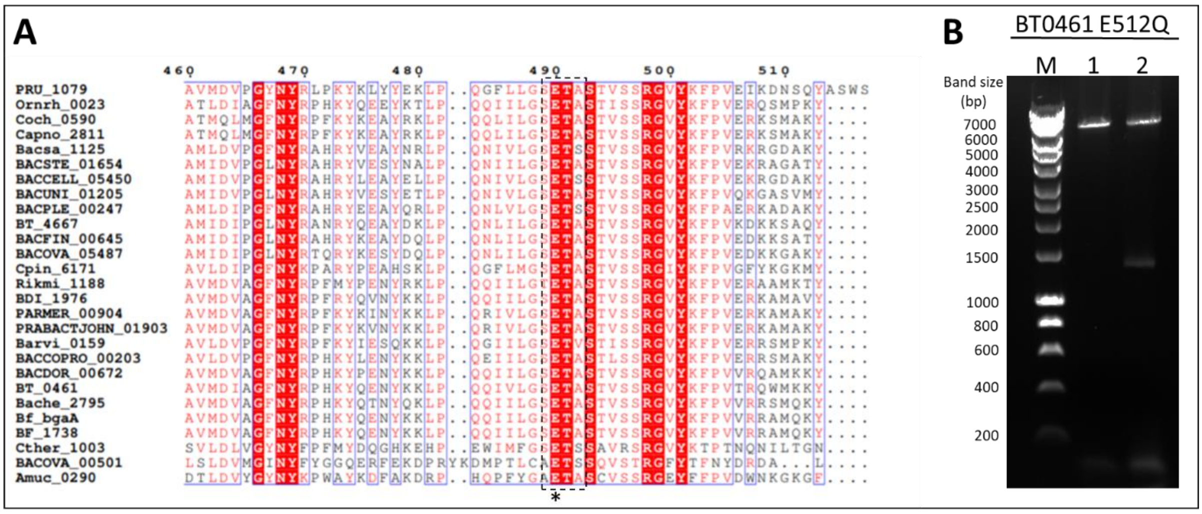


Figure 4.24: Generating a BT0461^{GH2} catalytic mutant. **A)** Amino acid sequence alignment of BT0461^{GH2} to the homologous GH2-family β-galactosidases. The sequence alignment was done using Clustal Omega and visualised using the Espript 3.0 too. Amino acid residues displayed in red boxes are fully conserved. Catalytic residue is highlighted in the black box and labelled with an asterisk (*). **B)** Agarose gel showing the PCR products of the site-directed mutagenesis. M is a high fidelity genomic DNA marker. Lane 1 shows a successful site-directed mutagenesis product that was sent to sequence.

The generated BT0461^{GH2} E512Q catalytic mutant was tested against a variety of β1,4-linked substrates and was found to be inactive (data not shown).

To understand the structural basis of the substrate specificity displayed by BT0461^{GH2}, the crystal trial of the BT0461^{GH2} E512Q catalytic mutant was set up in order to obtain a crystal of a ligand-bound form of the protein. The BT0461^{GH2} E512Q catalytic mutant was expressed and purified using the gel filtration following a protocol described in Section 2.1.25. The protein fractions collected from the gel filtration purification were visualised using the SDS-PAGE and combined (**Figure 4.25**).

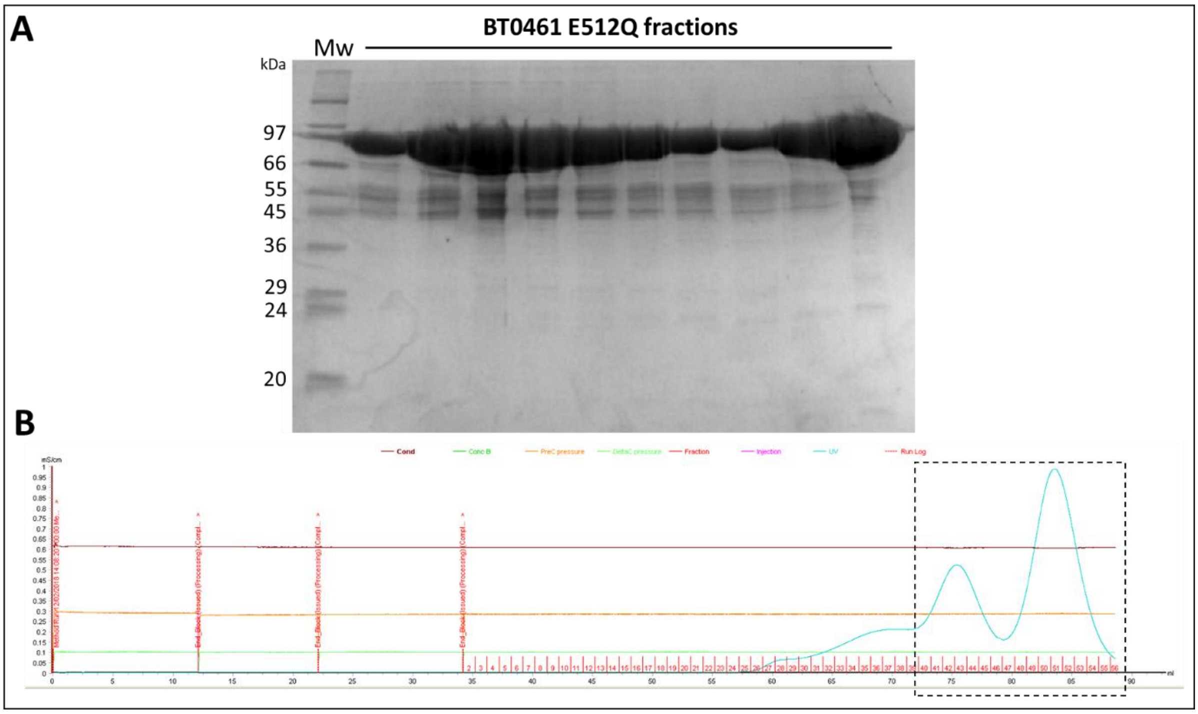


Figure 4.25: Purification of the BT0461^{GH2} E512Q catalytic mutant. A) SDS-PAGE gel of purified BT0461^{GH2} E512Q mutant. These fractions were pooled and concentrated. **B)** A graphical representation of the gel filtration procedure. Blue peaks highlighted in a black box represent the eluted protein fractions (displayed in red numbers) that were collected.

The purified protein was concentrated to 30mg/ml and co-crystallised with Gal-β1,4-GlcNAc using an automated Mosquito^{R} nano-dispensing robot and sitting-drop vapour diffusion method following a protocol described in Section 2.4.1. The conditions of crystal screens (Molecular Dimensions) in which crystals formed were optimised. The BT0461^{GH2} E512Q mutant was co-crystallised with Gal-β1,4-GlcNAc using the hanging-drop vapour diffusion method following a protocol described in Section 2.4.2. Numerous crystals were obtained throughout the duration of 2 months. The crystals that formed in 15% PEG3350, 50mM sodium nitrate, 0.1M Bis-Tris propane (pH 7.5) condition were harvested (**Figure 4.26**). The crystals were collected and analysed by Dr. Arnaud Baslé. Unfortunately, none of the crystals diffracted and due to the time constraints, the crystal screens could not be repeated.

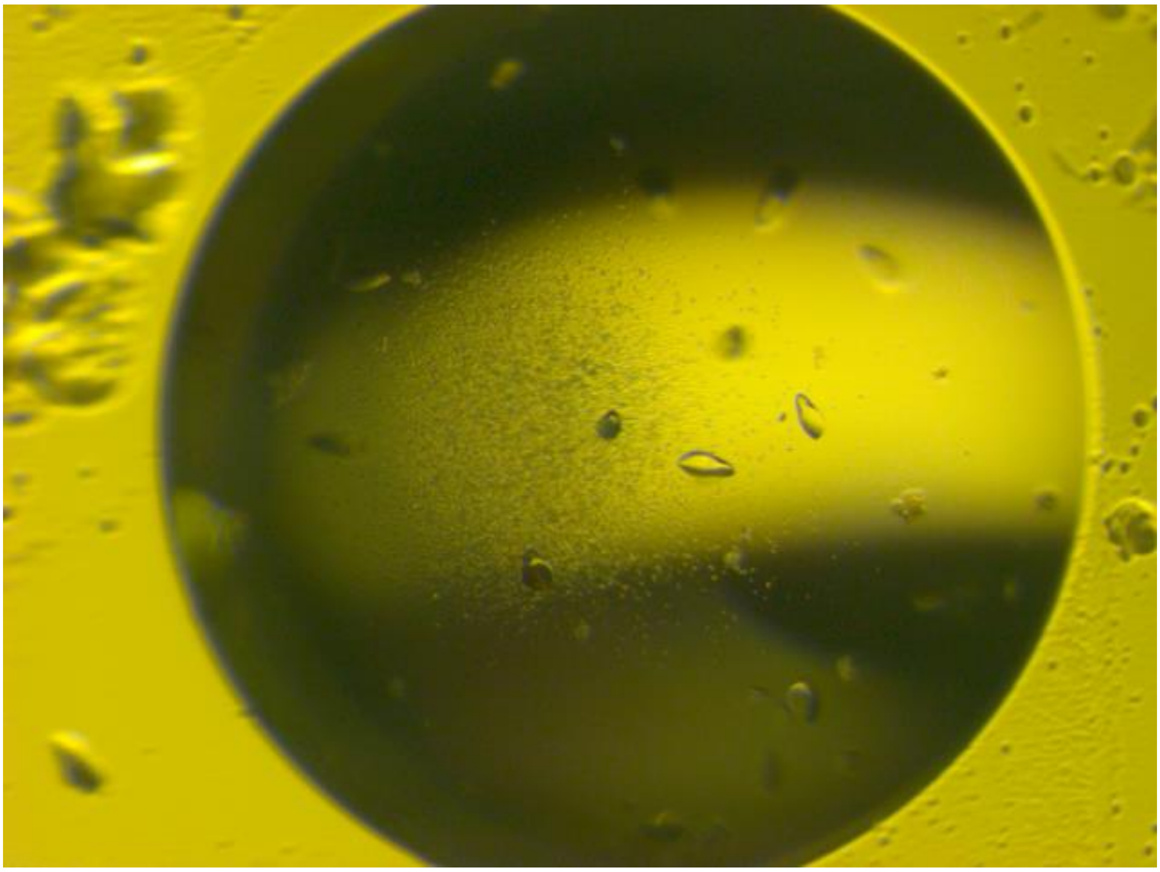


Figure 4.26: An image of the BT0461^{GH2} E512Q crystals. Crystals formed in 15% PEG3350, 50mM sodium nitrate, 0.1M Bis-Tris propane (pH 7.5) buffer using hanging-drop vapour diffusion method.

4.3.5. The four GH20s: Characterization and comparison of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}

4.3.5.1. Introduction

In the gastrointestinal tract (GI), D-GlcNAc and D-GalNAc residues are found in structural chains of carbohydrates decorating glycoconjugates, such as mucins, O-glycoproteins and N-glycoproteins. On N-glycoproteins, D-GlcNAc residues are usually found forming β1-2/4/6 and bisecting β1-4 glycosidic linkages (Stanley *et al.,* 2009). The enzymes that cleave GlcNAc and GalNAc monosaccharides from the non-reducing end of glycans are grouped into glycoside hydrolase (GH) family 20 (GH20) of the CAZy classification. *B. thetaiotaomicron*, a prominent member of human gut microbiota, encodes numerous GH20-family β-N-acetylhexosaminidases, each specialised in cleaving specific glycosidic bonds. Four of these enzymes - BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, were shown to be upregulated when *B. thetaiotaomicron* was grown on N-glycoproteins as sole carbon sources, compared to the glucose control (**Figure 3.7.1**). Despite the potential importance of N-glycan degradation to the host health (e.g. degradation of N-glycans present on secretory IgA), the knowledge of de-glycosylation pathways is lacking, particularly how commensal gut microbes can tackle different D-GlcNAc linkages found on complex N-glycan structures. Thus, the aim of this section was to characterize the activities of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, and investigate their involvement in complex N-glycan degradation.

4.3.5.2. Bioinformatics analysis and comparison of the four GH20s

To compare the sequence similarities of the four GH20s to the other known N-acetyl-β-hexosaminidases, the amino acid sequences of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} were obtained using the UniProt database and queried against BLASTP and KEGG search to identify the homologous proteins. The identified putative β-hexosaminidase homologs were used for phylogenetic and bioinformatics analysis.

Initially, the sequence similarity database (SSDB) tool hosted by Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to identify orthologues of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. The orthologues were found in a range of species of bacteria, including members of gut microbiota and other *Bacteroides* species. Closest related orthologues were selected for phylogenetic analysis. Based on the CAZy search, none of these enzymes have been characterised yet with an exception of BF_1730 (nanA), BF_1734 (nanB) and BF_1735 (nanM) that were shown to be involved in sialylated glycoconjugate utilisation in *B. fragilis* (Nakayama-Imaohji *et al.,* 2012).

In order to explore the sequence similarities between the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, a phylogenetic tree was constructed using a phylogeny.fr tool. The taxonomic relationships between the enzymes displayed in the phylogenetic tree in **Figure 4.27** indicate that the four GH20s encoded by the *B. thetaiotaomicron* could be functionally quite different. The enzymes in each cluster could be specialised in targeting different glycosidic linkages, e.g. ones found on complex N-glycans such as β1,2/4/6 or bisecting GlcNAc-β1,4- linkages.

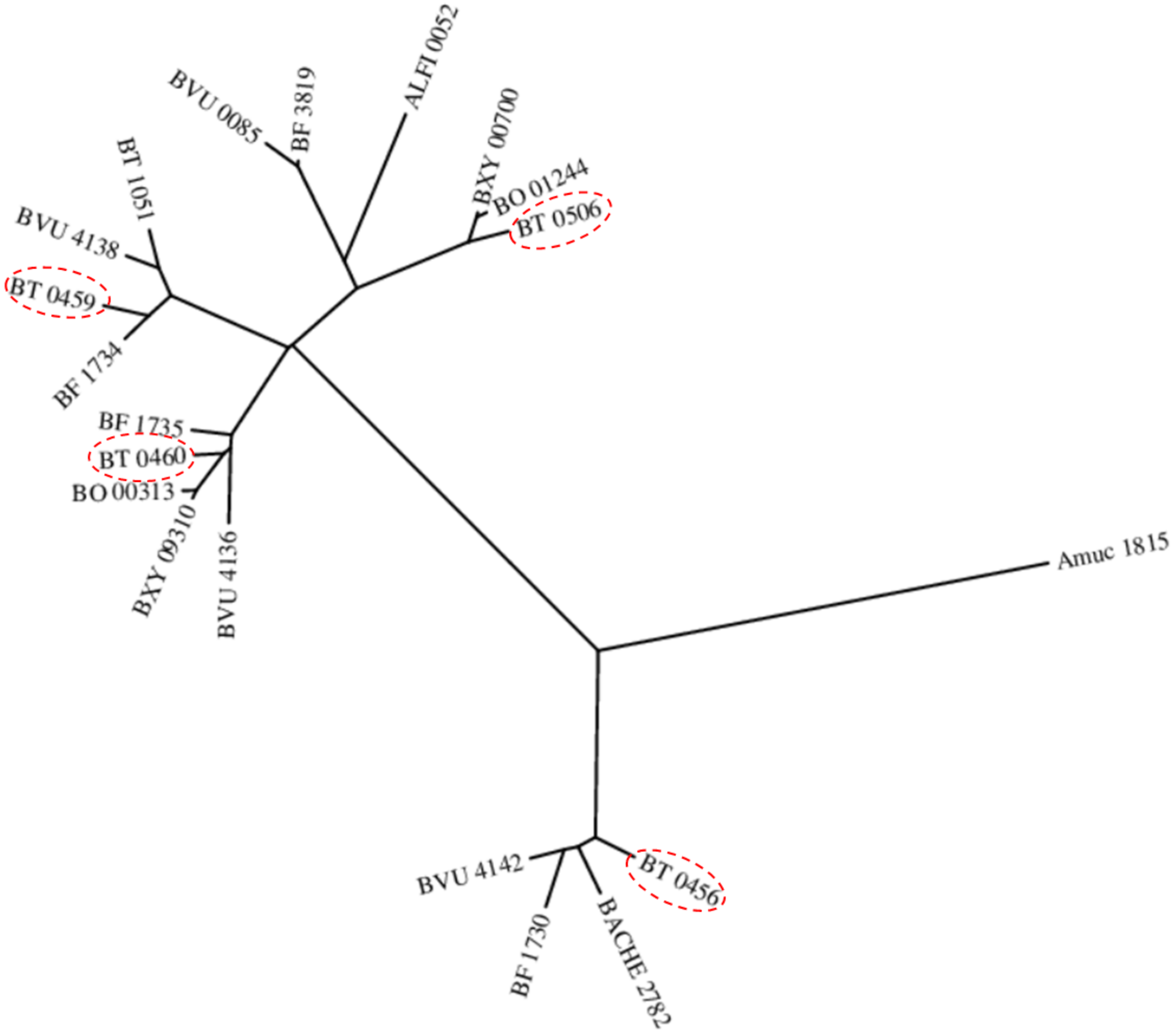


Figure 4.27: Phylogenetic tree of β-hexosaminidases encoded by the members of gut microbiota. Species included in the tree: *B. thetaiotaomicron, B. fragilis, B. vulgatus, A. muciniphila, B. xylanisolvens, B. ovatus, B. hellcogenes and A. finegoldii*. BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} are highlighted in red. The phylogenetic tree was constructed using the DynTree layout.

To investigate the taxonomic relationship of the four GH20s further, the BT0460^{GH20} amino acid sequence was used as a query on a KEGG database to identify paralogous genes in *B. thetaiotaomicron*. The taxonomic relationship of 15 identified genes was visualised using a dendrogram (**Figure 4.28**). The results confirmed that the four GH20s are not functionally related and suggest they may have different substrate or linkage specificities.

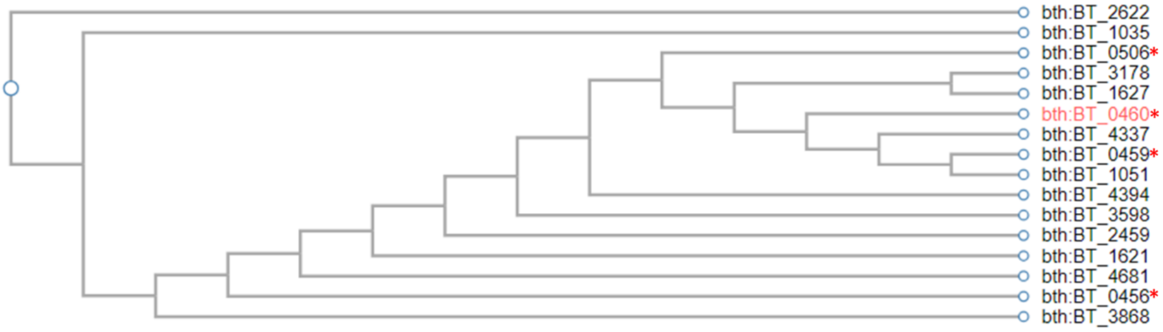


Figure 4.28: A dendrogram showing the relationship of BT0460^{GH20} β-hexosaminidase to paralogues encoded by *B. thetaiotaomicron.* The four GH20s are labelled with a red asterisk (*). KEGG SSDB tool was used to perform a sequence similarity search. The dendrogram was generated using the best matches identified by a Smith-Waterman similarity score.

To explore the similarities of the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases further, the amino acid sequences were aligned and compared. The results show that amino acid sequence of BT0459^{GH20} is 40% identical to BT0460^{GH20}, 36% identical to BT0506^{GH20} and 22% identical to BT0456^{GH20} whereas BT0460^{GH20} is 33% identical to BT0506^{GH20} and 22% identical to BT0456^{GH20} (**Figure 4.29**). Meanwhile, the amino acid sequence of BT0456^{GH20} is 24% identical to BT0506^{GH20}. The sequence alignment shows that the four GH20s have the characteristic catalytic residues of the GH20 retaining family of β-hexosaminidases, the D-E pair of amino acids that is usually preceded by the H-x-G-G motif (CAZypedia.org, GH20 family). Interestingly, BT0456^{GH20} has a H-I-G-T motif instead which may be involved in its substrate specificity.

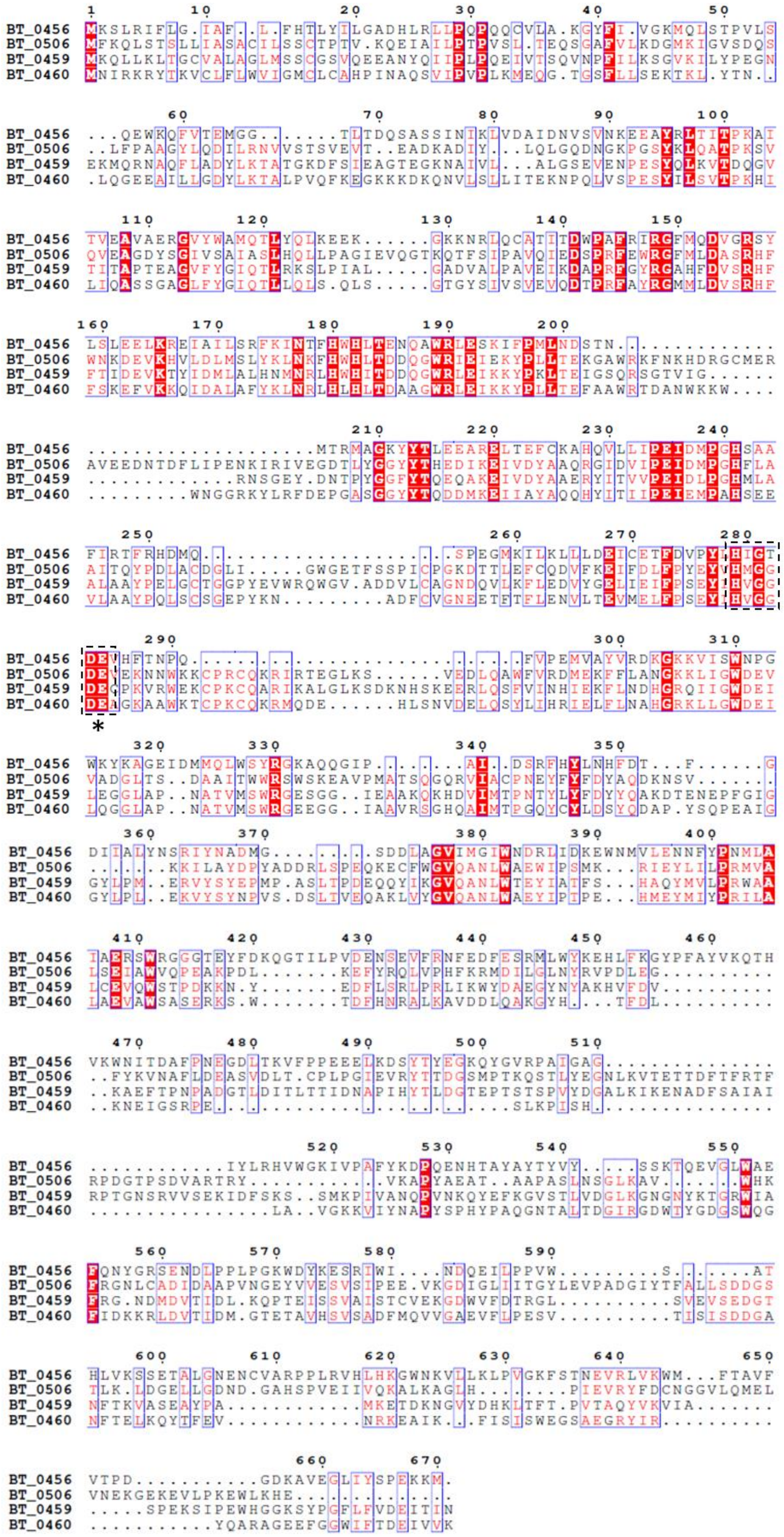


Figure 4.29: Amino acid sequence alignments of the four β-hexosaminidases: BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. Fully conserved amino acid residues are highlighted in solid red boxes. Amino acid residues framed in blue are 70% or more similar based on their physicochemical properties. Catalytic residue pair D-E is labelled with an asterisk (*).

The PFAM domain composition of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0460^{GH20} β-hexosaminidases is displayed in **Figure 4.30**. Despite having quite low sequence similarity scores displayed in **Figure 4.29**, the four β-hexosaminidases have very similar PFAM domain architectures. It is not surprising, considering multiple protein sequences can belong to a single PFAM domain family. All four of the GH20-family enzymes contain the non-catalytic Glyco_hydro_20b and catalytic Glyco_hydro_20 domains. Interestingly, BT0456^{GH20} appears to have a truncated version of a Glyco_hydro_20 domain, which may play a role in its activity. BT0459^{GH20} and BT0460^{GH20} also contain a predicted CBM domain labelled F5_F8_type_C (Pfam) whereas BT0506^{GH20} contains a predicted CBM domain labelled PA14 (Pfam). BT0459^{GH20} and BT0506^{GH20} also contain a predicted linker domain labelled as CHB_Hex_C_1 by Pfam.

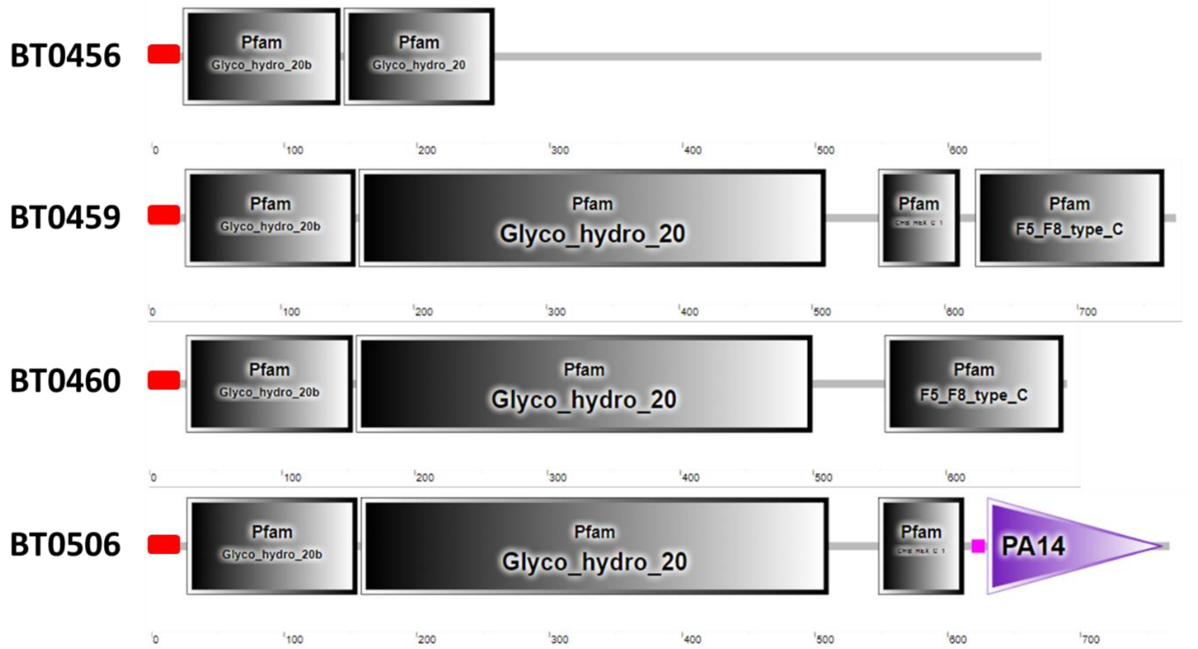
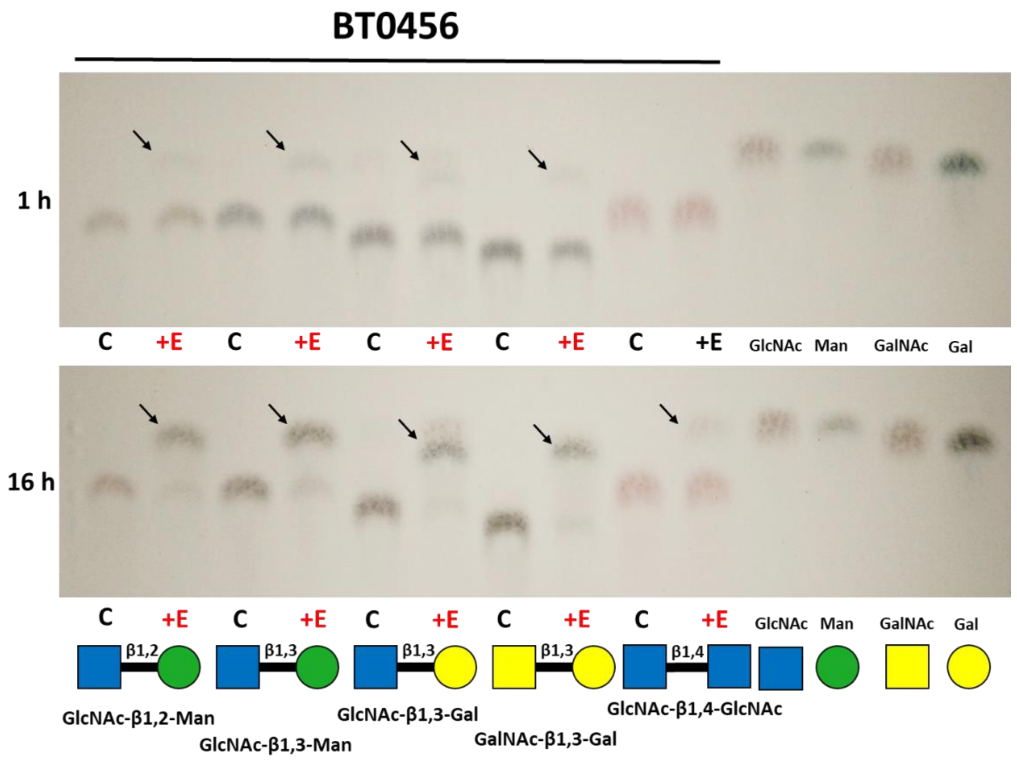


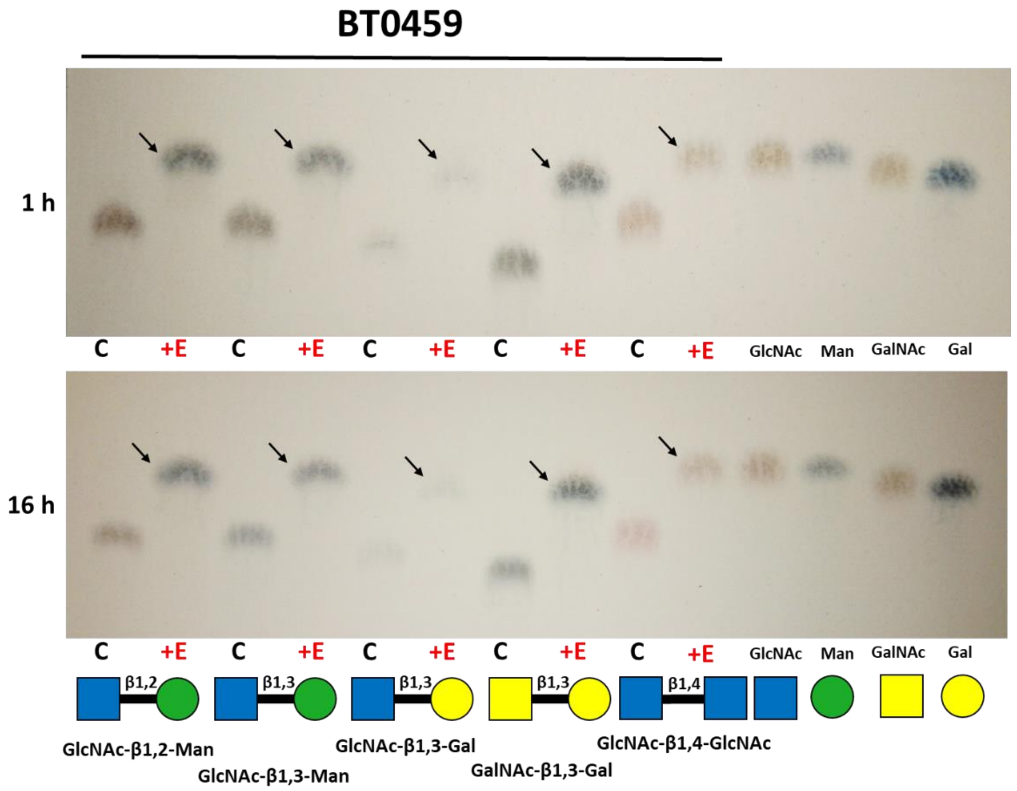
Figure 4.30: Domain composition of the GH20-family β-hexosaminidases. PFAM domain composition of β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} from *B. thetaiotaomicron* are displayed in grey boxes. Predicted signal peptide sequence is displayed in red. Amino acid sequence length is displayed below the domains.

4.3.5.3. Defining the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} substrate specificity

To initially investigate the activities of the four GH20 enzymes, assays with para-Nitrophenol (pNP)-linked substrates were performed. Data showed that BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} were all active against pNP-β-GlcNAc and pNP-β-GalNAc, confirming they are β-hexosaminidases. To gain more insight into the substrate specificities of the four GH20s further, enzyme assays against various defined disaccharides were performed and visualised using thin-layer chromatography (TLC) (see Section 2.3.2). The tested disaccharides include GlcNAc-β1,2-Man, GlcNAc-β1,3-Man, GlcNAc-β1,3-Gal, GalNAc-β1,3-Gal and GlcNAc-β1,4-GlcNAc.

Based on the observed release of monosaccharides, the initial results showed that all four β-hexosaminidases are active against most of the disaccharides tested but displayed a variation in their enzymatic efficiencies (**Figure 4.31**). BT0459^{GH20} appeared to be the most active enzyme of the four, completely degrading all tested disaccharides into monosaccharides within 1 hour of incubation at 37°C. Meanwhile, BT0456^{GH20} and BT0460^{GH20} displayed only a partial degradation of the disaccharides after 1 hour of incubation whereas BT0506^{GH20} appears to be the least efficient enzyme, showing no activity against any of the disaccharides after 1h of incubation. Furthermore, only partial degradation of the tested disaccharides was observed by BT0506 after 16 hours of incubation. BT0506^{GH20} showed no activity against GlcNAc-β1,4-GlcNAc (chitobiose). These results suggest an importance of a +2 subsite for this enzyme. In overall, these results indicate that BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases have different glycosidic linkage specificities. Based on the inefficiency observed in GlcNAc-β1,4-GlcNAc (chitobiose) degradation by BT0456^{GH20}, BT0460^{GH20} and BT0506^{GH20}, these enzymes may have a preference for a different aldohexaose, such as mannose, over D-GlcNAc in the +1 subsite.


BT0456
1 h
C
+E
C
+E
C
+E
C
+E
C
+E
GlcNAc
Man
GalNAc
Gal
16 h
C
+E
C
+E
C
+E
C
+E
C
+E
GlcNAc
Man
GalNAc
Gal
β1,2
β1,3
β1,3
β1,3
β1,4
GlcNAc-β1,2-Man
GlcNAc-β1,3-Man
GlcNAc-β1,3-Gal
GalNAc-β1,3-Gal
GlcNAc-β1,4-GlcNAc


BT0459
1 h
C
+E
C
+E
C
+E
C
+E
C
+E
GlcNAc
Man
GalNAc
Gal
16 h
C
+E
C
+E
C
+E
C
+E
C
+E
GlcNAc
Man
GalNAc
Gal
β1,2
β1,3
β1,3
β1,3
β1,4
GlcNAc-β1,2-Man
GlcNAc-β1,3-Man
GlcNAc-β1,3-Gal
GalNAc-β1,3-Gal
GlcNAc-β1,4-GlcNAc

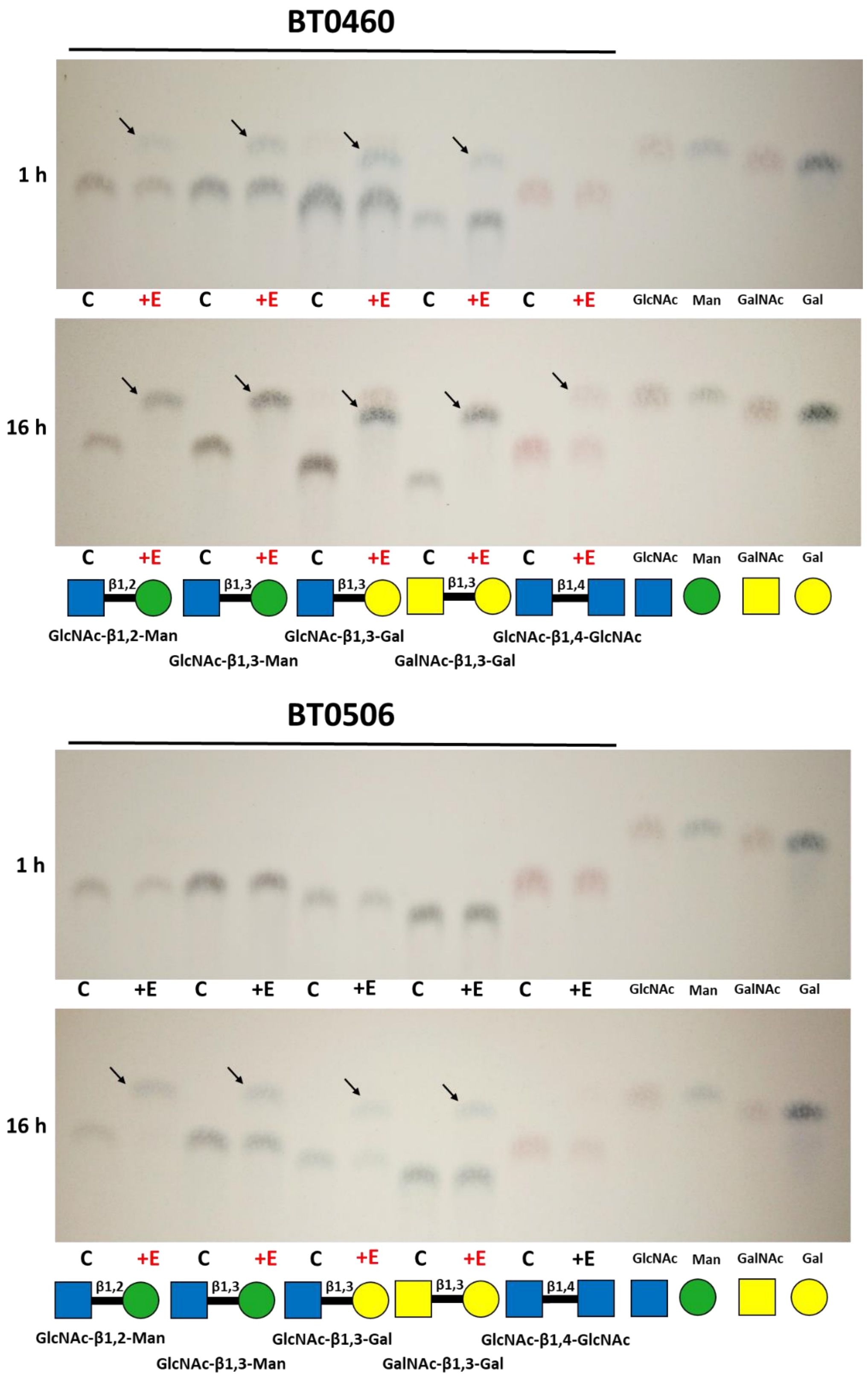


Figure 4.31: Screening for the enzyme activity of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} against simple disaccharides. 1µM of enzymes was incubated with 1mM of disaccharides. C – disaccharide control; +E – substrate incubated with the enzyme. Red highlight indicates an active enzyme. 1mM of N-Acetyl-Glucosamine (GlcNAc), mannose (Man), N-acetyl-galactosamine (GalNAc) and galactose (Gal) standards were used. TLC plates were stained using the DPA stain. Monosaccharide products are labelled with arrows.

Interestingly, assays of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} against chito-oligosaccharides revealed that BT0459^{GH20} appears to be the only enzyme that can fully degrade chitobiose, chitotriose, chitotetrose and chitopentose during 1 hour of incubation at 37°C (**Figure 4.32**). Due to a presence of a band representative of a chitobiose, it was observed that neither BT0456^{GH20}, BT0460^{GH20} nor BT0506^{GH20} showed any activity against this disaccharide after 1 hour of incubation. Based on these results, all four GH20s seem to prefer chitotriose, chitotetrose and chitopentose, suggesting they may have a preference for oligosaccharide structures over disaccharides. Furthermore, none of the enzymes showed any activity on de-acetylated chito-oligosaccharides, indicating the importance of the acetyl group for their activity (data not shown).

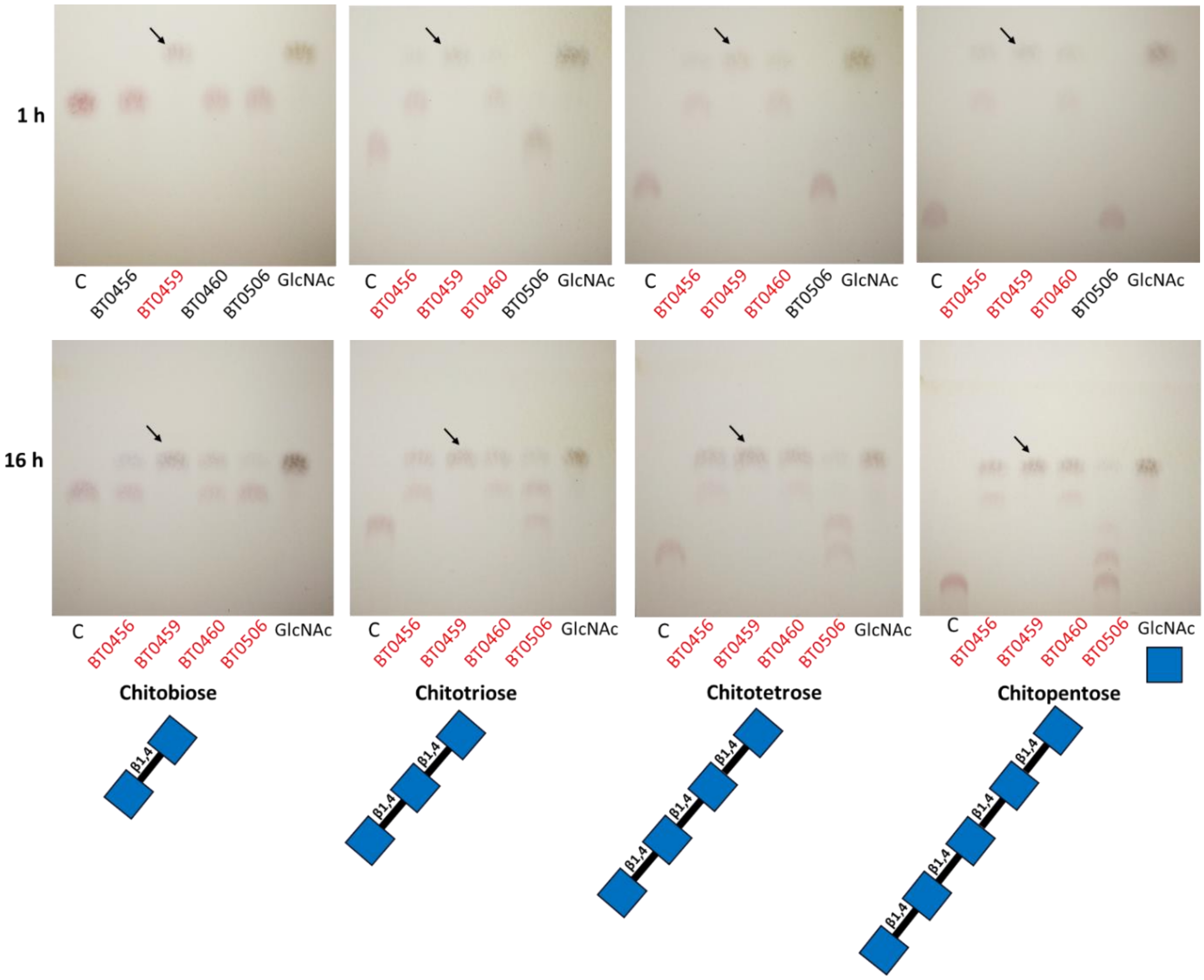


Figure 4.32: Investigating the enzyme activities of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} against chito-oligosaccharides. 1µM of enzymes was incubated with 1mM of substrates. C – substrate control. Red highlight indicates enzyme was active. 1mM of N-Acetyl-Glucosamine (GlcNAc) standard was used. TLC plates were stained with DPA stain. GlcNAc products are labelled with arrows.

To investigate whether BT0459^{GH20}, BT0456^{GH20}, BT0460^{GH20} and BT0506^{GH20} have the ability to cleave the Gal-β1,4-GlcNAc (LacNAc) disaccharide off, which would be indicative of an endo-glycosidase activity, enzyme assays against Lacto-N-neotetraose were set up following a protocol described in Section 2.3.2. The enzyme assays were incubated for 16 hours at 37 °C. The results displayed in **Figure 4.33** show that none of the enzymes can hydrolyse LacNAc linkages, suggesting they are exo-glycosidases.

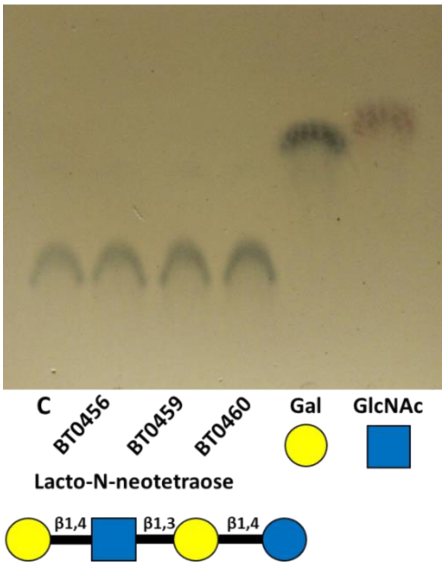
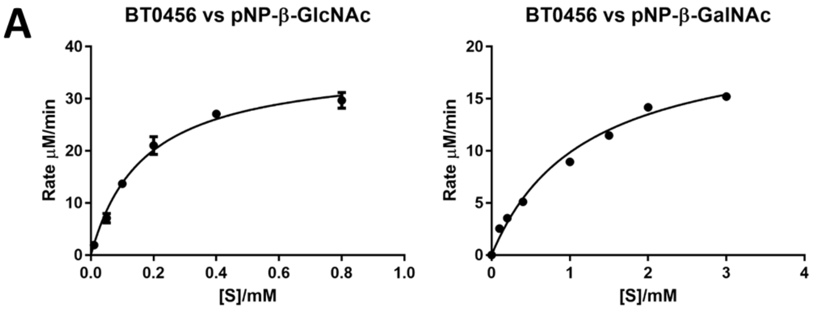
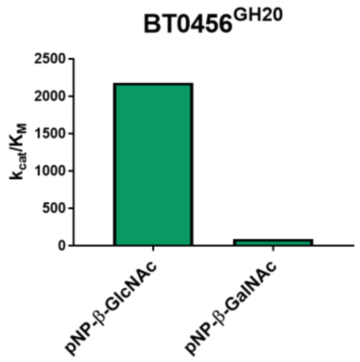


Figure 4.33: Screening for the activity of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} against Lacto-N-neotetraose. 1μM of enzymes was incubated with 1mM of substrate. C –substrate control. 1mM of galactose (Gal) and N-Acetyl-Glucosamine (GlcNAc) standards were used. TLC plates were stained using the DPA stain.

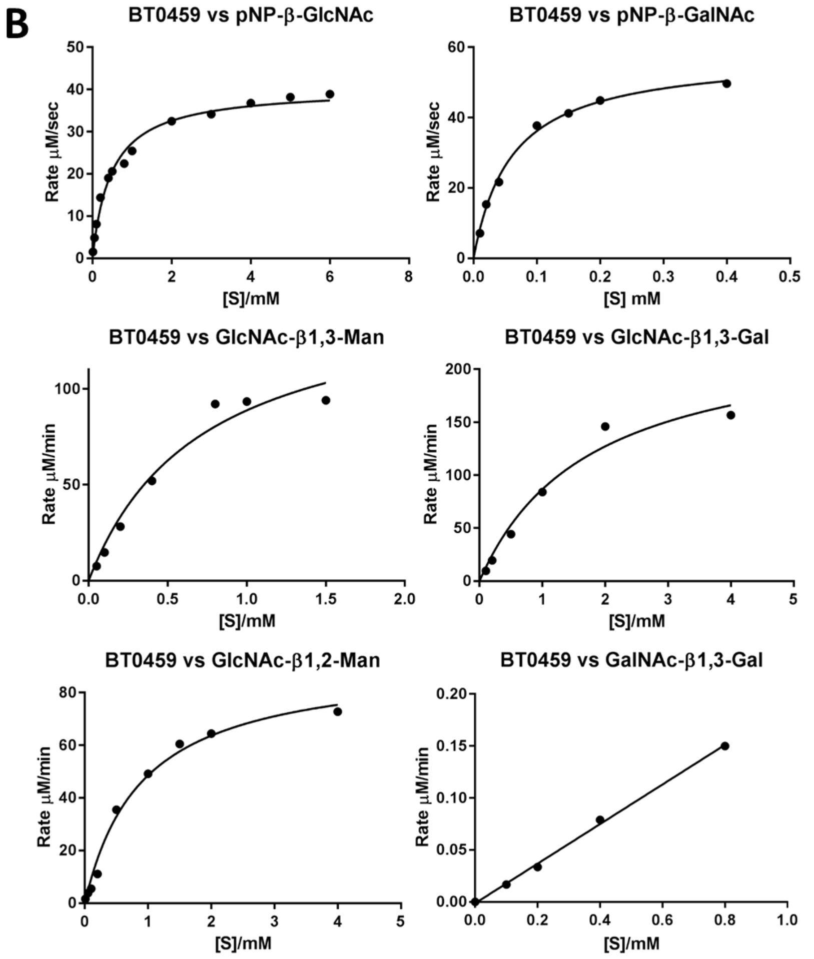
To determine the substrate preference of the four retaining β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, the catalytic activities of the enzymes against defined disaccharides were measured at 340nm using a D-Mannose/D-Galactose detection kit (Megazyme) following the manufacturer's instructions (see Section 2.3.5.2). The enzyme kinetics results showed that all four glycosidases displayed the highest activity against the pNP-β-GlcNAc and pNP-β-GalNAc, however these are synthetic substrates that are not found in nature (**Figure 4.34**). Based on the k_{cat}/K_M, BT0460^{GH20} and BT0506^{GH20} show a preference for GlcNAc-β1,2-Man disaccharide over the rest of the substrates tested. This disaccharide is commonly found in complex N-glycan structures.

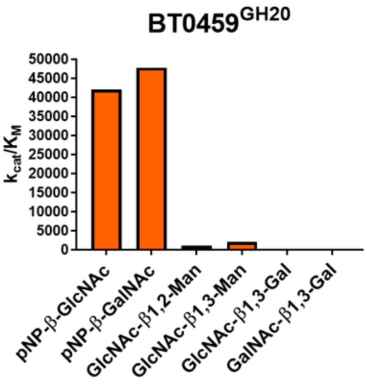
Interestingly, BT0459^{GH20} shows a 2-fold preference for a synthetic GlcNAc-β1,3-Man disaccharide over GlcNAc-β1,2-Man.



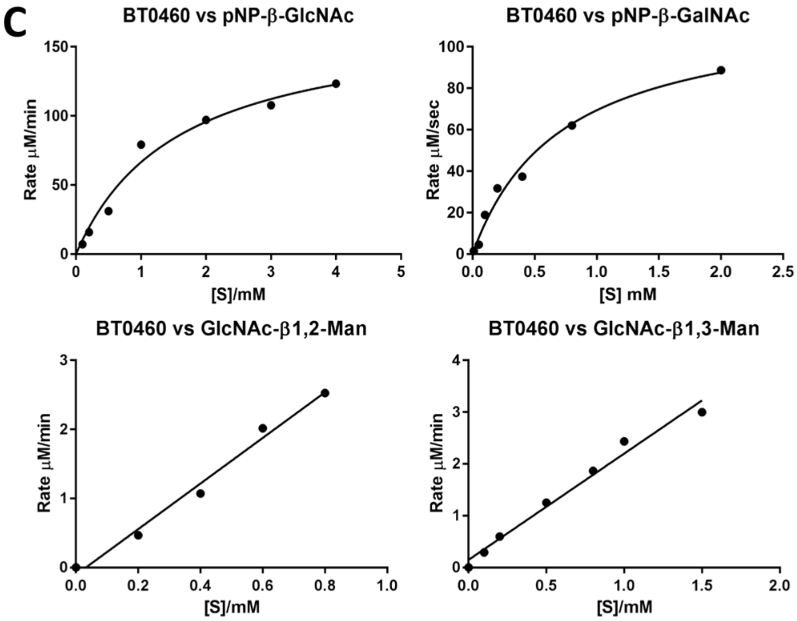


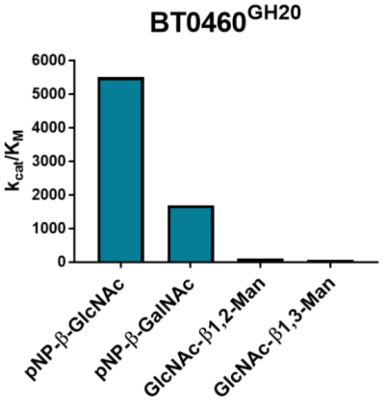
	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0456	pNP-β-GlcNAc	382	0.175±0.013	2183
	pNP-β-GalNAc	105	1.145±0.17	92
	GlcNAc-β1,2-Man	-	-	-
	GlcNAc-β1,3-Gal	-	-	-
	GlcNAc-β1,3-Man	-	-	-



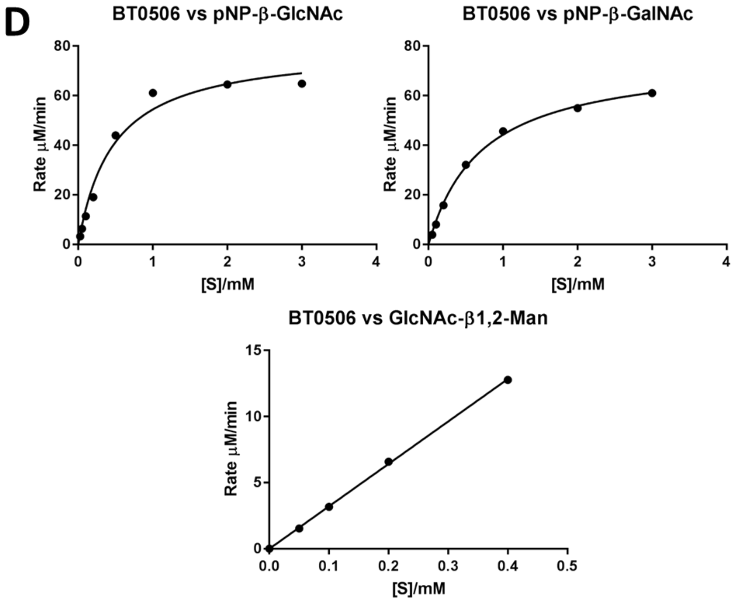


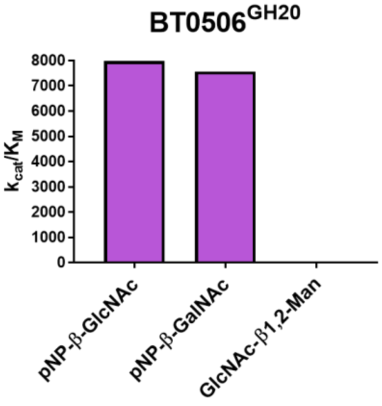
	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0459	pNP-β-GlcNAc	20185	0.48±0.03	42052
	pNP-β-GalNAc	2868	0.06±0.003	47800
	GlcNAc-β1,2-Man	888	0.78±0.053	1138
	GlcNAc-β1,3-Gal	597	1.76±0.29	339
	GlcNAc-β1,3-Man	1523	0.72±0.127	2115
	GalNAc-β1,3-Gal	-	-	101





	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0460	pNP-β-GlcNAc	8690	1.581±0.23	5496
	pNP-β-GalNAc	1218	0.72±0.07	1692
	GlcNAc-β1,2-Man	-		103
	GlcNAc-β1,3-Gal	-	-	-
	GlcNAc-β1,3-Man	-		60





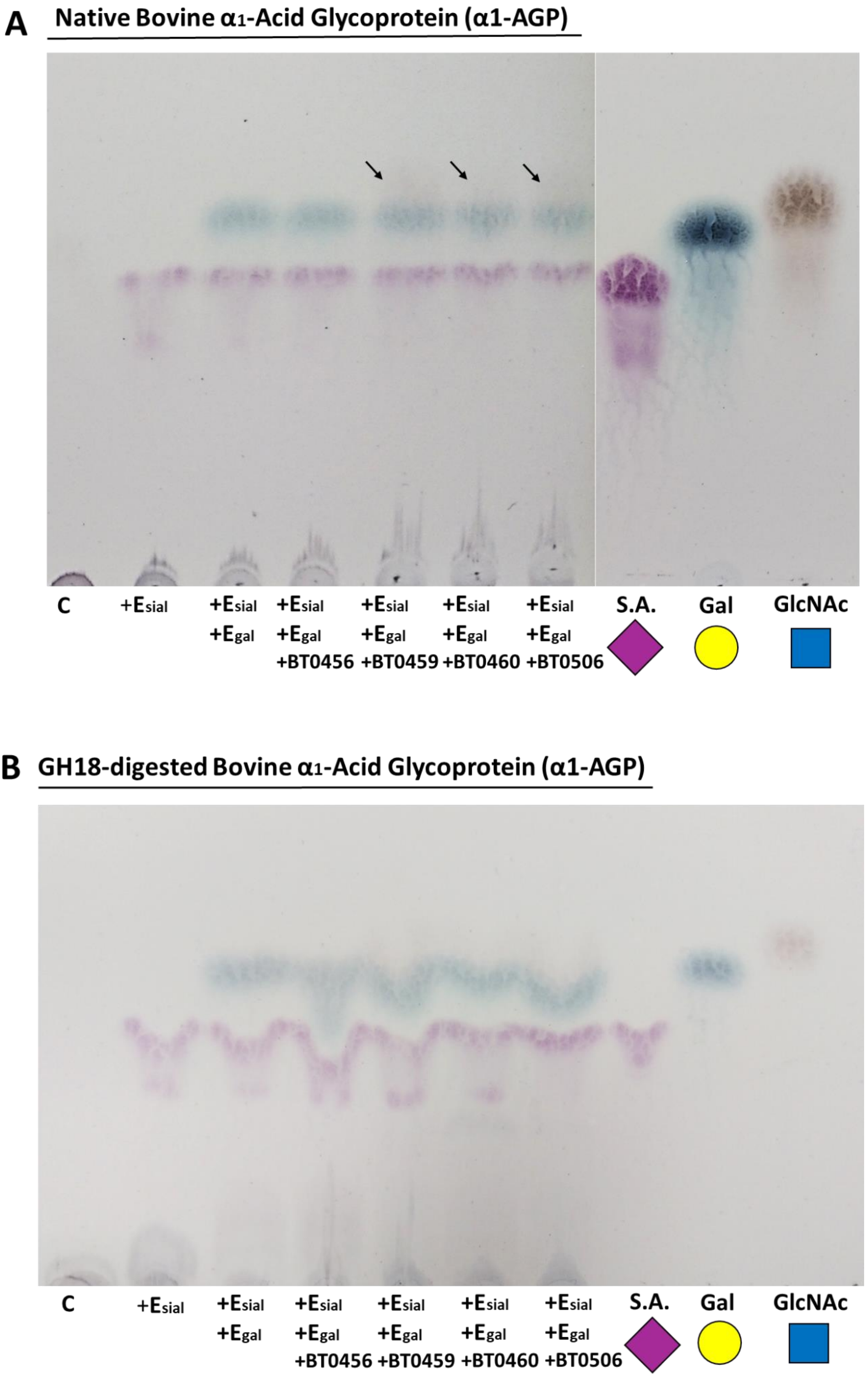
	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0506	pNP-β-GlcNAc	3989	0.47±0.05	8487
	pNP-β-GalNAc	378	0.7±0.029	540
	GlcNAc-β1,2-Man	-	-	2.55
	GlcNAc-β1,3-Gal	-	-	-
	GlcNAc-β1,3-Man	-	-	-

Figure 4.34: Graphical representation of the enzyme activity of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. A) BT0456^{GH20} kinetic parameters. **B)** BT0459^{GH20} kinetic parameters. **C)** BT0460^{GH20} kinetic parameters. **D)** BT0506^{GH20} kinetic parameters. All reactions were performed at 37 °C in 20 mM NaH_2PO_4 (pH 7.5). Rates were measured at 320 nm for pNP-linked substrates and at 340 nm for the disaccharides. The results were plotted in GraphPad Prism 7.0 software using non-regression analysis (see Section 2.3.5). The standard errors were generated from biological triplicates. For some substrates, labelled “–“ on the table, no kinetic parameters could be obtained.

4.3.5.4. Investigating the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} activities against N-glycans

To further explore the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} substrate specificities, enzyme assays were set up against complex N-glycoproteins, such as bovine α_1-AGP and bovine fetuin. These N-glycoproteins were chosen because they contain different complex N-glycan structures (**Figure 3.6**).

Speculating that the GH18-liberated N-glycan structures are the likely targets for the GH20-family enzymes because *B. thetaiotaomicron* encodes numerous GH18 glycosidases but does not produce a PNGaseF-like enzyme, enzyme assays against native α_1-AGP, GH18-digested α_1-AGP and PNGAseF-digested bovine fetuin were set up following a protocol described in Section 2.3.2. The products were visualised using the TLC (**Figure 4.35**) (see Section 2.3.5). D-GlcNAc migrates very closely to the galactose on the TLC plate, thus a very the faint band is visible after the staining with the DPA (labelled with black arrows). Despite problems with the staining, the results show that all four GH20-family enzymes appear to be active on all N-glycoproteins tested. Compared to BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, BT0456^{GH20} appears to release the least of D-GlcNAc off the native α_1-AGP (**Figure 4.35-A**).


A
Native Bovine α1-Acid Glycoprotein (α1-AGP)
C
+Esial
+Esial +Egal
+Esial +Egal +BT0456
+Esial +Egal +BT0459
+Esial +Egal +BT0460
+Esial +Egal +BT0506
S.A.
Gal
GlcNAc
B
GH18-digested Bovine α1-Acid Glycoprotein (α1-AGP)
C
+Esial
+Esial +Egal
+Esial +Egal +BT0456
+Esial +Egal +BT0459
+Esial +Egal +BT0460
+Esial +Egal +BT0506
S.A.
Gal
GlcNAc

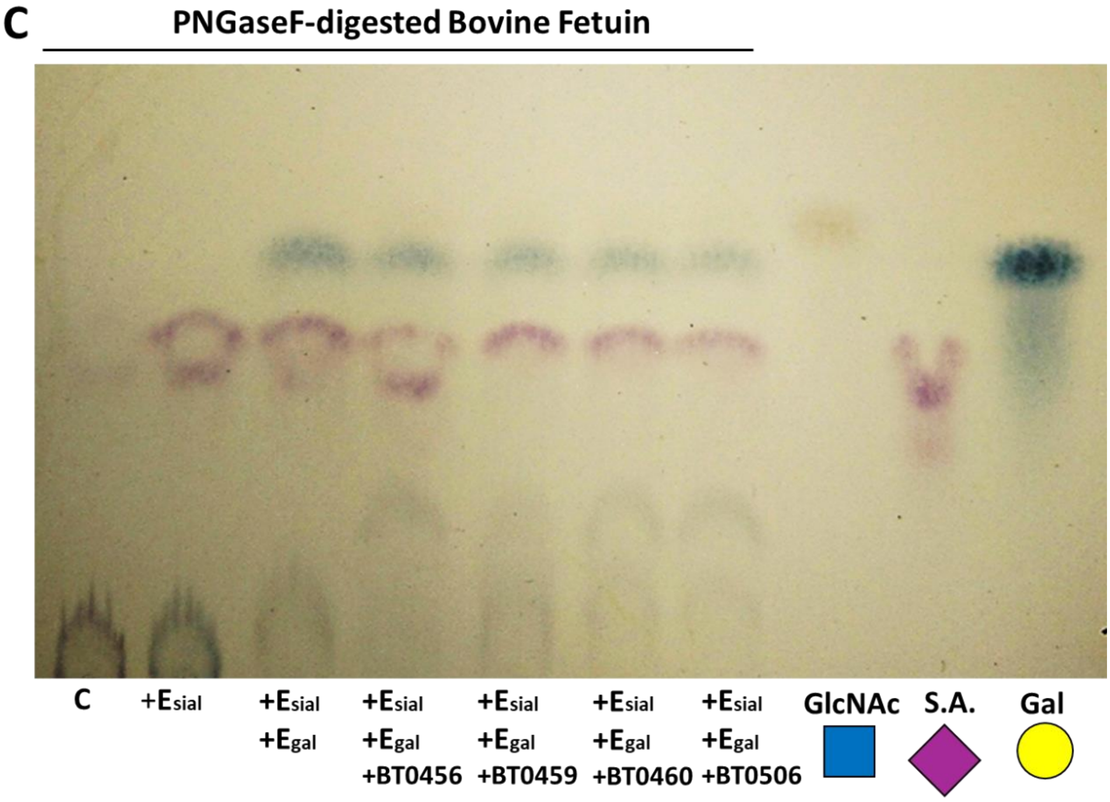
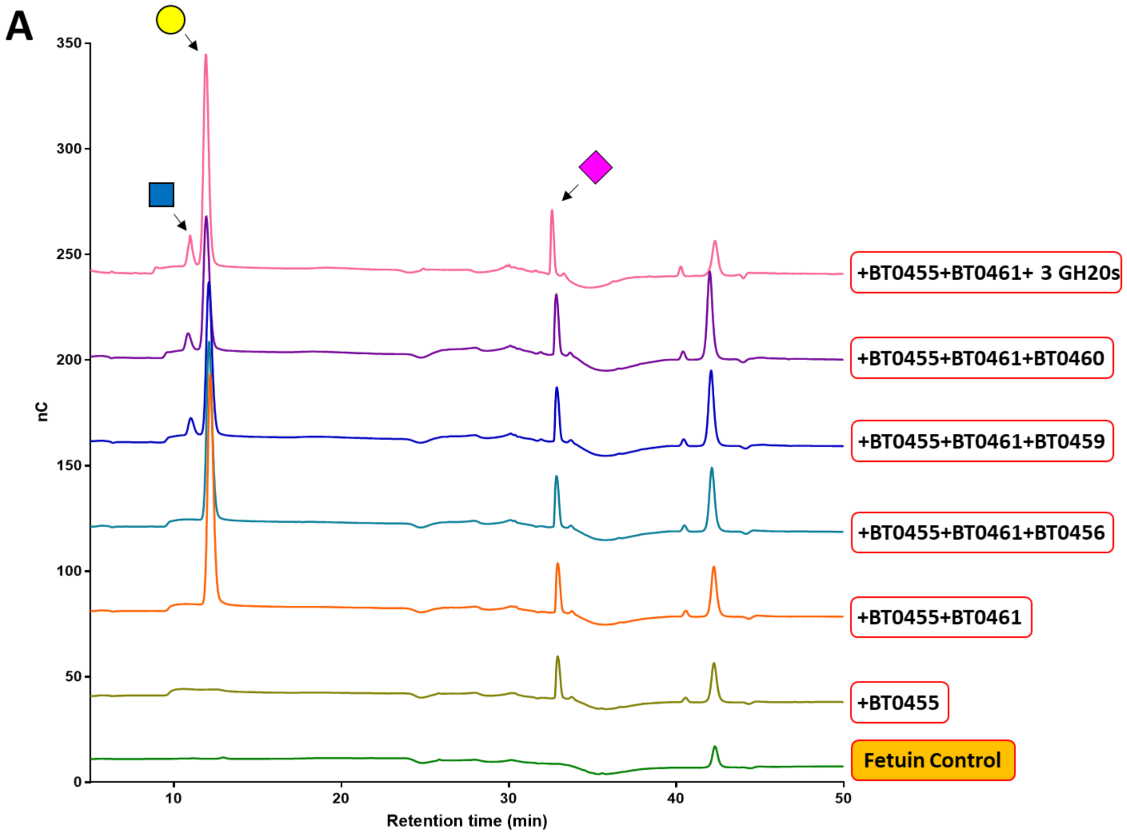

C
PNGaseF-digested Bovine Fetuin
C
+E_{sial}
+E_{sial} +E_{gal}
+E_{sial} +E_{gal} +BT0456
+E_{sial} +E_{gal} +BT0459
+E_{sial} +E_{gal} +BT0460
+E_{sial} +E_{gal} +BT0506
GlcNAc
S.A.
Gal

Figure 4.35: Degradation of N-glycan structures by the β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. A) Native bovine α$_1$-AGP. Arrows indicate a band representative of a GlcNAc product. **B)** GH18-digested bovine α$_1$-AGP. **C)** PNGaseF-digested bovine fetuin. E_{sial} is BT0455; E_{gal} is BT0461; C is control. TLC plate was stained using the DPA stain. GlcNAc is 1mM N-Acetyl-Glucosamine standard. S.A. is 2mM sialic acid (Neu5Ac) standard; Gal is 1mM galactose standard.

To investigate the N-glycan degradation by BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} glycosidases in more detail, the enzyme assays against de-sialylated and de-galactosylated N-glycoproteins were performed and the products were visualised using two different HPAEC-PAD programs (Section 2.3.3). At the time these assays were set up, BT0506^{GH20} had not been cloned. Based on the appearance of a peak representative of a GlcNAc, these results show that all three BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} enzymes release GlcNAc from native and PNGAseF-digested bovine fetuin pre-treated with the sialidase BT0455^{GH20} and galactosidase BT0461^{GH20} (**Figure 4.36**). Based on the size of the peak in **Panel A** when all three GH20s were incubated with native fetuin, more GlcNAc release is observed than in individual GH20 assays, suggesting that each of the BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} enzymes may target different GlcNAc linkages. Similar pattern was observed in **Panel B**. GH20 treatment of the PNGaseF-liberated fetuin N-glycans revealed a range of peaks that, based on their absence in **Panel A**, correspond to different N-glycan products (**Figure 4.36-B**). However, due to the lack of N-glycan standards, they could not be identified. These results indicate that BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} could target different GlcNAc linkages present on tetra-antennary bovine fetuin N-glycans, such as β1,2/4/6 and bisecting β1,4- linkage. Judging by the size of a GlcNAc peak, a notable difference in BT0456^{GH20} activity on native fetuin compared to PNGaseF-cleaved fetuin N-glycans was observed, indicating a preference for N-glycan structures liberated from protein backbones (**Figure 4.36-A and B**). Interestingly, no such difference in activity was observed in BT0459^{GH20} and BT0460^{GH20} assays. Combined, these results support the suggestion that the N-glycan de-glycosylation by the B.theta$^{0455\text{-}0461}$ enzymes is a sequential process



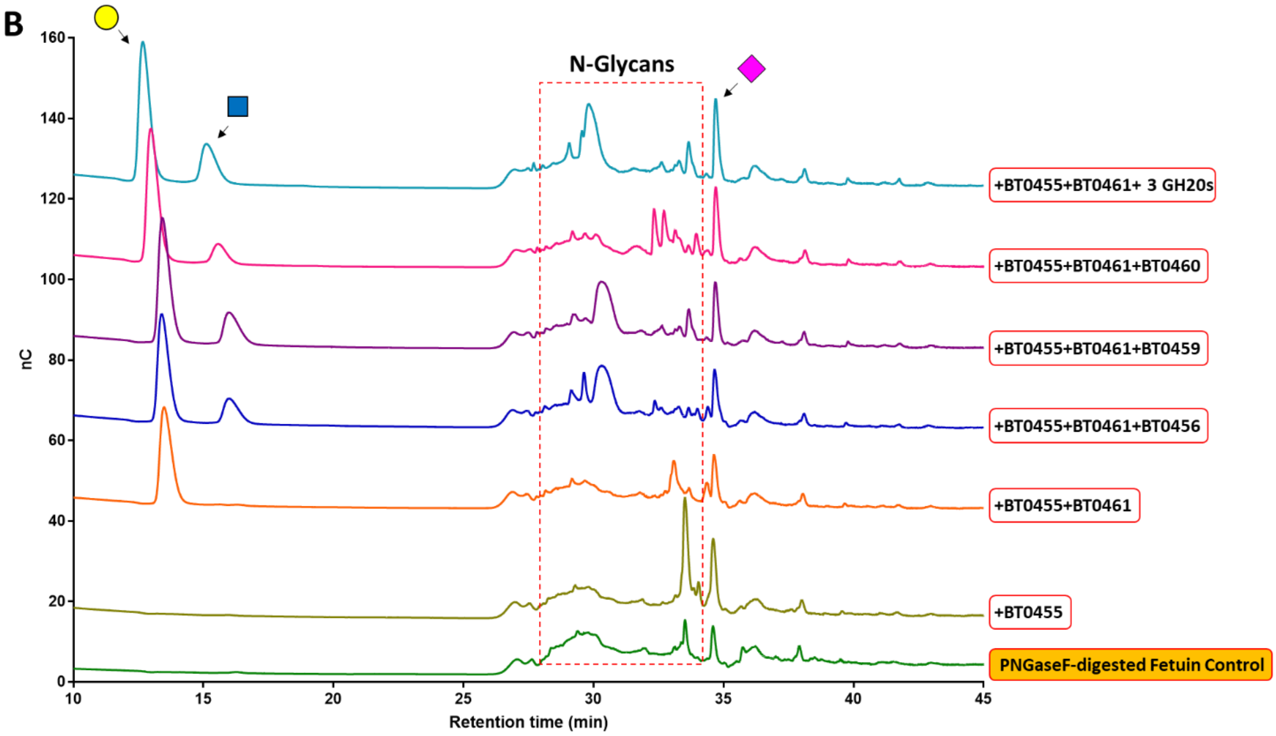
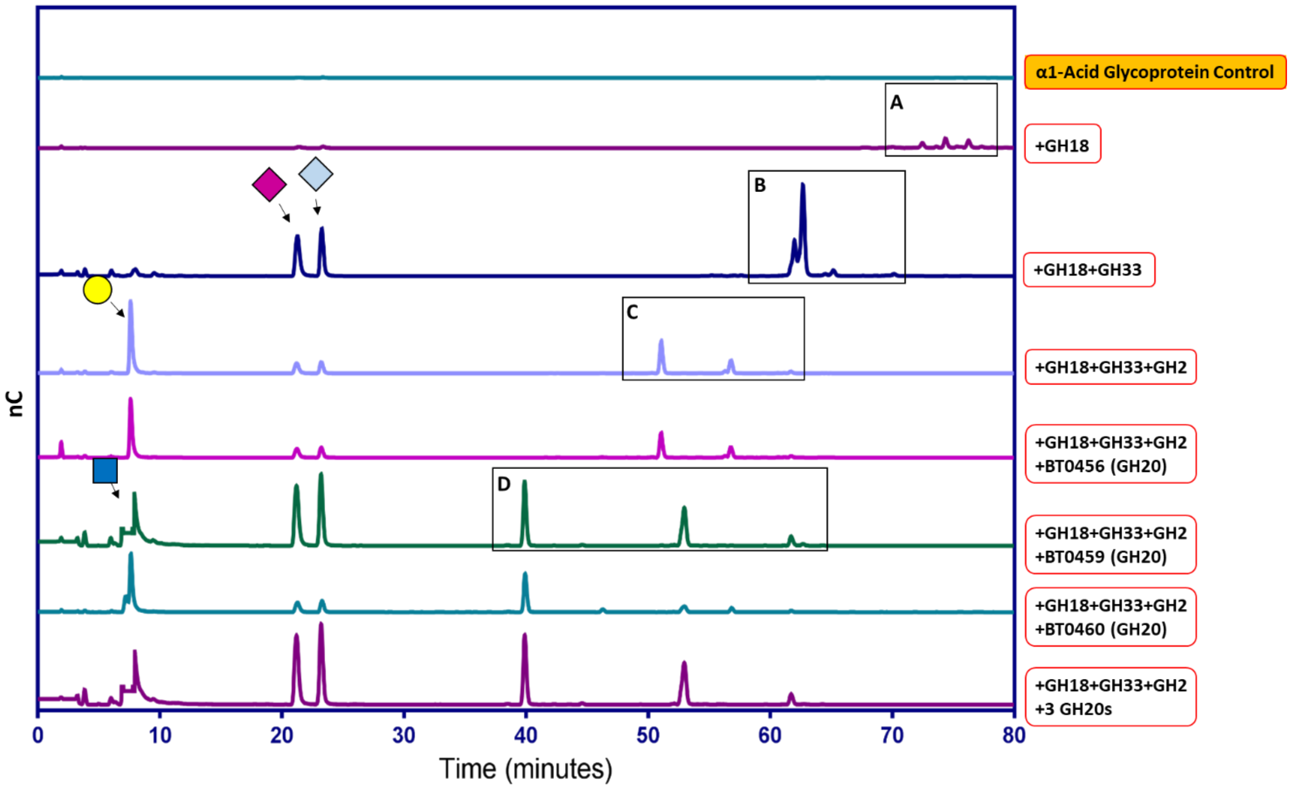


Figure 4.36: Exploring the activity of BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} against bovine fetuin. HPAEC-PAD chromatograms showing the release of N-Acetylglucosamine (GlcNAc). N-glycoproteins were pre-treated with the sialidase BT0455^{GH33} and galactosidase BT0461^{GH2} because GH20s were previously confirmed to be exo-glycosidases. **A)** Activity on native bovine fetuin. **B)** Activity on PNGaseF-digested bovine fetuin. 10mg/ml of N-glycoproteins were incubated with 1µM of enzymes in 20mM sodium phosphate (pH 7) buffer for 16 h at 37°C. 1:10 dilution was loaded onto the HPAEC-PAD. The concentration of monosaccharide standards used is 1mM. Predicted N-glycan elution peaks are highlighted in red box. The difference in size of galactose and GlcNAc elution peaks is due to detection response difference. Two different elution programs were used due to the requirement of the N-glycan peak separation of PNGaseF-digested sample.

To investigate the degradation of N-glycan structures in more detail, the GH20 enzyme assays against α_1-AGP were performed and products were visualised using the U-HPLC combined with the LC-ESI-MS analysis (see Section 2.3.8). The results displayed in **Figure 4.37** show a sequential degradation of α_1-AGP N-glycans by BT1044^{GH18}, BT0455^{GH33}, BT0461^{GH2} and the three GH20s. Once again, BT0506^{GH20} had not been cloned at this stage. Based on no change observed between the peaks present in +GH18+GH33+GH2 sample (**Figure 4.37-C**) and the BT0456-treated sample, these results indicate that BT0456^{GH20} does not appear to display any activity against biantennary N-glycans present on α_1-AGP. Meanwhile, BT0459^{GH20} and BT0460^{GH20} appear to have similar substrate specificities, with BT0459^{GH20} producing more end-product. Judging by the mass spectrometry results displayed in **Figure 4.37-B**, endo-β-N-acetylglucosaminidase BT1044^{GH18} cleaves predominantly sialylated biantennary complex N-glycan structures off α_1-AGP. These structures mainly contain GlcNAc-β1,2-Man linkages, thus explaining similar activities off BT0459^{GH20} and BT0460^{GH20}. These results also suggest that BT0456^{GH20} cannot target GlcNAc-β1,2-Man linkages of N-glycan structures, explaining low upregulation data observed in the RNA seq analysis on α_1-AGP (**Figure 3.8**).



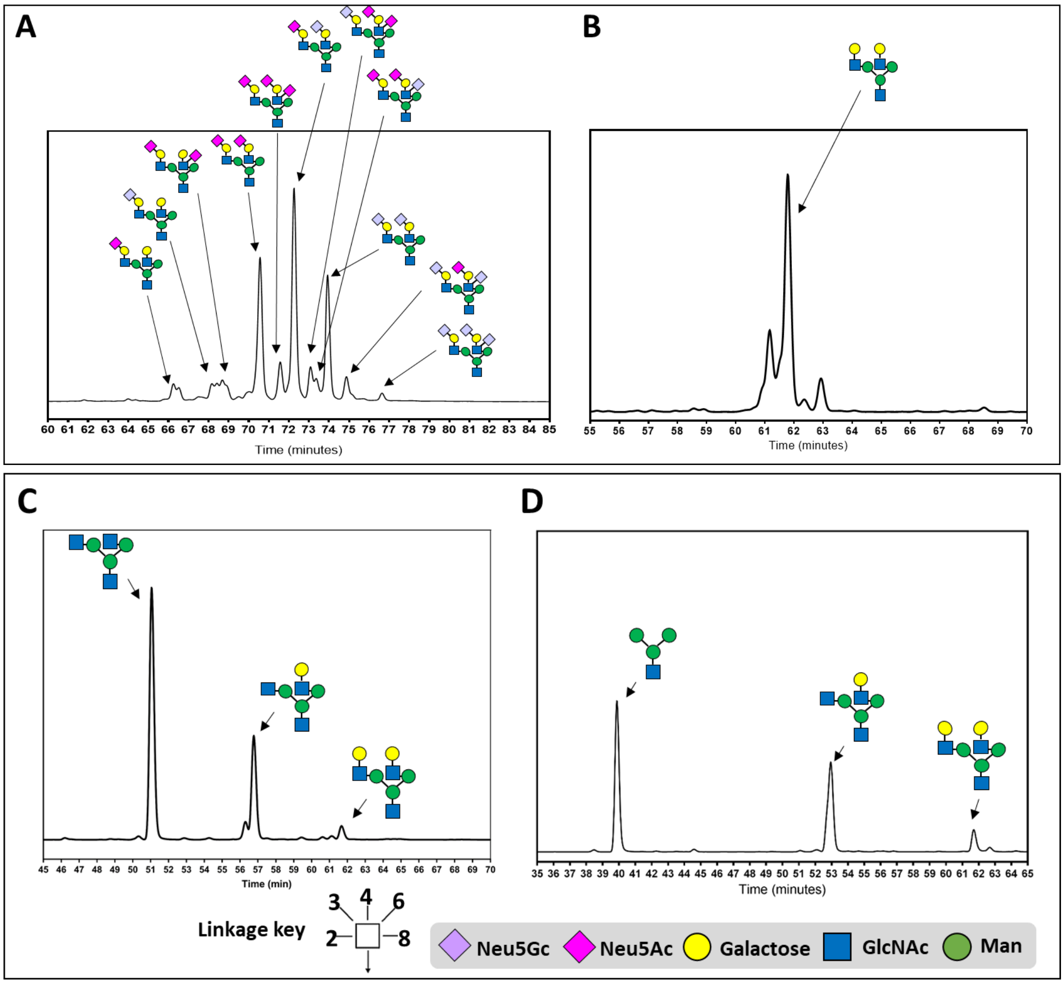


Figure 4.37: Degradation of α$_1$-AGP N-glycans. Top panel: HPLC analyses of α$_1$-AGP degradation by BT1044^{GH18}, BT0455^{GH33}, BT0461^{GH2}, BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20}. Based on the height of sialic acid peaks, loss of HPLC detection sensitivity was observed in some of the samples. Bottom panel: Figures A, B, C and D show N-glycan structures corresponding to each U-HPLC peak that were constructed using the mass spectrometry data (See Table 4.1 for details).

To increase our understanding of the roles of BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} in N-glycan breakdown, enzyme assays against NGA4 N-glycan were set up. NGA4 is a complex N-glycan structure found on several mammalian glycoproteins, including human α$_1$-AGP (Ludger). It contains GlcNAc-β1,2/4/6-Man linkages. The assay was incubated for 16h at 37°C, products were labelled with procainamide and visualised using U-HPLC and mass spectrometry (see Section 2.3.8). The results displayed in **Figure 4.38** show that BT0456^{GH20} can cleave two GlcNAc linkages off the NGA4 structure, however the precise linkages could not be determined. Considering previously obtained data (**Figure 4.37**), these results suggest that BT0456^{GH20} cannot target GlcNAc-β1,2-Man linkages of complex N-glycan structures. In comparison, BT0459^{GH20} can hydrolyse all GlcNAc-β1,2/4/6-Man linkages present on the NGA4 structure, confirming it is a broad-acting enzyme. Whereas a partial degradation of the NGA4 N-glycan structure was observed following the incubation with BT0460^{GH20}. The results suggest that BT0460^{GH20} can cleave all GlcNAc linkages off NGA4 N-glycan structure much like BT0459^{GH20}, but at significantly lower efficiency. Interestingly, it also leaves an N-glycan structure with one GlcNAc linkage still in-tact, supporting a potential importance of a +2 binding sub-site.

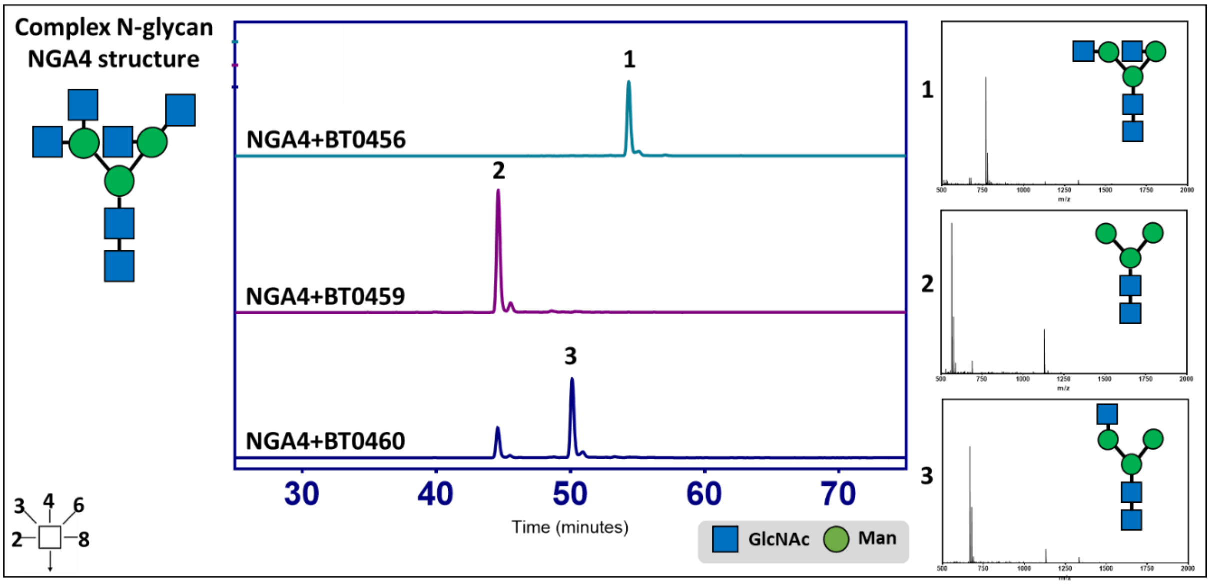


Figure 4.38: Investigating the activity of BT0456^{GH20}, BT0459^{GH20} and BT0460^{GH20} against complex N-glycan structure NGA4. U-HPLC chromatograms of NGA4 ($GlcNAc_4Man_3GlcNAc_2$) degradation are displayed. The predicted N-glycan structures corresponding to the 1, 2 and 3 of HPLC peaks have been constructed using the mass spectrometry data. NGA4 N-glycan was purchased from and enzyme assay products were analysed by Ludger Ltd.

To quantitatively assess the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} degradation of the bovine α_1-AGP N-glycan structures, relative rates of GlcNAc release were measured by setting up timed enzyme assays following a protocol described in Section 2.3.3. BT0456^{GH20} showed no detectable activity against bovine α_1-AGP, so it could not be assessed in this experiment. To stay consistent, the α_1-AGP was pre-treated with BT1044^{GH18}, BT0455^{GH33} and BT0461^{GH2} prior to incubating with the GH20 enzymes. Based on the appearance of a peak representative of the N-glycan tetrasaccharide, these results show BT0459^{GH20} and BT0506^{GH20} are the most effective enzymes against the GlcNAc linkages present on α_1-AGP N-glycans, capable of removing all GlcNAc caps within 40 minutes of incubation (**Figure 4.39**). In comparison, BT0460^{GH20} appeared to be relatively inefficient. Consistent with previously observed results (**Figure 4.38**), BT0460^{GH20} leaves N-glycan structure with one GlcNAc linkage still present after 16 hours of incubation, suggesting it cannot access it.

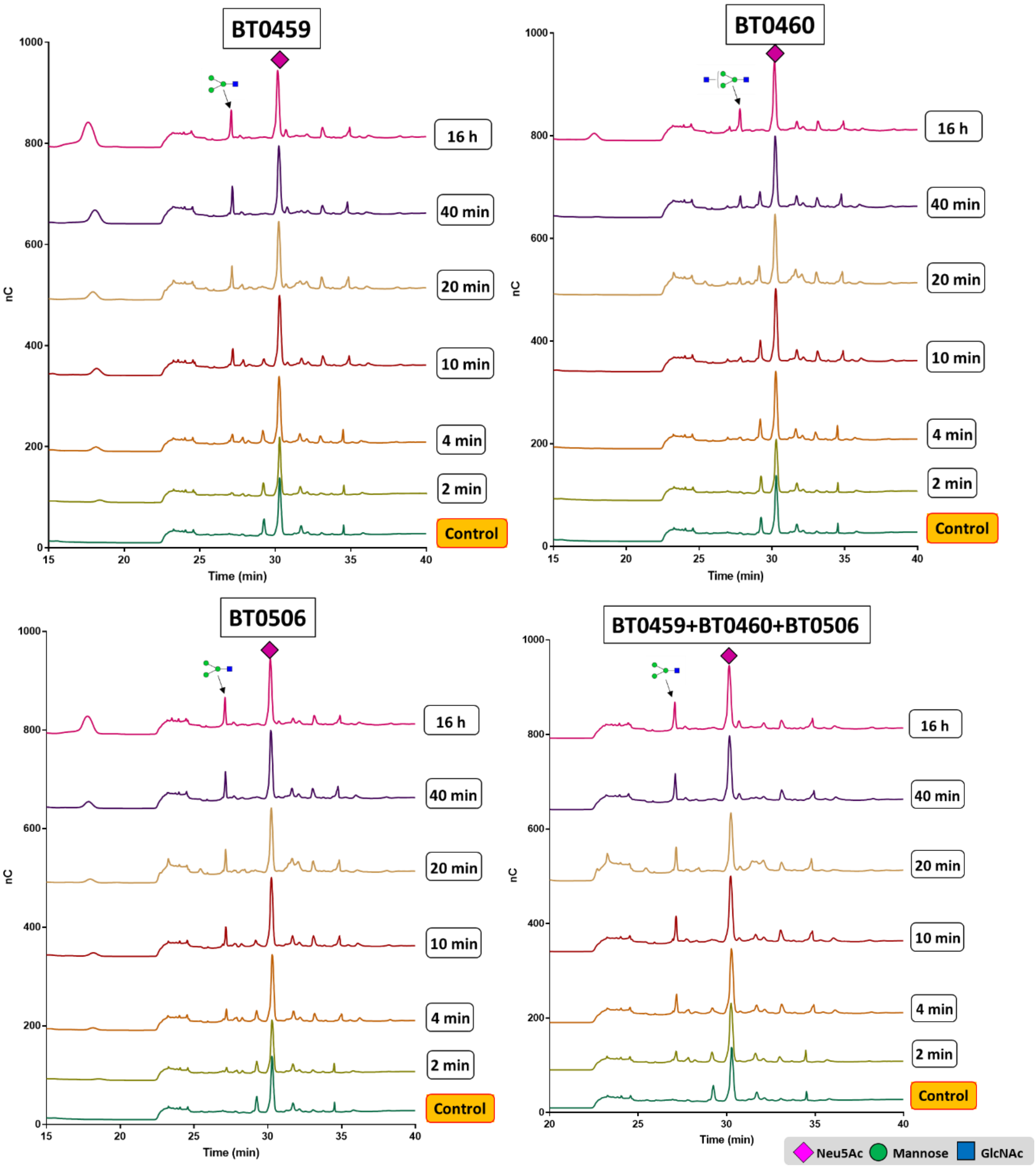


Figure 4.39: Investigating the efficiency of N-glycan degradation by BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. HPAEC-PAD chromatograms showing the timed degradation of N-glycans by the three GH20s. 10mg/ml of N-glycoprotein was incubated with 1µM of each of enzymes in 20mM sodium phosphate (pH 7) buffer for up to 16 h at 37°C. Samples were diluted 1:10 before loading. Elution peaks representative of the N-glycan tetrasaccharide ($Man_3GlcNAc_1$) and N-glycan pentasaccharide ($GlcNAc_1Man_3GlcNAc_1$) products are labelled with arrows.

The products of the bovine α_1-AGP N-glycan degradation by the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} after 1 hour of incubation at 37 °C are displayed in **Figure 4.40**. Control is 10mg/ml of liberated α_1-AGP N-glycan mixture pre-treated with BT0455^{GH33} and BT0461^{GH2} enzymes. Based on the disappearance of the peaks representative of complex N-glycan structures that contain exposed GlcNAc linkages, these results show that BT0459^{GH20} and BT0506^{GH20} catalyse the hydrolysis of all exposed GlcNAc linkages. Consistent with previous results, BT0460^{GH20} appears to cleave only one of the GlcNAc linkages present on the biantennary N-glycan structures. Interestingly, the data also suggests that BT0460^{GH20} prefers the PNGaseF-cleaved N-glycan over the GH18-cleaved N-glycan structure, a structure it would not normally see. None of the enzymes appear to have any activity on Gal-β1,4-GlcNAc (LacNAc) linkages, further confirming that they are exo-glycosidases.

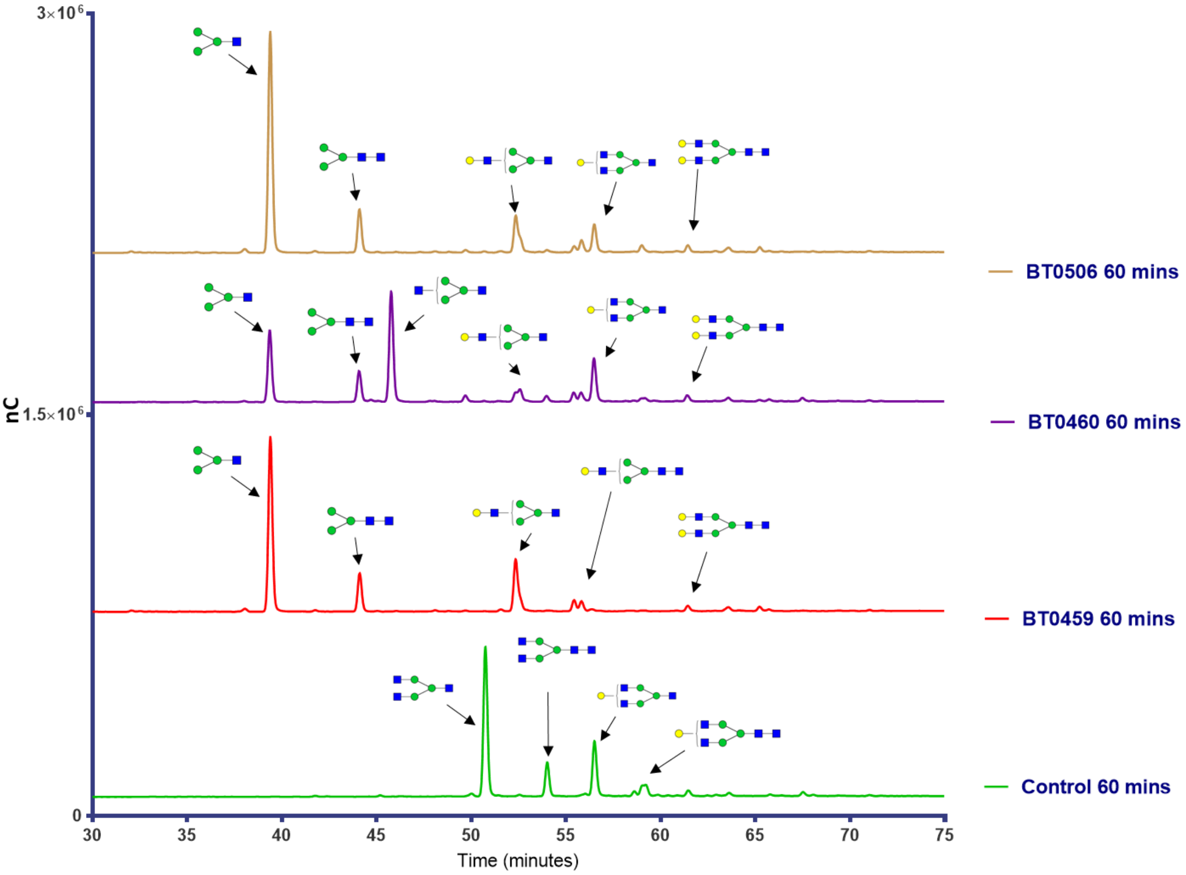


Figure 4.40: The degradation of bovine α_1-AGP N-glycans by BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. Enzyme assays were set up using 10mg/ml of substrate and 1µM of each of enzymes. The reaction was incubated for 1 hour, inactivated and products were labelled with procainamide before being analysed using U-HPLC and mass spectrometry. The N-glycan structures corresponding to the U-HPLC peaks have been constructed using the mass spectrometry data (Ludger) and are labelled with arrows.

To explore the catalytic efficiency of the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases in more detail, the complex GH18-digested N-glycan degradation and the appearance of the N-glycan core tetrasaccharide (Manα1-3(Manα1-6)Manβ1-4GlcNAc) product was monitored. GH18-digested N-glycan mixture was chosen because *B. thetaiotaomicron* does not produce a PNGaseF-like enzyme. Considering that GH20s were previously confirmed to be exo-glycosidases, N-glycan mixture was also pre-treated with the BT0455^{GH33} and BT0461^{GH2} enzymes. The N-glycan core tetrasaccharide peak was identified using the mass spectrometry data. The amount of tetrasaccharide produced by each enzyme was estimated based on the peak area (nC*min) obtained in the HPAEC-PAD chromatogram using the Chromeleon software. The results were plotted using GraphPad Prism 7.0 software and analysed (**Figure 4.41**). When incubated with the N-glycans derived from bovine α_1-AGP, BT0459^{GH20} and BT0506^{GH20} displayed similar enzymatic efficiency and over time generated a higher amount of tetrasaccharide (Manα1-3(Manα1-6)Manβ1-4GlcNAc) compared to the BT0460^{GH20}. Instead, BT0460^{GH20} was found to generate a high amount of an N-glycan pentasaccharide (GlcNAcβ1-2Manα1-3(Manα1-6)Manβ1-4GlcNAc)(**Figure 4.41-B**). These results are consistent with the degradation pattern of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} observed in **Figures 4.38, 4.39 and 4.40**.

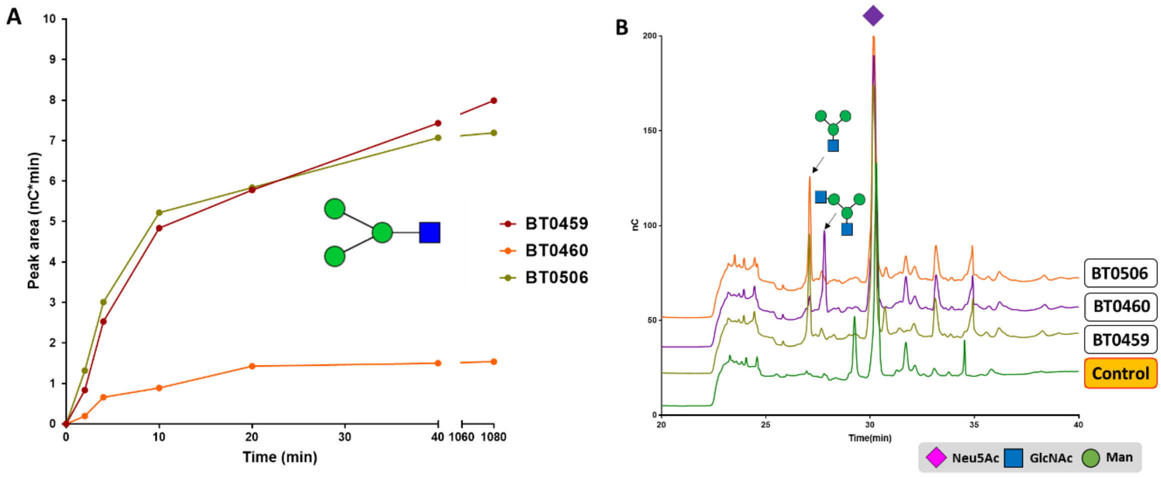
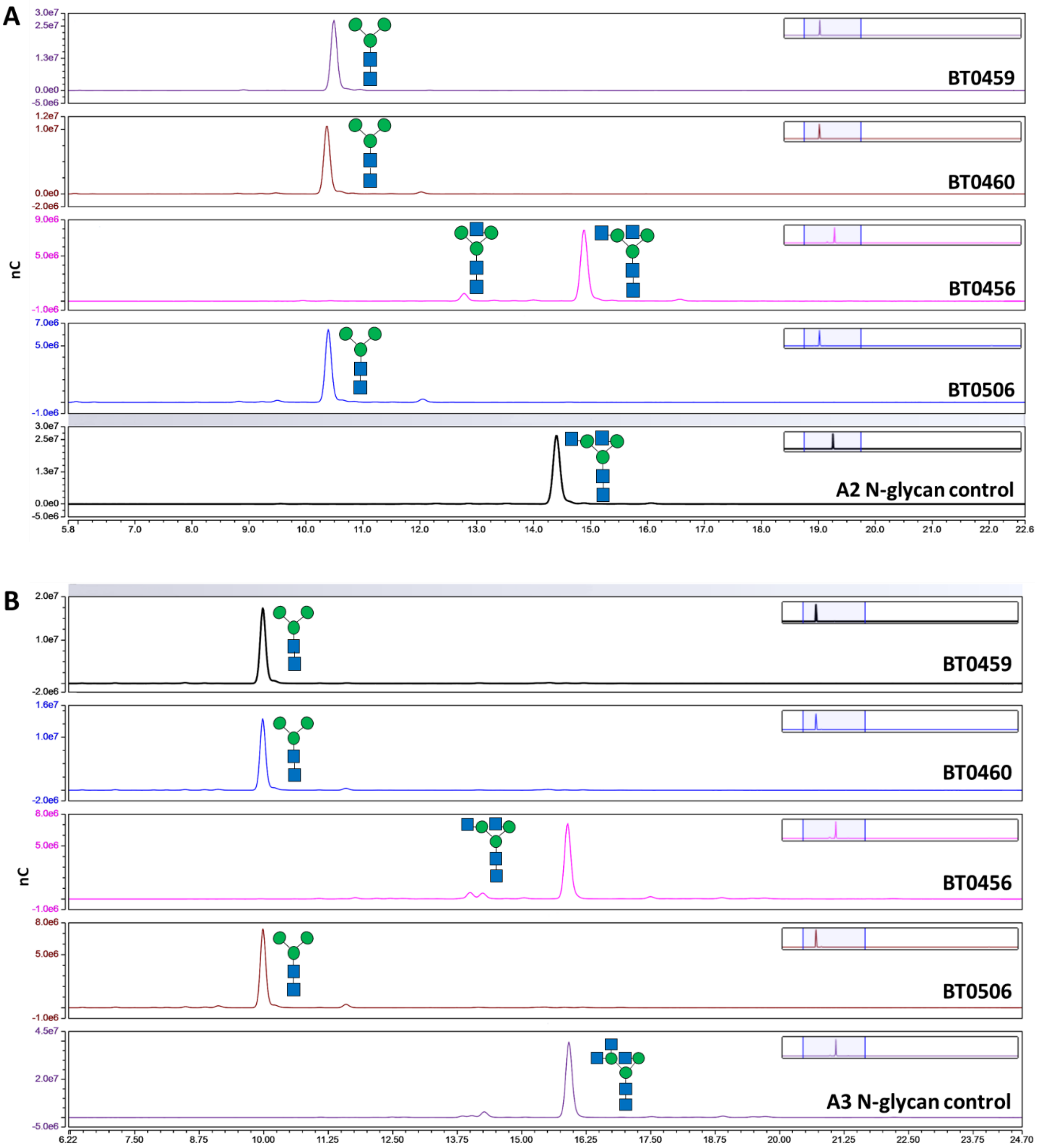


Figure 4.41: Monitoring the N-glycan core tetrasaccharide production by BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. **A)** The production of N-glycan core tetrasaccharide (Manα1-3(Manα1-6)Manβ1-4GlcNAc) over time. Peak area (nC*min) corresponding to the N-glycan core tetrasaccharide was recorded and plotted. Peaks corresponding to these structures are labelled with arrows. **B)** HPAEC-PAD chromatograms showing the N-glycan degradation products generated by BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} after 18 hours of incubation. The experiment was repeated three times and same pattern of degradation was observed.

To clarify the linkage specificity of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, enzyme assays against known structures of N-glycans available commercially at our industrial partner Ludger Ltd were set up following a protocol described in Section 2.3.8. The procainamide-labelled products were visualised using the U-HPLC combined with the mass spectrometry (LC-ESI-MS) analysis to accurately identify the N-glycan structures. The results displayed in **Figure 4.42** suggest that BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} have an overlapping complex N-glycan specificities and can fully deglycosylate A2, A3 and A4 N-glycan structures. Based on the changes in the size of the peaks representative of N-glycan structures, BT0456^{GH20} seems to display the lowest efficiency, struggling to cleave GlcNAc-β1,2-Man linkages of A2 and A3 N-glycan structures. BT0456^{GH20} can hydrolyse the GlcNAc-β1,4-Man linkage of A3 N-glycan structure but shows no activity on A4 N-glycan structure. These results are consistent with the previous findings that BT0456^{GH20} is not specialised in targetting GlcNAc-β1,2-Man linkages (**Figures 4.37 and 4.38**). Interestingly, BT0460^{GH20} seems to prefer bisecting GlcNAc linkages on IgG N-glycans whereas BT0506^{GH20} cannot target them at all. Meanwhile, BT0456^{GH20} and BT0459^{GH20} show a trace activity against bisecting GlcNAc, although the activity of BT0459^{GH20} against it seems to be impaired by the presence of core fucose on IgG N-glycans (**Figure 4.42-D**). Consistent with the previously obtained results, these data show that BT0459^{GH20} is a very efficient, broad-acting β-hexosaminidase.


A
nC
3.0e7
2.5e7
1.3e7
0.0e0
-5.0e6
BT0459
1.2e7
1.0e7
0.0e0
-2.0e6
BT0460
9.0e6
5.0e6
-1.0e6
BT0456
7.0e6
5.0e6
-1.0e6
BT0506
3.0e7
2.5e7
1.3e7
0.0e0
-5.0e6
A2 N-glycan control
5.8
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
22.6
B
nC
2.0e7
1.0e7
-2.0e6
BT0459
1.6e7
1.0e7
-2.0e6
BT0460
8.0e6
5.0e6
-1.0e6
BT0456
8.0e6
5.0e6
-1.0e6
BT0506
4.5e7
2.0e7
-5.0e6
A3 N-glycan control
6.22
7.50
8.75
10.00
11.25
12.50
13.75
15.00
16.25
17.50
18.75
20.00
21.25
22.50
23.75
24.70

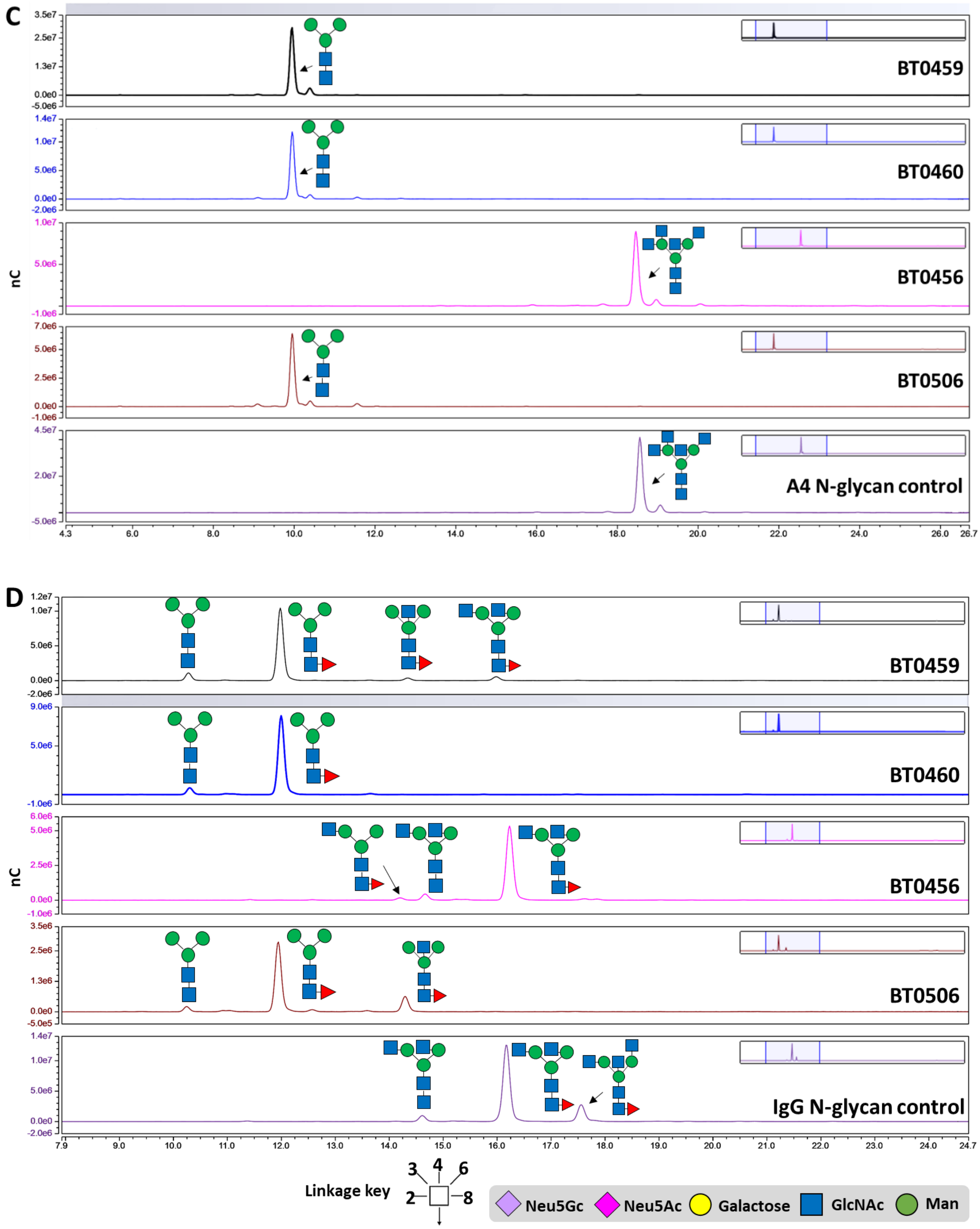
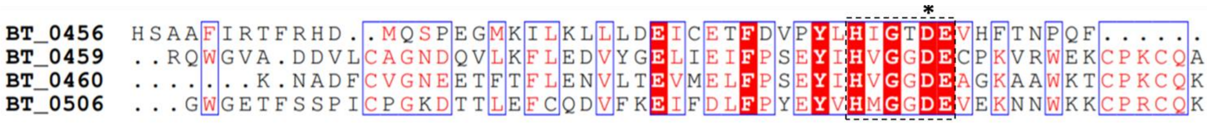


Figure 4.42: Degradation of known N-glycan structures by BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. A) A2 N-glycan (GlcNAc$_2$Man$_3$GlcNAc$_2$) degradation. **B)** A3 N-gycan (GlcNAc$_3$Man$_3$GlcNAc$_2$) degradation. **C)** A4 N-glycan (GlcNAc$_4$Man$_3$GlcNAc$_2$) degradation. **D)** IgG N-glycan (GlcNAc$_3$Man$_3$GlcNAc$_2$Fuc$_1$) degradation. IgG was pre-treated with sialidase BT0455^{GH33} and galactosidase BT0461^{GH2}.

4.3.5.5. Crystallography and structure analysis

4.3.5.5.1. Generating catalytic mutants of the four GH20s

One of the key characteristics of the GH20-family β-hexosaminidases is the presence of the conserved catalytic residue pair – D-E that is usually preceded by the H-x-G-G motif. Amino acid sequence alignment of the four GH20-family enzymes allowed us to identify the conserved catalytic residues of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} (**Figure 4.43)**. Interestingly, BT0456^{GH20} appears to have a H-I-G-T motif instead of H-x-G-G motif seen in other GH20s, suggesting it may have an effect on the catalytic activity of this enzyme. Aspartic acid (D) marked with an asterisk (*) was successfully mutated to generate a catalytic mutant.



```
                                                       *
BT_0456  HSAAFIRTFRHD..MQSPEGMKILKLLLDEICETFDVPYLHIGTDEVHFTNPQF......
BT_0459  ..RQWGVA.DDVLCAGNDQVLKFLEDVYGELIEIFPSEYIHVGGDECPKVRWEKCPKCQA
BT_0460  .......K.NADFCVGNEETFTFLENVLTEVMELFPSEYIHVGGDEAGKAAWKTCPKCQK
BT_0506  ...GWGETFSSPICPGKDTTLEFCQDVFKEIFDLFPYEYVHMGGDEVEKNNWKKCPRCQK
```

Figure 4.43: Identifying catalytic residues of BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases. The sequence alignment was done using Clustal Omega and visualised using the Espript 3.0. Amino acid residues displayed in red boxes are fully conserved. Amino acid residues framed in blue are 70% or more similar based on their physicochemical properties. Catalytic residue is highlighted with an asterisk (*).

The catalytic mutants D332N of BT0459, D328N of BT0460 and D347N of BT0506 were generated by site-direct mutagenesis following a protocol described in Section 2.1.10 and sequenced (**Figure 4.44-A**). The primers used to generate these mutants are displayed in Supplemental table 3.2. The attempt to generate the BT0456^{GH20} catalytic mutant was unsuccessful. The mutant enzymes were expressed and purified following a protocol described in Section 2.1.25. The protein fractions collected from the gel filtration purification were visualised using SDS-PAGE and combined (**Figure 4.44-B**). The purified mutant enzymes were tested and showed no activity against pNP-β-GlcNAc (data not shown). An E333A mutant of BT0459 and E329A mutant of BT0460 were also generated. Surprisingly, after testing them against pNP-β-GlcNAc, it was discovered that they retained their catalytic activity (data not shown).

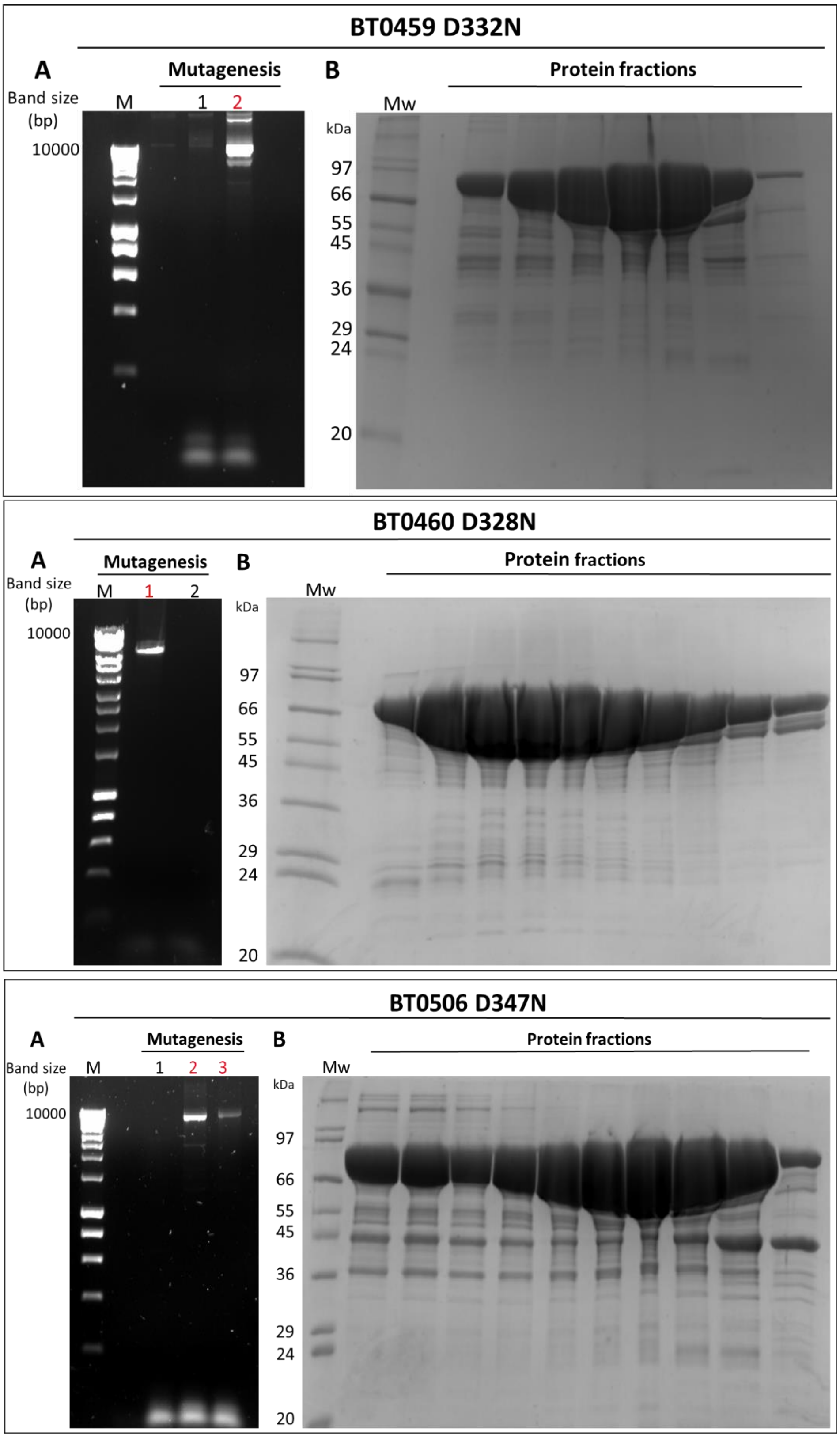


Figure 4.44: Generating catalytic mutants of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. A) Agarose gel showing the PCR products of the site-directed mutagenesis. M is a high-fidelity genomic DNA marker. Site-directed mutagenesis products highlighted in red were sent for sequencing. **B)** SDS-PAGE analysis of purified protein fractions. Mw is a wide-range molecular weight marker (Sigma).

4.3.5.5.2. Generating catalytic mutant crystals

Initially, it was attempted to generate the GH20 catalytic mutant crystals in a bound form with a ligand. The purified BT0459$_{D332N}$, BT0460$_{D328N}$ and BT0506$_{D347N}$ proteins were concentrated to 30mg/ml and co-crystallised with either 100mM GlcNAc-β1,2-Man or high-concentration of complex N-glycan derived from bovine α_1-AGP. This N-glycan substrate was produced by digestion with BT1044^{GH18}, BT0455^{GH33} and BT0461^{GH2} enzymes (see Section 2.3.7). Equal volume (100 nL) of protein and crystal screen solutions (Molecular Dynamics) were mixed using a sitting-drop vapour-diffusion method and the automated Mosquito^{R} nanodrop dispensing robot (see Section 2.4.1.) Crystals were grown at 20 °C. Initial hit conditions were optimised and proteins were co-crystallised with either GlcNAc-β1,2-Man or the N-glycan mixture using the hanging-drop vapour diffusion method following a protocol described in Section 2.4.2. Numerous crystals of BT0459$_{D332N}$, BT0460$_{D328N}$ and BT0506$_{D347N}$ catalytic mutants were obtained from various conditions over the months. The **Figure 4.45** illustrates a BT0459$_{D332N}$ crystal obtained by this approach. The crystal collection and analysis was done by Dr. Arnaud Baslé (ICaMB, NU). Unfortunately, none of the crystals diffracted and no data could be obtained. It was also attempted to soak BT0459$_{D332N}$ apo crystals in a high concentration of N-glycan mixture derived from α_1-AGP using the same method as described above. Unfortunately, it was not successful and due to time constraints it could not be repeated.

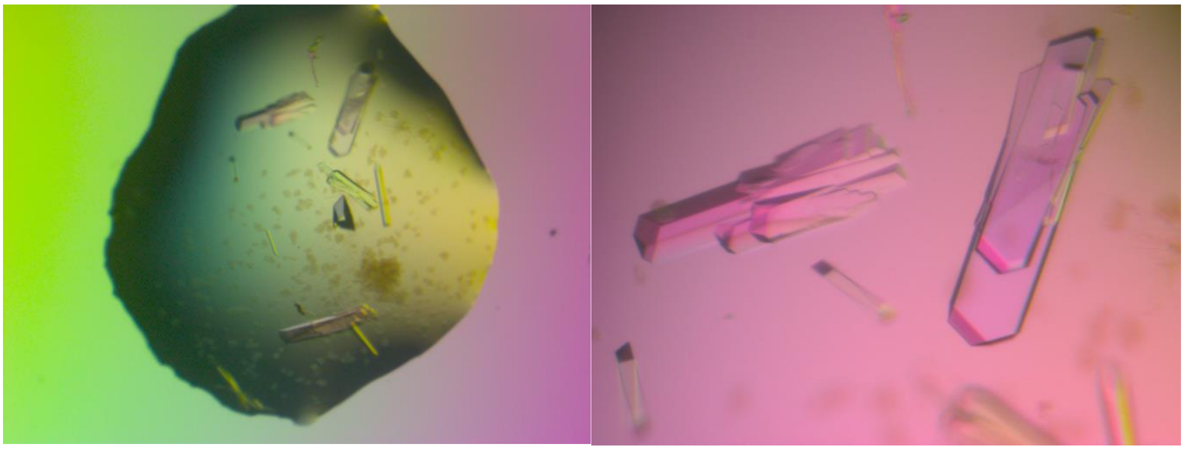


Figure 4.45: Examples of BT0459$_{D332N}$ crystals. Crystals were obtained by co-crystallization of BT0459$_{D332N}$ mutant with GlcNAc-β1,2-Man substrate. Crystals formed in 15% (w/v) PEG3350, 75mM sodium acetate, 0.1M of Bis-Tris (pH 7.5).

4.3.5.5.3. BT0459^{GH20} crystal structure

After the attempt to obtain the crystals of the GH20 catalytic mutants in a bound form with a ligand failed, it was decided to generate the crystal structures of the wild type BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} proteins. The purified wild type proteins were concentrated and initial crystal screens (Molecular Dimensions) were set up using a Mosquito crystallisation robot (TTP Labtech) and sitting-drop crystallisation technique following a protocol described in Section 2.4.1. Initial hits were optimised and set up using the hanging-drop vapour diffusion method following a protocol described in Section 2.4.2. Only high-quality BT0459^{GH20} crystals were obtained after 10 days in 10% (w/v) PEG3350, 0.1M sodium acetate and 100mM Bis-Tris (pH 7.5) (**Figure 4.46**). Due to time constraints, crystals of the other GH20s could not be obtained.

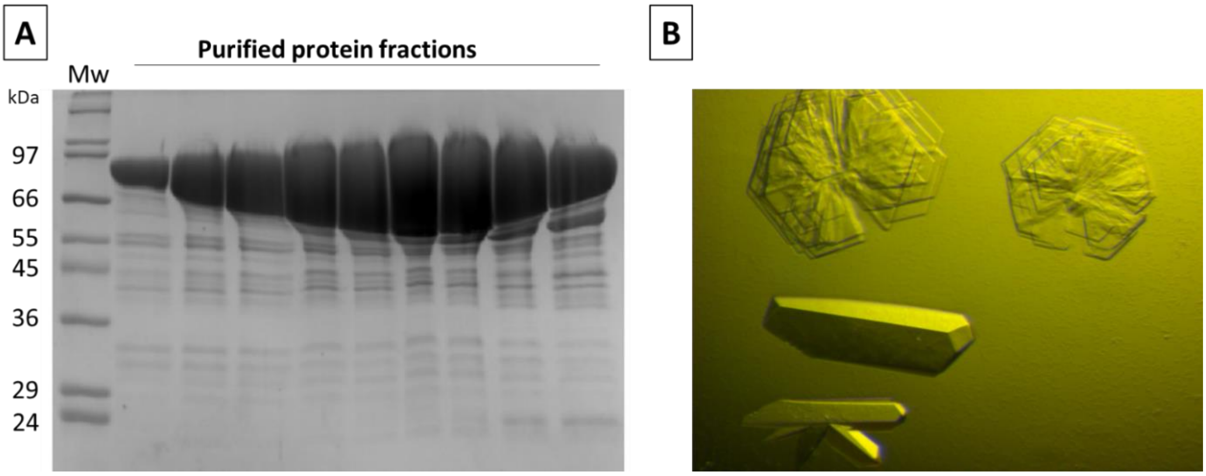


Figure 4.46: Generating BT0459^{GH20} crystal. A) SDS-PAGE gel showing the purification of BT0459^{GH20} protein expressed in *E.coli* Tuner cells. The fractions were combined and concentrated to 30mg/ml. **B)** BT0459^{GH20} crystals formed in 10% (w/v) PEG3350, 0.1M sodium acetate and 100mM Bis-Tris buffer (pH 7.5).

The crystals were harvested and the data sets were collected by Dr. Arnaud Baslé at the Diamond Light Source (DLS). The structure of the wild type BT0459^{GH20} was solved by molecular replacement to a resolution of 2.4 Å. BT0459^{GH20} protein is 87kDa in size and consists of four domains. The BT0459^{GH20} structure presented a C-Terminal putative CBM (F5/8 type C) domain comprised of a β-sandwich fold connected to the β-stranded linker domain (CHB_HEX_C1) followed by a GH20 domain that folds into $(\beta/\alpha)_8$-barrel topology and GH20b domain comprised of an β-sandwich fold and two α-helix strands (**Figure 4.47**).

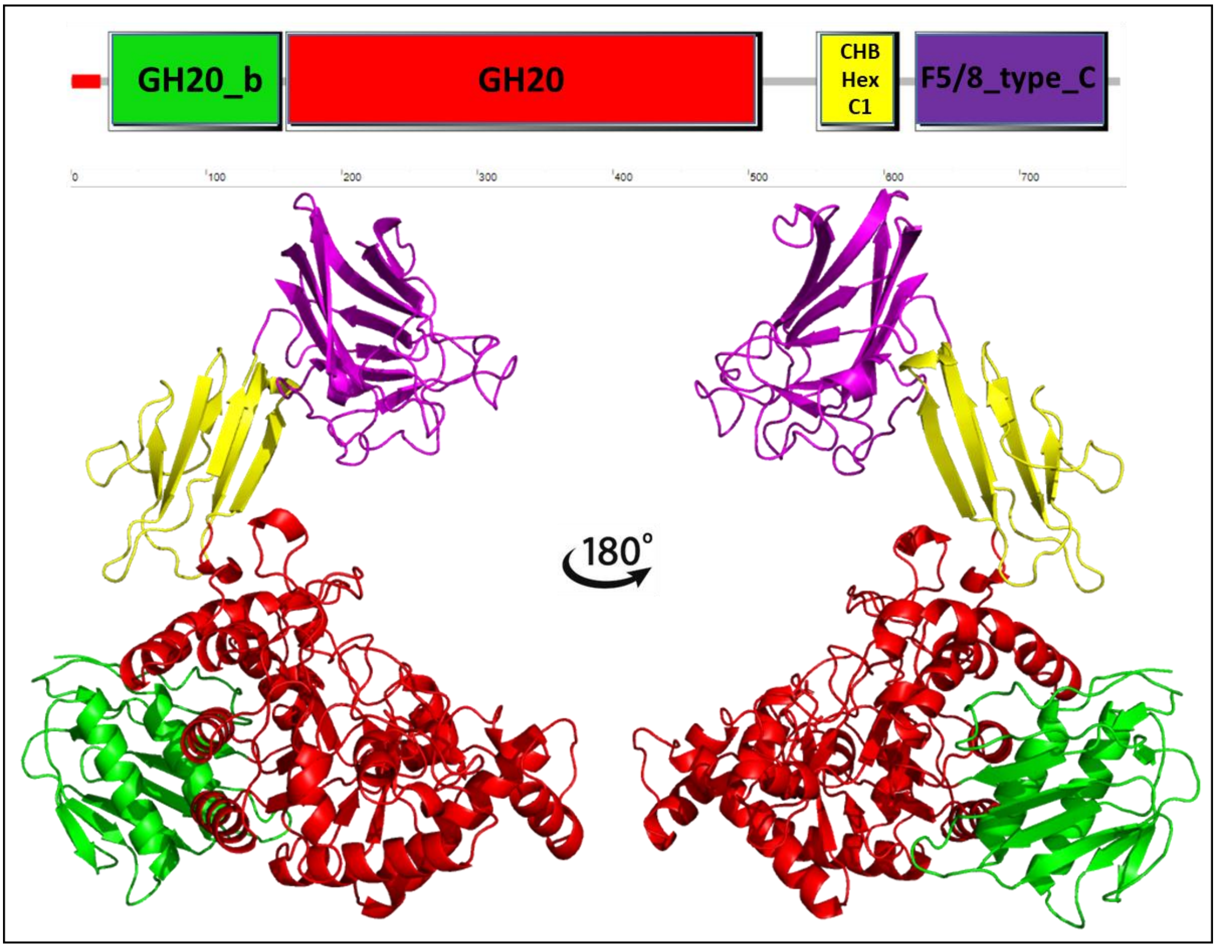


Figure 4.47: Wild type BT0459^{GH20} crystal structure. X-ray structure of BT0459^{GH20} was coloured according to the domain composition. The structure was rotated 180 degrees between the two images.

Because co-crystallisation of BT0459$_{D332N}$ catalytic mutant with an N-glycan or GlcNAc-β1,2-Man ligand failed, it was decided to investigate the structural similarities of the wild type BT0459^{GH20} to a structure of an *exo*-β-D-N-acetylglucosaminidase StrH from *S. pneumoniae* that was previously shown to target N-glycans (Pulvinage *et al.,* 2011). Protein structure comparison server (DALI) and literature review was used to identify this enzyme. StrH is one of the virulence factors of *S.pneumoniae* required for effective human compex N-glycan utilisation. Because StrH is a much larger enzyme and contains two catalytic domains – GH20a and GH20b, the BT0459^{GH20} only shares 21% sequence similarity to StrH. These GH20a and GH20b domains enable StrH to recognise GlcNAc residues on a range of complex N-glycan structures, such as bi-antennary, tetra-antennary and bisecting (Pulvinage *et al.,* 2011). To compare the similarities, the BT0459^{GH20} crystal structure was overlaid with the GH20a and GH20b domains of StrH (PDB ID: 2YLA). The catalytic cores of both of the GH20a and GH20b domains are structurally very similar to the GH20 domain of BT0459^{GH20} (**Figure 4.48**). Furthermore, the GlcNAc-bound -1 subsite of StrH GH20b domain is fully conserved in BT0459^{GH20}, suggesting the similar substrate specificity of these enzymes.

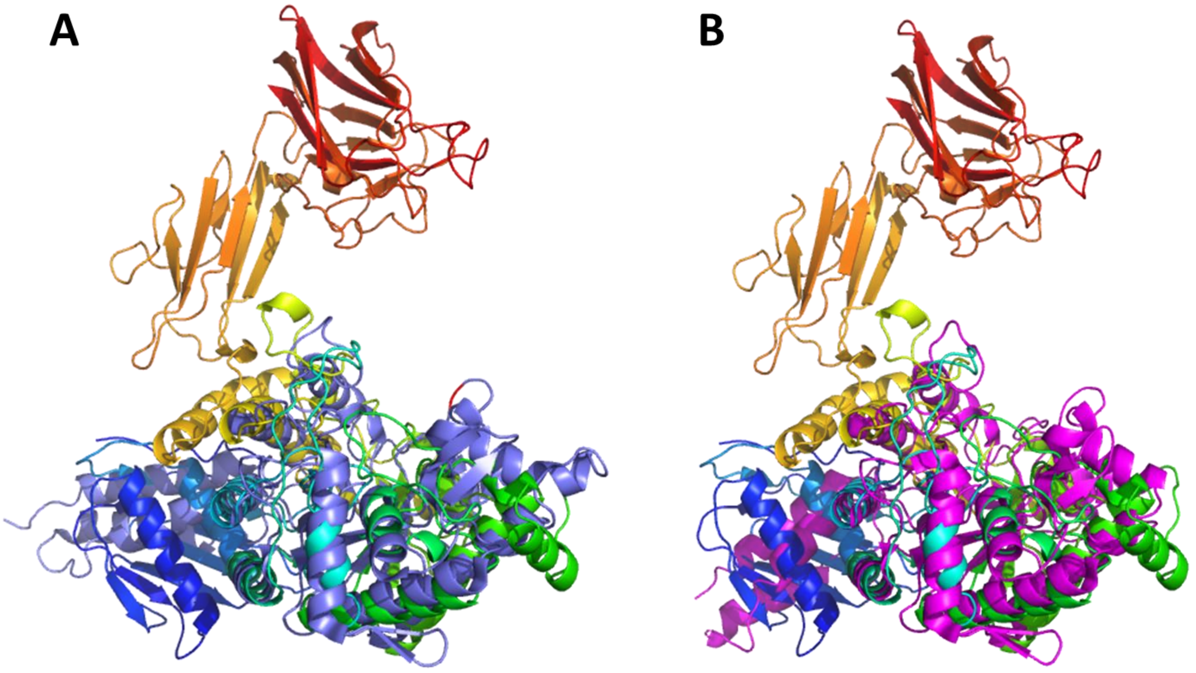
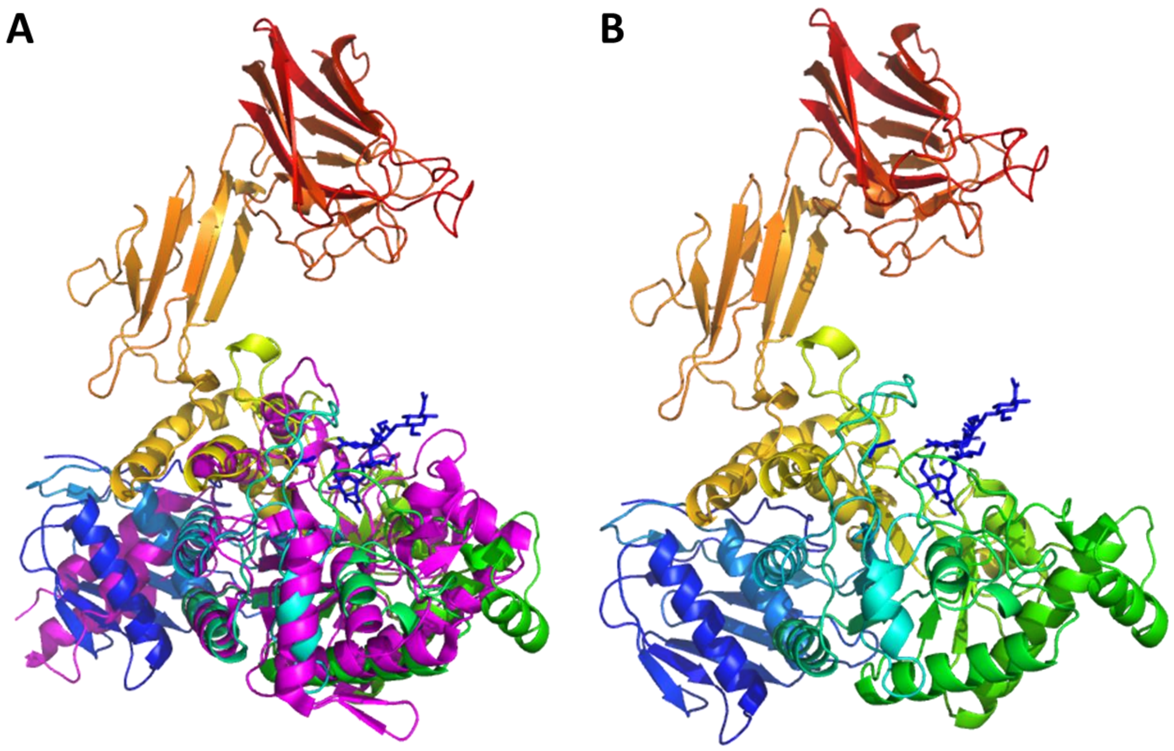


Figure 4.48: Structure of BT0459^{GH20} overlaid with GH20 domains from StrH of S. *pneumoniae*. X-ray structures of BT0459^{GH20} overlaid with **A)** StrepGH20a and **B)** StrepGH20b. X-ray structure of BT0459^{GH20} was coloured rainbow from red (C-terminus) to blue (N-terminus). StrepGH20a was coloured in purple and StrepGH20b was coloured in pink.

Pluvinage *et al.,* 2011 has successfully co-crystallised the *exo*-β-D-N-acetylglucosaminidase StrH from *S. pneumonia* with an N-glycan ligand bound to the active site of the GH20b domain. The crystal structure of BT0459^{GH20} was overlaid with the structure of StrH GH20B$_{E805Q}$ bound to the bisecting N-glycan structure NGA2B (GlcNAcβ1,2Manα1,3(GlcNAcβ1,2Manα1,6(GlcNAcβ1,4))Manβ1,4GlcNAc) (**Figure 4.49**). The results showed that StrepGH20b has an aromatic ramp (W877, yellow) structure that serves as a platform to anchor the N-glycan in place but BT0459^{GH20} seem to lack such loop structure (**Figure 4.49-C**). BT0459^{GH20} appears to have a very open binding site, with the catalytic residues D332 and E333 in a close proximity of the modelled N-glycan structure. There also appears to be an aromatic tryptophan (W405) close to the active site, suggesting it may be required for the BT0459^{GH20} interactions with the complex N-glycan structures. Combined, these results suggest BT0459^{GH20} may employ a different N-glycan binding strategy.



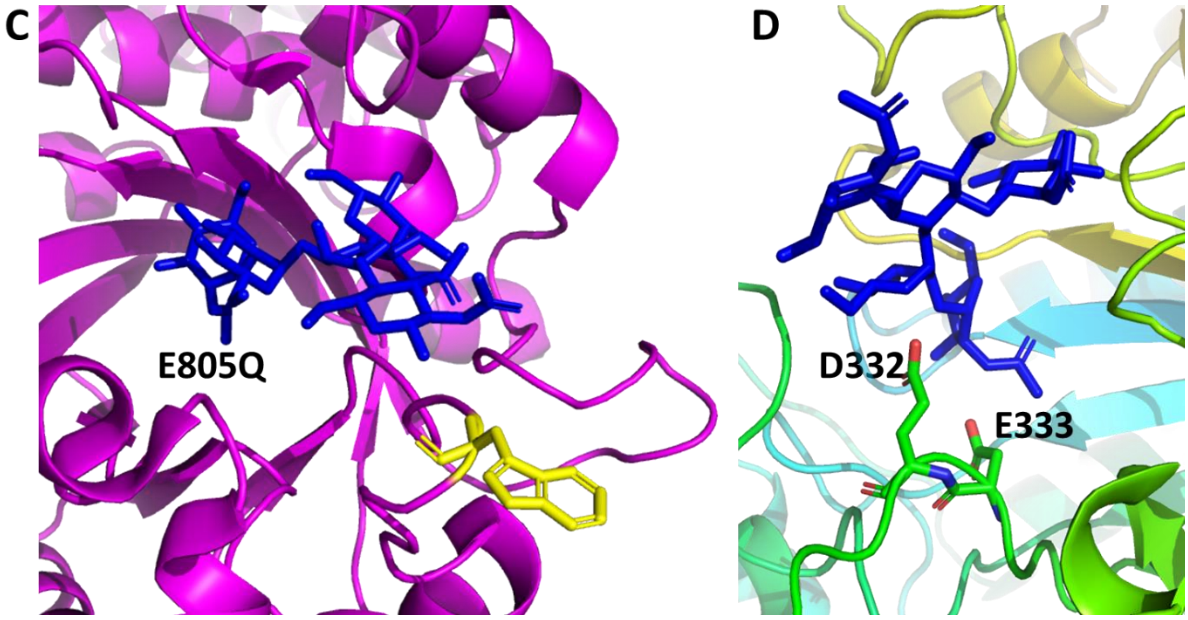


Figure 4.49: Investigating the predicted N-glycan binding site of BT0459^{GH20} crystal structure. A) X-ray structure of BT0459^{GH20} overlaid with StrepGH20b bound to the N-Glycan structure. **B)** Model of BT0459^{GH20} bound to the N-glycan structure. **C)** StrH GH20B catalytic mutant E805Q bound to the model N-glycan structure. The aromatic ramp structure is coloured yellow. **D)** BT0459^{GH20} catalytic residues D332 and E333 in close proximity of the model N-glycan structure. X-ray structure of BT0459^{GH20} was coloured rainbow from red (C-terminus) to blue (N-terminus) and StrepGH20b was coloured in pink. N-glycan structure is shown as blue sticks.

The crystal structure of the BT0459^{GH20} suggests that it is in an open-state conformation. The relative position of the putative CBM domain (F5/8 type C) compared to the active site indicates that BT0459^{GH20} may adopt a conformational change in order to bind large carbohydrate substrates, such as N-glycans. A protein structure comparison server (DALI) search revealed that the BT0459^{GH20} F5/8 type C domain was structurally related to the CBM32 from a *Clostridium perfringens* β-hexosaminidase (Ficko-Blean *et al.,* 2006; PDB ID: 1TVG), suggesting the F5/8 type C domain may be a CBM. A structural alignment showed that BT0459^{GH20} F5/8 type C domain is almost identical to the CpCBM32 but it contains large loop structures that may be required to anchor N-glycans in place (**Figure 4.50**).

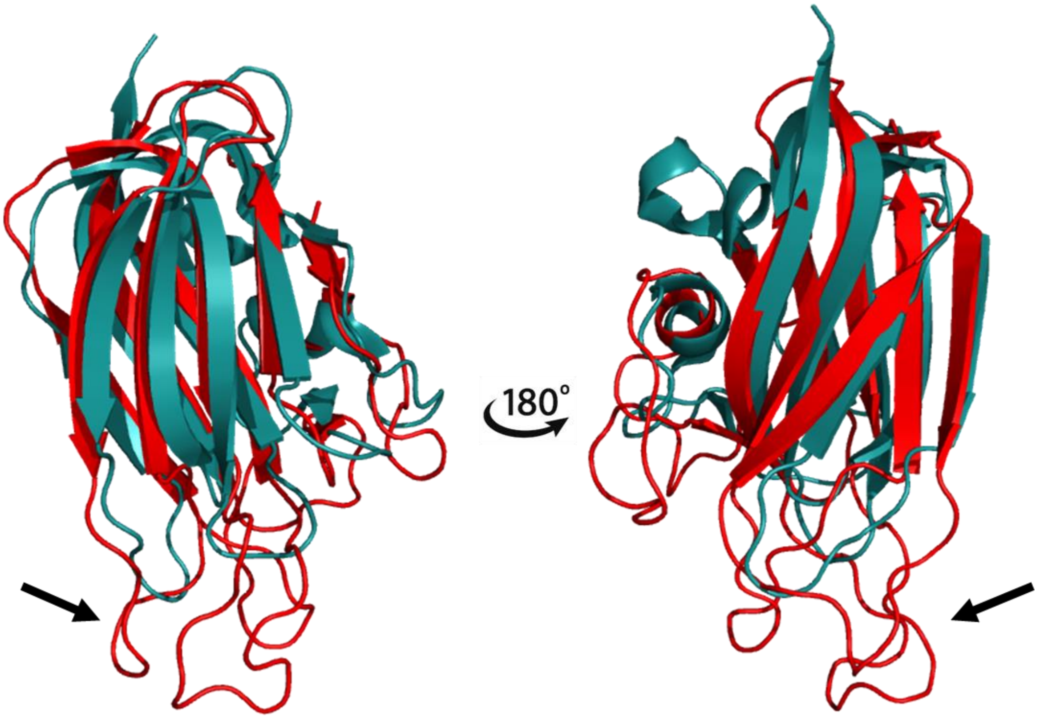


Figure 4.50: The comparison of the BT0459^{GH20} CBM domain to the CBM32. X-ray structure of a putative BT0459^{GH20} CBM domain (F5/8 type_C) was coloured red and CBM32 (PDB ID: 1TVG) was coloured in teal. The structure was rotated 180 degrees between the two images. The large loop structures are highlighted with black arrows.

Based on the search results using the Conserved Domain Architecture Retrieval Tool (NCBI.org), the domain composition of the BT0459^{GH20} is quite unique. Out of 27196 β-hexosaminidases produced by bacteria, only 1449 contained this domain composition with majority being GH20 glycosidases encoded by members of *Bacteroides* species. Due to the lack of aromatic ramp present in StrH and relative positioning of the model N-glycan, it was speculated that the CBM domain (F5/8 type C) together with the linker domain (CHB_HEX_C1) may be involved in the N-glycan binding. The binding model showing the conformational change of the open state BT0459^{GH20} to the closed state the BT0459^{GH20} was made. The model predicts the CBM domain movement towards the BT0459^{GH20} active site to bind and anchor the glycan structures in place (**Figure 4.51**). The similar kind of movement has been observed by the CBM domain of a GH5_4 family enzyme from *B. licheniformis* (Liberato *et al.,* 2016).

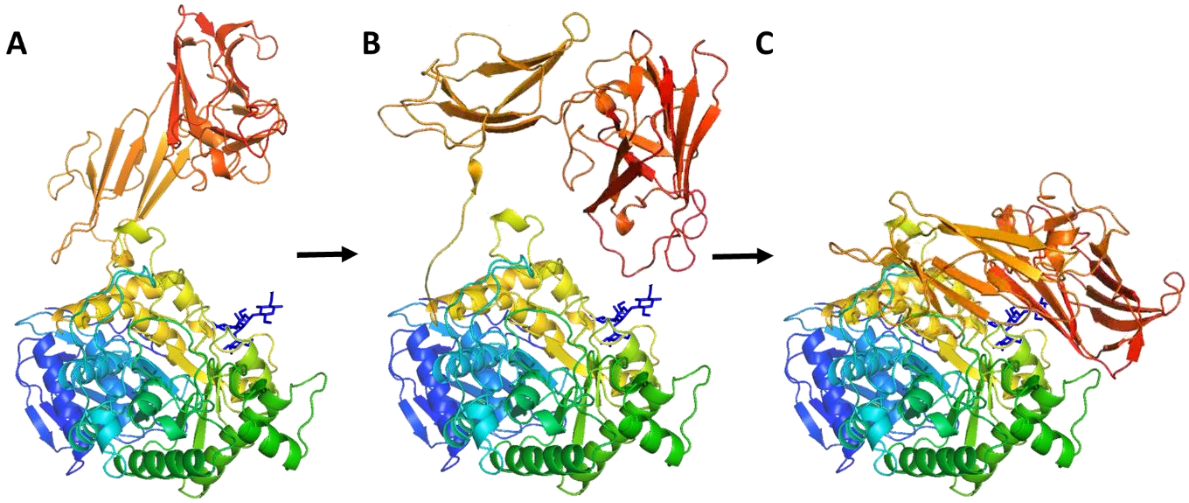
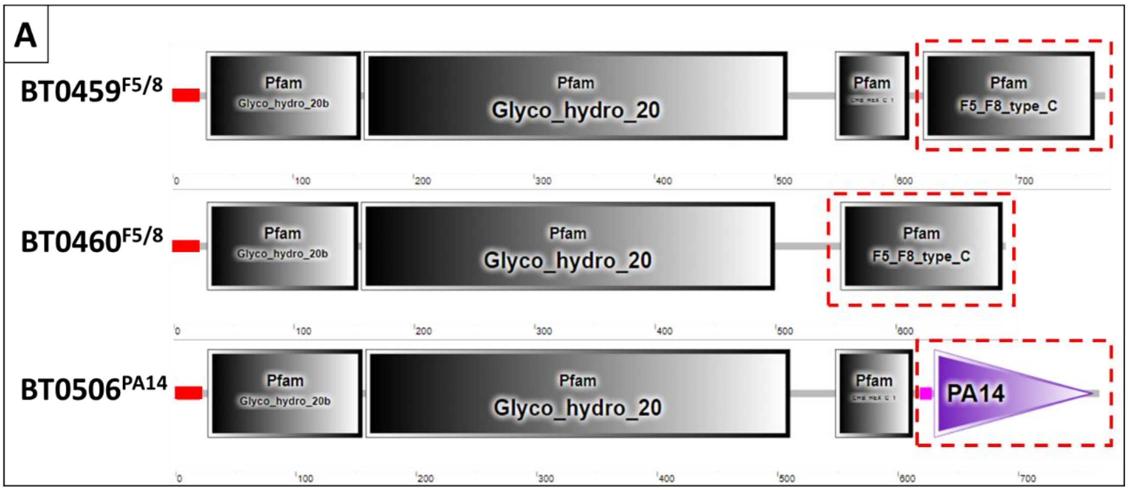


Figure 4.51: The predicted BT0459^{GH20} binding model. A) The open conformation of BT0459^{GH20}. **B)** CBM domain (red) starts to move towards the active site. **C)** The predicted closed form of BT0459^{GH20}. CBM domain fully covers the active site, which would potentially help to anchor the bound glycans in place. X-ray structure of BT0459^{GH20} was coloured rainbow from red (C-terminus) to blue (N-terminus). N-glycan model is coloured in blue.

4.3.5.6. Investigating the CBM domain involvement in N-glycan binding

4.3.5.6.1. Generating the predicted CBM mutants of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}

The F5_F8_type_C domain of BT0459^{GH20} was previously shown to be structurally related to the CBM32 (**Figure 4.50**). Thus, in this study we wanted to investigate the C-terminal domains of the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} proteins in more detail and determine whether they are carbohydrate-binding domains that could be involved in the N-glycan binding. BT0456^{GH20} was not investigated because it did not have a defined C-terminal domain sequence. The F5/8_type_C domain of BT0459^{GH20} only shares 26% sequence similarity to the F5/8_type_C domain of BT0460^{GH20} and 25% sequence similarity to the PA14 domain of BT0506^{GH20} whereas F5/8_type_C domain of BT0460^{GH20} is 22% similar to the PA14 domain of BT0506^{GH20} (**Figure 4.29**). Although F5/8_type_C domain appears to be related to the PA14 domain, these results suggest that the domains present in these three GH20s may have very different roles. The cloned C-terminal domains were labelled BT0459$^{F5/8}$, BT0460$^{F5/8}$ and BT0506^{PA14}, analysed using the SDS-PAGE and compared to the full length proteins. The generated mutants expressed well (**Figure 4.52**).



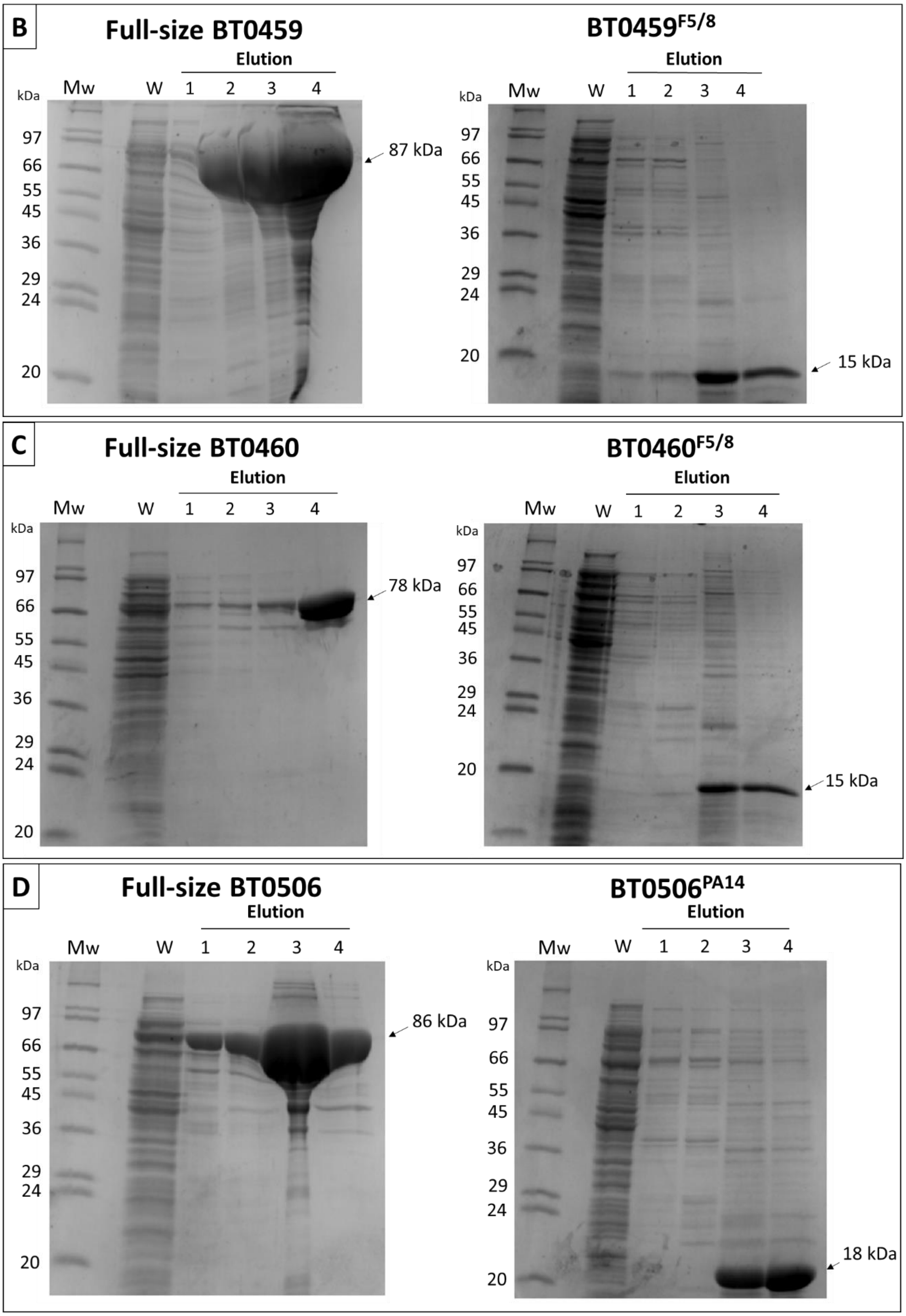
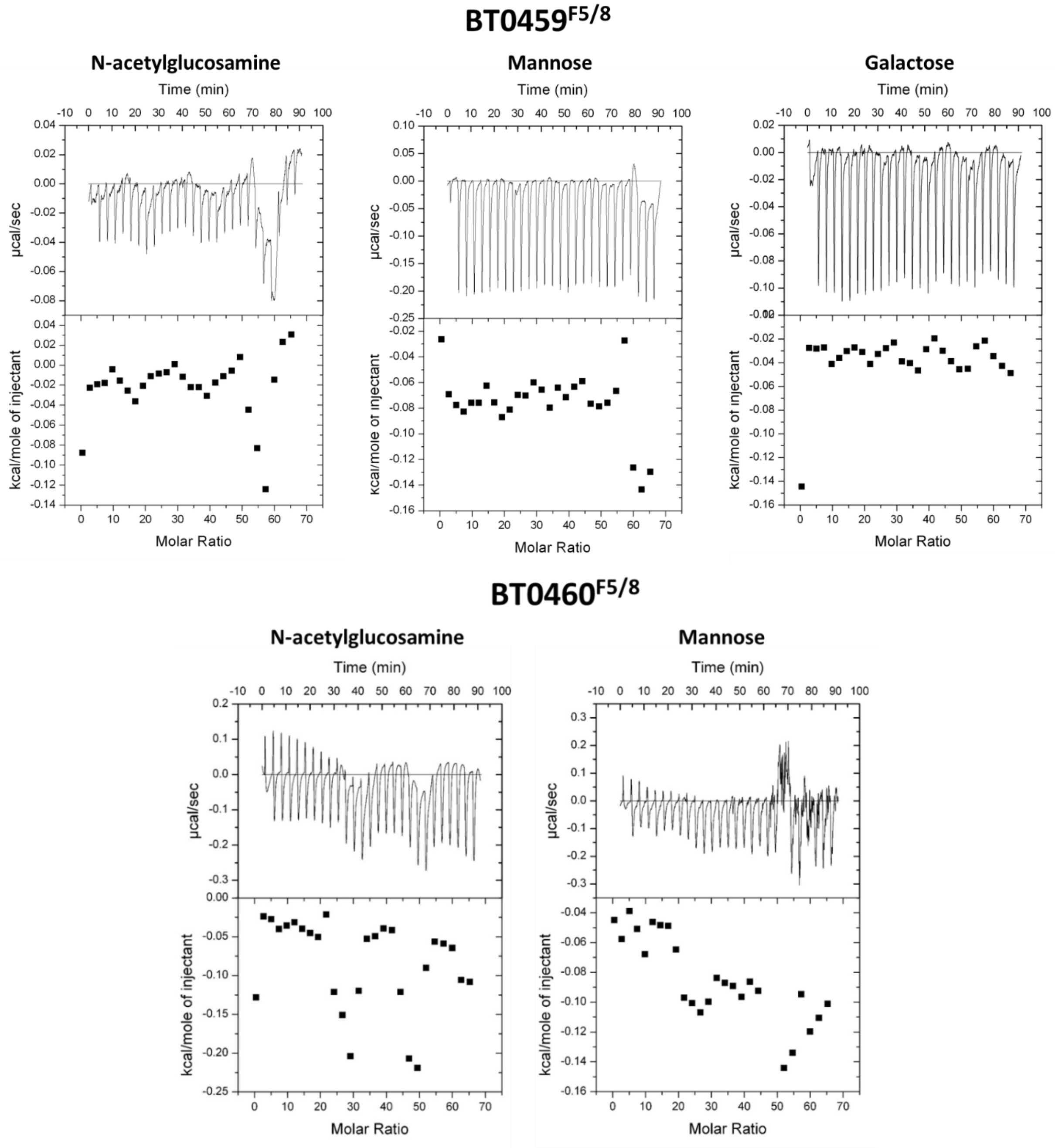


Figure 4.52: Expression of the C-terminal domains of the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. **A)** The domain composition of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. The region of each protein encoding the predicted C-terminal CBM is highlighted in red. **B)** Purification of the BT0459$^{F5/8}$. **C)** Purification of the BT0460$^{F5/8}$. **D)** Purification of the BT0506^{pA14}. Mw is a wide-range molecular weight marker (Sigma). Full-size protein purification is also displayed in Figure 4.2.

4.3.5.6.2. Biochemical characterization of the GH20 C-terminal domains

The key role of CBM domains is binding to the substrates. The binding affinities of the recombinant C-terminal domains from BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} were initially investigated using the isothermal titration calorimetry (ITC; Section 2.31). Galactose (Gal), N-acetylglucosamine (GlcNAc) and mannose (Man) were tested as they are major structural components of complex N-glycans but no binding was observed (**Figure 4.53**).



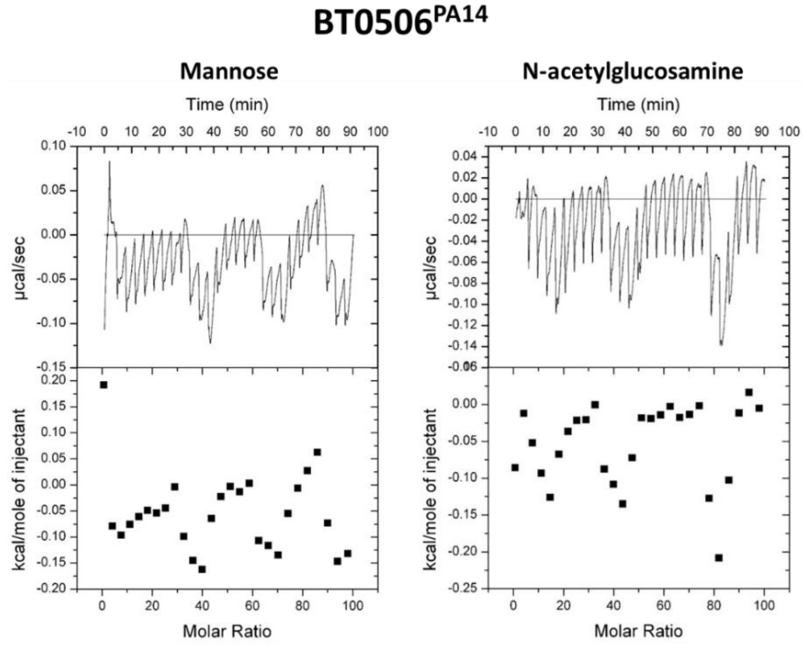


Figure 4.53: ITC analysis of monosaccharide binding by the putative CBMs of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. For these experiments, 10 mM of each of the monosaccharides and 30 µM of purified recombinant putative CBMs were used.

Because ITC requires very high concentrations of glycans that can get prohibitively expensive to test, it was decided to investigate and quantify the bio-molecular interactions between the putative carbohydrate-binding modules of GH20-family β-hexosaminidases and complex N-glycans using the Microscale Thermophoresis (MST) (Section 2.34). The N-glycans used for MST were derived from bovine α_1-AGP using the BT1044^{GH18}, BT0455^{GH33} and BT0461^{GH2} treatment following a protocol described in Section 2.3.7. The MST studies on the purified amine labelled BT0459$^{F5/8}$, BT0460$^{F5/8}$ and BT0506^{PA14} revealed that all three displayed affinities for complex N-glycan structures over simple mono- or di-saccharides (**Figure 4.54**), supporting the prediction that these domains are carbohydrate binding modules. However, the binding observed in all cases was in a relatively low affinity range. The binding model and K_d values could only be estimated for carbohydrate-binding modules of BT0459 (K_d=804.4 ± 175.1) and BT0460 (K_d=554.9 ± 123.4). These results suggest that these novel CBM domains of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} are involved in N-glycan binding.

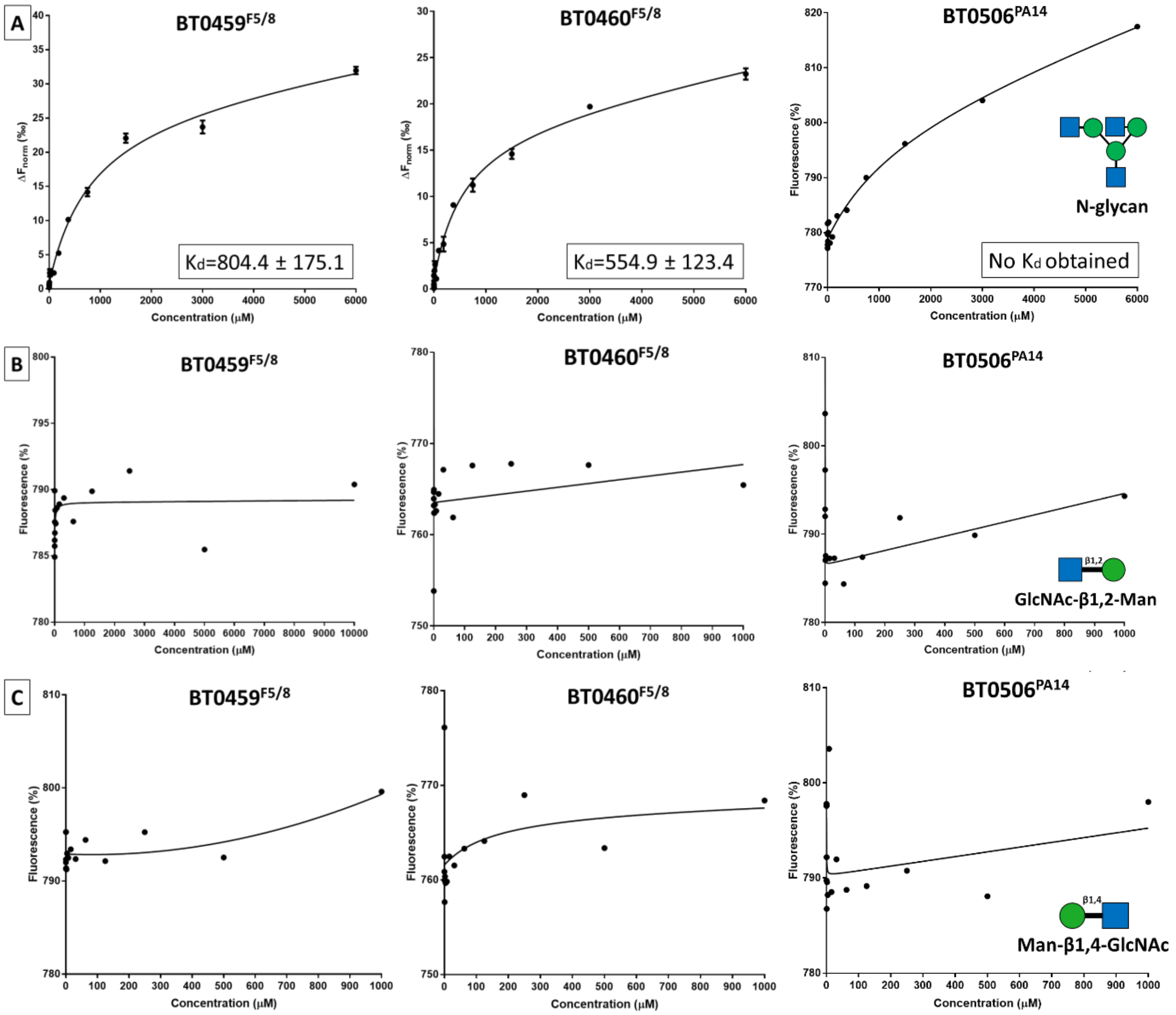


Figure 4.54: MST analyses of carbohydrate binding by recombinant CBM domains of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. 100 nM of fluorescently labelled proteins was used in this assay. **A)** CBM binding curves to N-glycans (6mM). K$_d$ of BT0459 CBM is 804.4 ± 175.1 whereas K$_d$ of BT0460 CBM is 554.9 ± 123.4. **B)** CBM binding to GlcNAc-β1,2-Man (10mM). No binding was observed. **C)** CBM binding to Man-β1,4-GlcNAc (10mM). No binding was observed. The experiments were repeated three times to ensure the data was consistent.

4.3.5.6.3. Generating the BT0459ΔCBM mutant

The solving of the BT0459^{GH20} crystal structure provided insight into how this enzyme could work. It was previously shown that C-terminal domain (F5/8 type_C) is indeed a carbohydrate-binding domain (**Figure 4.54**). In this study, we wanted to further investigate the involvement of the CBM32-like domain of BT0459^{GH20} in the N-glycan degradation and to determine what effect removing this domain would have on the enzyme activity and substrate specificity. A truncated form of BT0459^{GH20} lacking the CBM domain, BT0459ΔCBM, was generated using the data obtained from the crystal structure (**Figure 4.50**) and a protocol described in Section 2.1.13. Mutant was expressed and purified using the protocols described in Sections 2.1.20 and 2.1.21 (**Figure 4.55**). A generated BT0459ΔCBM mutant lacks the F5_F8_type_C (Pfam) domain and is predicted to be of 66 kDa in size.

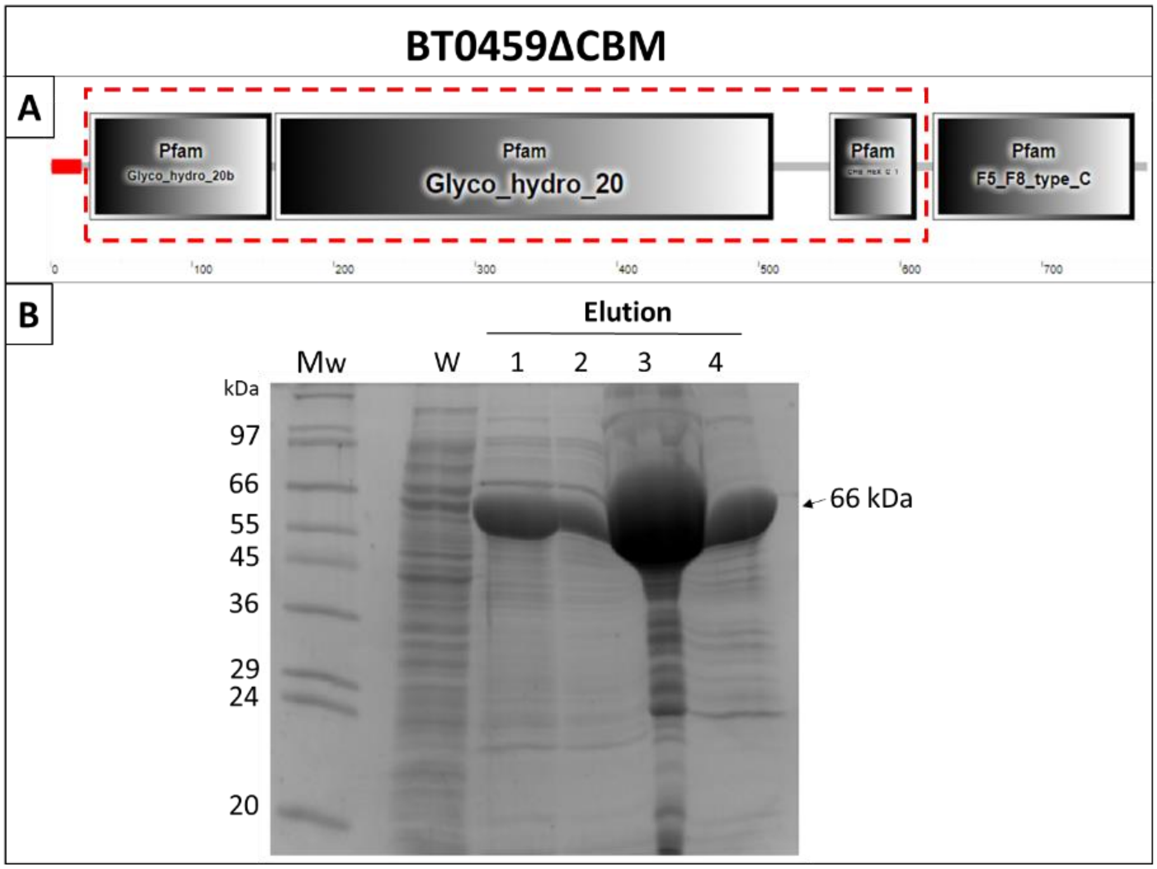


Figure 4.55: Generating a BT0459ΔCBM mutant. A) The domain composition of BT0459^{GH20}. The truncated BT0459ΔCBM mutant lacks the F5_F8_type_C domain. **B)** SDS-PAGE gel of BT0459ΔCBM purification. The predicted size of the mutant is ~66 kDa. Mw is a wide-range molecular weight marker (Sigma-Aldrich).

4.3.5.6.4. Investigating the enzyme activity of the BT0459ΔCBM

To initially investigate the activity of the BT0459ΔCBM, colorimetric assays against pNP-β-GlcNAc and pNP-β-GalNAc substrates were set up and incubated for 30 minutes at 37 °C. The positive results were observed via the change in assay colour to yellow and showed that BT0459ΔCBM mutant retained its activity against these substrates. To investigate BT0459ΔCBM activity further, enzyme assays against the simple disaccharide structures that BT0459^{GH20} was previously shown to be active against (**see Figure 4.31**) were set up and incubated for 1h at 37 °C (Section 2.3.5). Based on the release of monosaccharides, the results revealed no notable differences in BT0459ΔCBM activity compared to the full-length BT0459^{GH20} against these substrates (**Figure 4.56**).

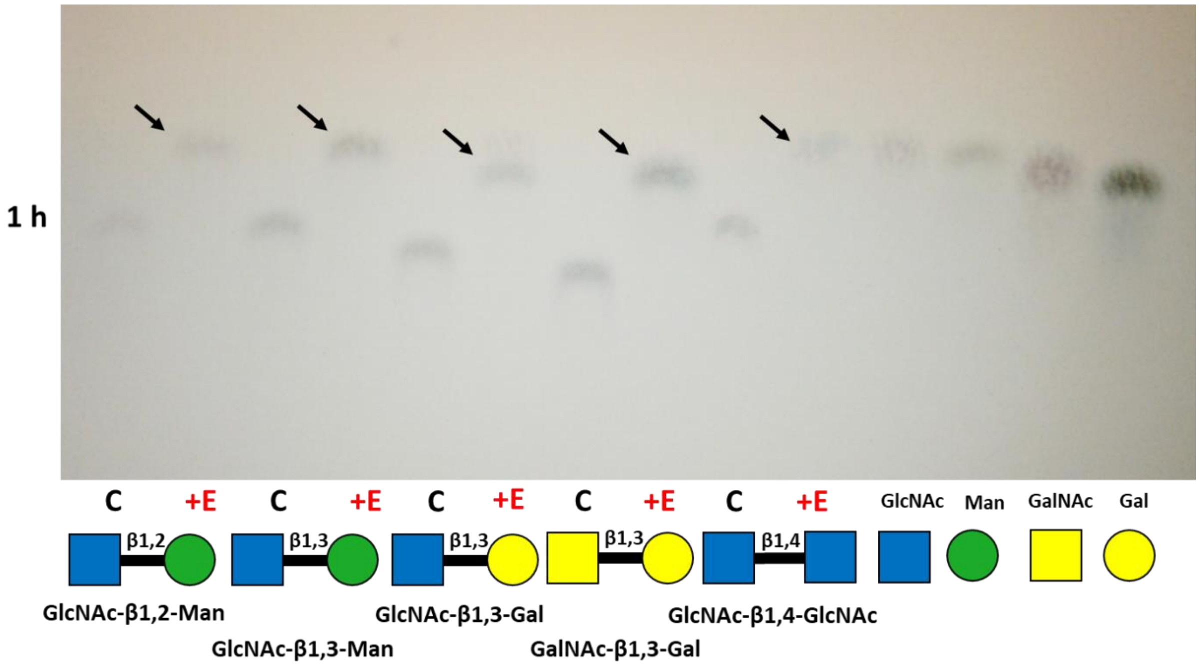


Figure 4.56: Screening for BT0459ΔCBM activity against simple disaccharides. 1µM of enzyme was incubated with 1mM of disaccharides. C – disaccharide control; +E– substrate incubated with the enzyme. 1mM of monosaccharide standards were used. Thin-layer chromatography (TLC) plate was stained using the DPA stain. Monosaccharide products are indicated by arrows.

No notable difference in BT0459ΔCBM activity compared to the BT0459^{GH20} was observed when it was incubated with chito-oligosaccharides, suggesting the CBM32-like domain may not be required for the BT0459^{GH20} activity against these substrates (**Figure 4.57**).

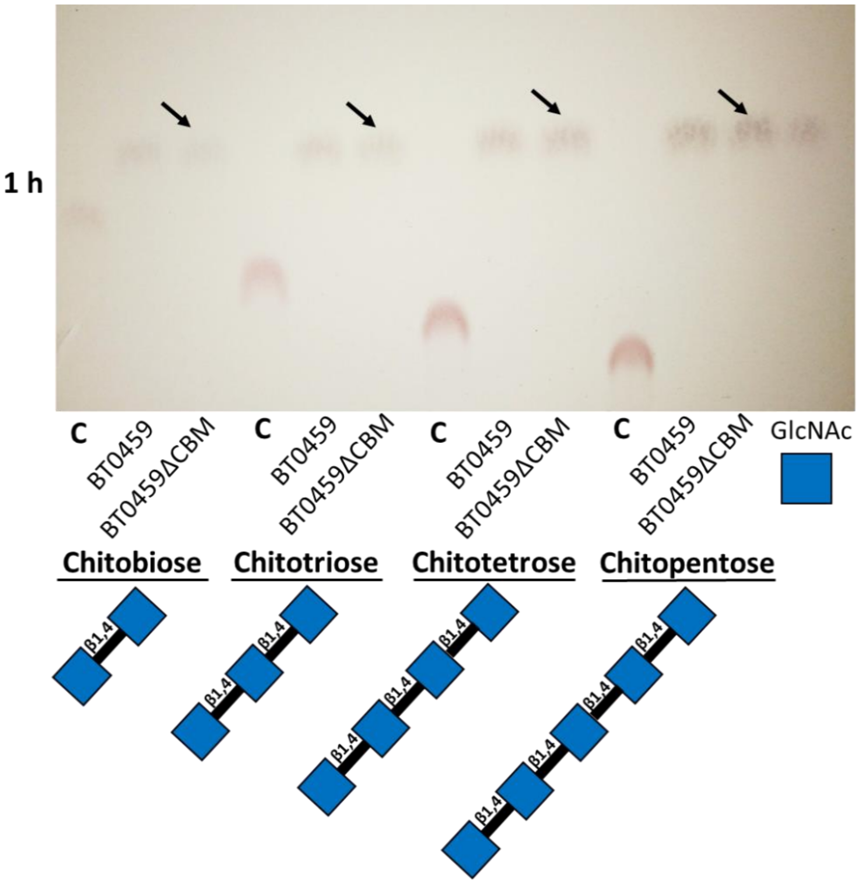
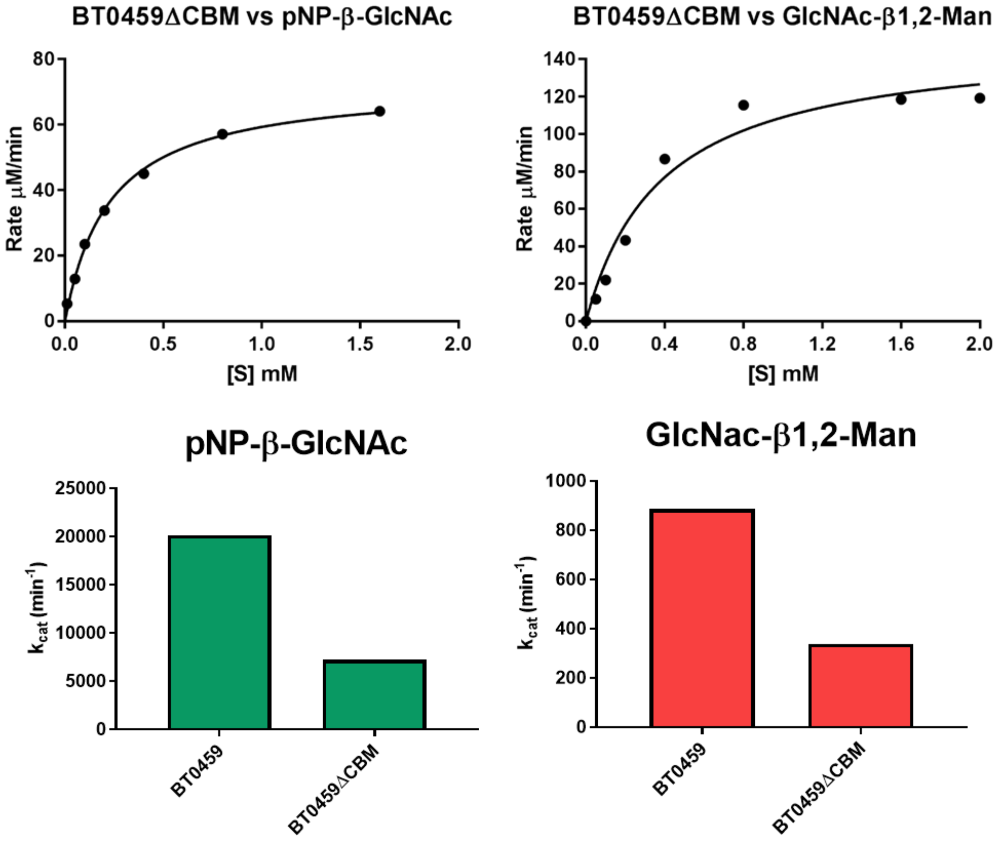


Figure 4.57: Comparing the BT0459ΔCBM activity versus BT0459^{GH20} against chito-oligosaccharides. 1µM of enzyme was incubated with 1mM of chito-oligosaccharides. C – substrate control. 1mM of GlcNAc standard was used. Thin-layer chromatography (TLC) plate was stained using the DPA stain. Monosaccharide products are indicated by arrows.

To investigate the differences in the BT0459^{GH20} and BT0459ΔCBM catalytic efficiency, kinetic parameters were measured and compared (Section 2.3.5.1) The kinetic parameters and respective kinetic curves for each substrate are shown in **Figure 4.58**. These results suggest that BT0459^{GH20} is more catalytically efficient than BT0459ΔCBM. Based on the k_{cat}/K_M values, BT0459^{GH20} is 1.2-fold more efficient in catalysing the hydrolysis of GlcNAc-β1,2-Man than BT0459ΔCBM. The same pattern was observed in activity against pNP-β-GlcNAc substrate, suggesting the CBM domain may be involved in enhancing of the catalytic activity of the BT0459.



BT0459ΔCBM	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
	pNP-β-GlcNAc	7270	0.23±0.009	31608
	GlcNAc-β1,2-Man	338	0.38±0.05	889

BT0459	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
	pNP-β-GlcNAc	20185	0.48±0.03	42052
	GlcNAc-β1,2-Man	888	0.78±0.053	1138

Figure 4.58: Kinetic parameters of BT0459ΔCBM mutant. The reactions were performed at 37 °C in 20 mM NaH_2PO_4 (pH 7.5). Rates were measured at 320 nm for pNP-linked substrates and 340 nm for GlcNAc-β1,2-Man. The data was plotted in GraphPad Prism 7.0 software using non-regression analysis (see Section 2.3.5). The standard errors were generated from technical triplicates.

To investigate the differences in BT0459^{GH20} and BT0459ΔCBM activity against complex N-glycans, the enzyme assays against α_1-AGP pre-treated with the BT1044^{GH18}, BT0455^{GH18} and BT0461^{GH2} were set up by incubating 1µM of enzymes with 10mg/ml of N-glycan mixture for 20 minutes at 37 °C. The products were visualised using the HPAEC-PAD following a protocol described in Section 2.3.3. After quantifying the relative peak area (nC*min) of the tetrasaccharide product, the results show that there was no difference in complex N-glycan degradation pattern between the BT0459^{GH20} and BT0459ΔCBM mutant, suggesting CBM domain may not be required for α_1-AGP complex bi-antennary N-glycan degradation (**Figure 4.59**). The experiment was repeated three times using freshly purified enzymes and different incubation timescales (0-16h) (**Figure 4.60**). Based on the appearance of a peak representative of a tetrasaccharide N-glycan structure, same pattern of N-glycan degradation was observed.

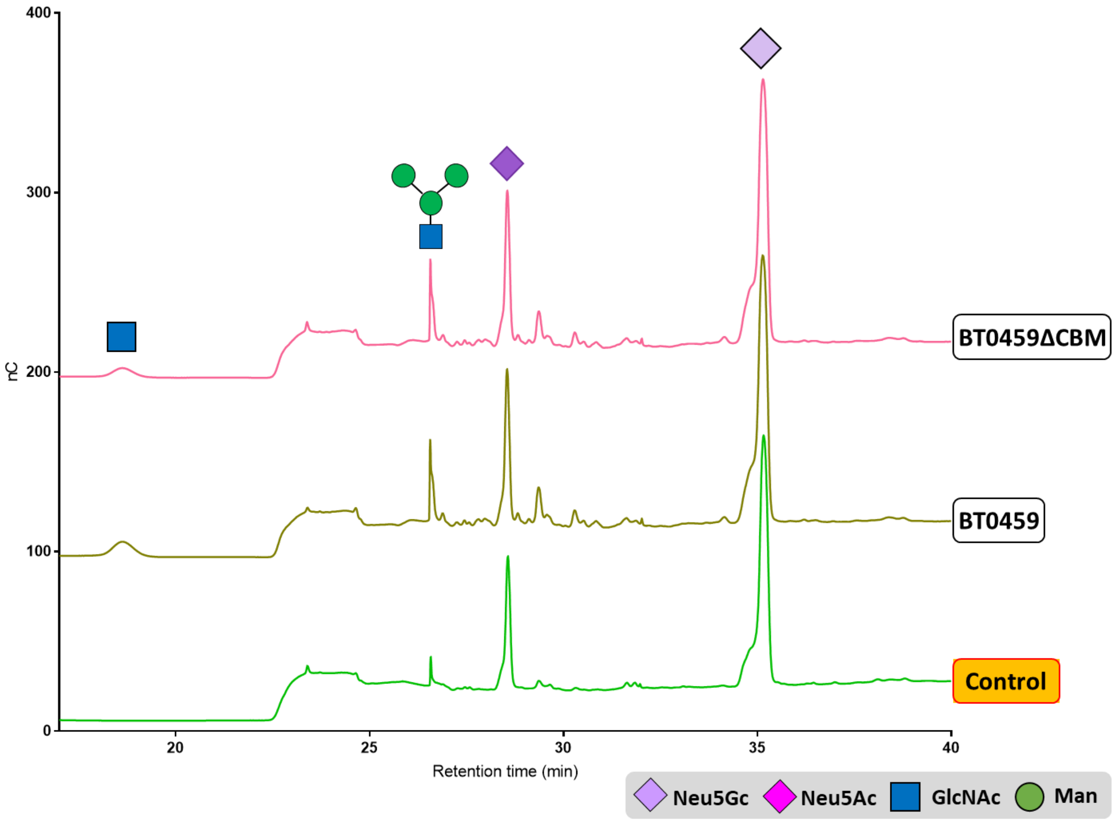


Figure 4.59: Investigating the differences in N-Glycan (α_1-AGP) degradation by BT0459^{GH20} and BT0459ΔCBM. 10mg/ml of N-glycoprotein was incubated with 1µM of enzymes in 20mM sodium phosphate (pH 7) for 1 h at 37°C. 1:10 dilution was loaded onto the machine. The concentration of monosaccharide standards used is 1mM.

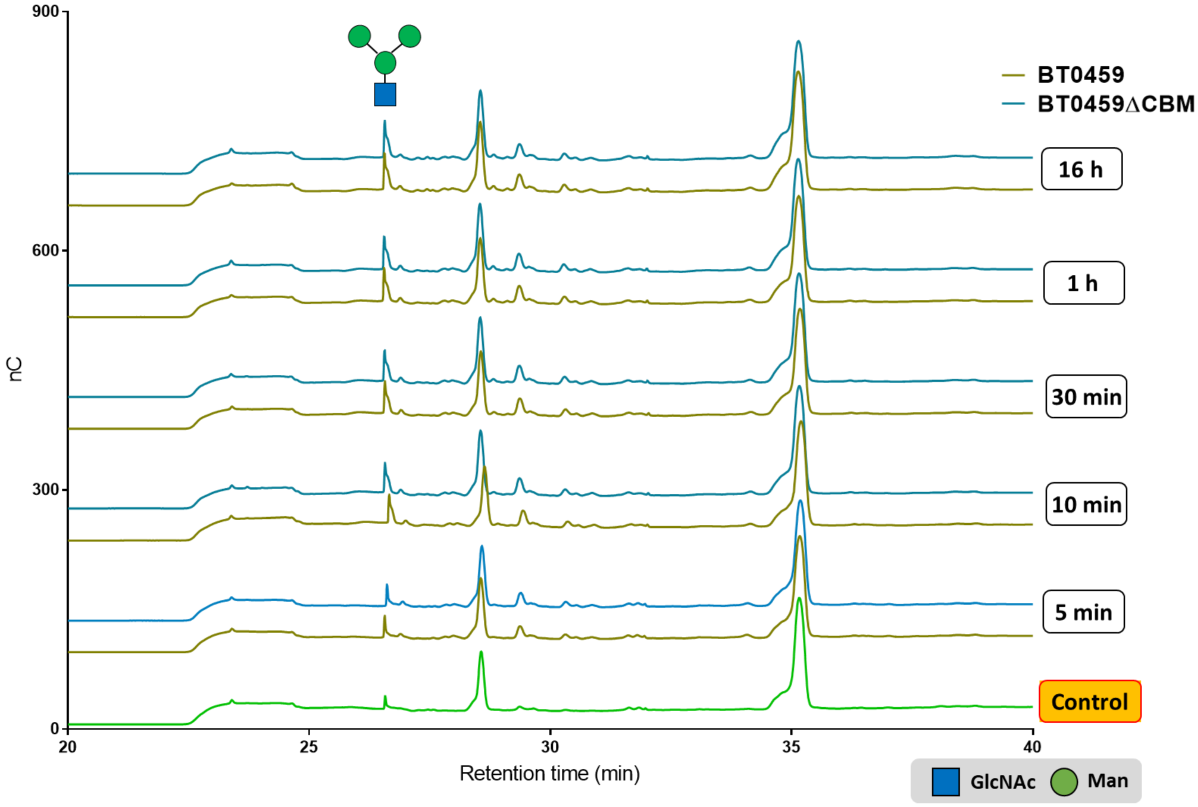


Figure 4.60: Time course of N-Glycan (α_1-AGP) degradation by BT0459^{GH20} and BT0459ΔCBM. 10mg/ml of N-glycoprotein was incubated with 1µM of enzymes in 20mM sodium phosphate (pH 7) at 37°C. 1:10 dilution was loaded onto the machine. The concentration of monosaccharide standards used is 1mM.

4.3.6. BT0458^{GH2}: investigating a mannosidase activity

4.3.6.1. Introduction

D-Mannose is a simple hexose sugar that forms a six-carbon ring. It is a 2-epimer of glucose that is primarily found as a sweet-flavoured α- or bitter-flavoured β-linked anomer of the pyranose. Mannose is widely spread in nature – it is often present as a structural component of carbohydrates found in plant cell walls, fungi, bacteria and mammals (Sharma *et al.,* 2014).

Mannosidases are a large group of enzymes that are specialised in catalysing the hydrolysis of the glycosidic mannose linkages found in complex carbohydrates. They are grouped into α-mannosidases and β-mannosidases. Majority of β-mannosidases are classified into the glycoside hydrolase (GH) family number 2 (GH2) of the CAZy classification, along with β-galactosidases and β-glucuronidases (www.cazy.org). To date, it is known that *B. thetaiotaomicron* produces 33 GH2-family enzymes, 5 of which are annotated as β-mannosidases. Like the rest of GH2 family of enzymes, β-mannosidases are retaining enzymes that follow a double-displacement Koshland mechanism (CAZypedia.org, GH2 family).

In gastrointestinal tract (GI), mannose can be found in complex carbohydrate structures decorating glycoconjugates, such as N-glycoproteins. On N-glycoproteins, mannose residues usually form α3/6- and β1,4- glycosidic linkages. *B. thetaiotaomicron*, a dominant member of human gut microbiota, possesses a β1,4-mannosidase BT0458^{GH2} (BtMan2A) that has been previously characterised by Dr. Louise Tailford. It is a five-domain β-mannosidase that is structurally similar to *E. coli* LacZ β-galactosidase (Tailford *et al.,* 2007). Although BT0458^{GH2} has been structurally and biochemically characterised, it was not investigated in the context of N-glycan degradation. As previously discussed (**Figure 3.7.1**), BT0458^{GH2} is upregulated when *B. thetaiotaomicron* is grown on N-glycoproteins as sole carbon sources, compared to the glucose control. Therefore, the aim of this study was to investigate the BT0458^{GH2} involvement in N-glycan degradation.

4.3.6.2. Bioinformatics analysis of BT0458^{GH2}

To investigate the similarities of BT0458^{GH2} to other known β-mannosidases, the BT0458^{GH2} amino acid sequence was used as a query for KEGG database search to identify the orthologous proteins. Sequence similarity database (SSDB) tool hosted by Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to generate a dendrogram of top 50 best protein matches to BT0458^{GH2} (**Figure 4.61**). The results show that BT0458^{GH2} type of GH2 β-mannosidase is highly conserved in *Bacteroides* and has closely-related orthologues in numerous members of gut microbiota, including β-mannosidase from *B. fragilis* that was previously shown to be involved in complex N-glycan degradation (Cao *et al.,* 2014)

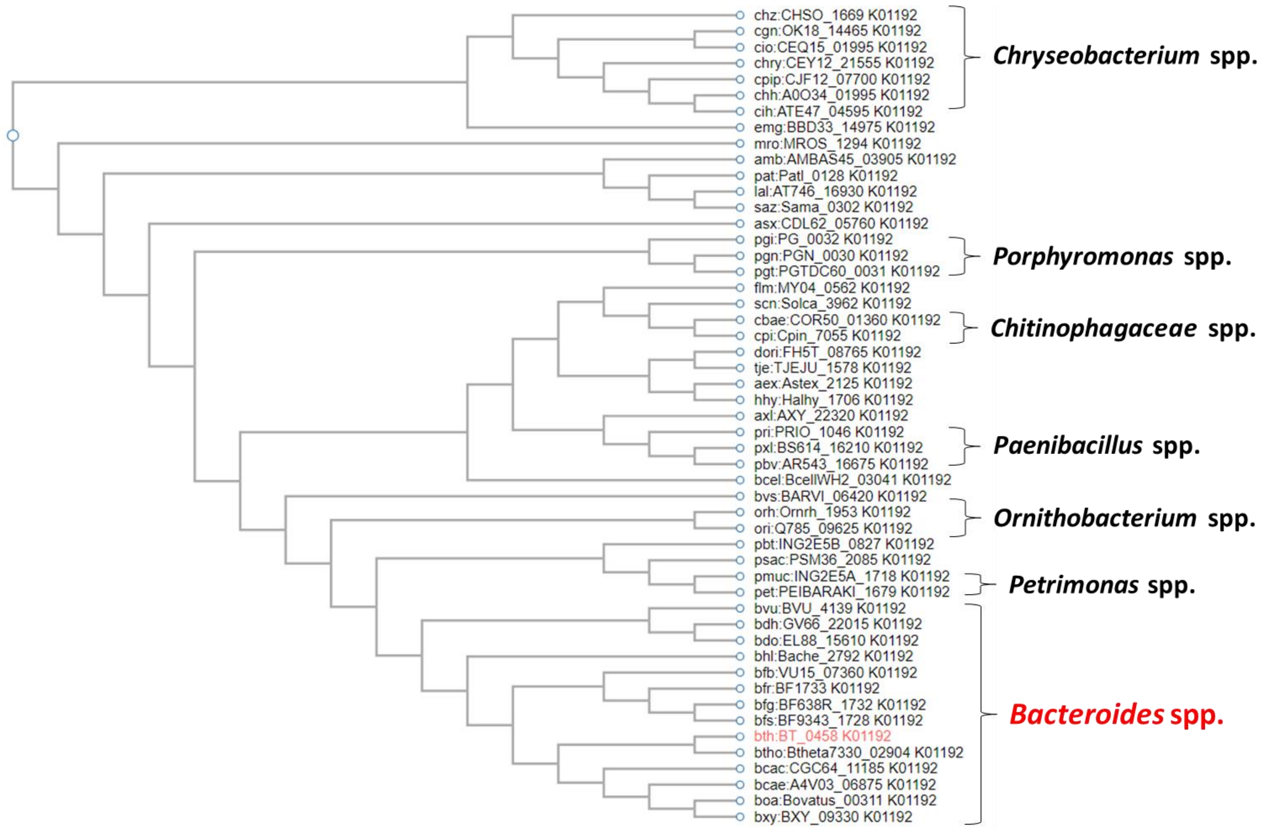


Figure 4.61: A dendrogram showing the taxonomic relationship of BT0458^{GH2} to the 50 best orthologous β-mannosidases. KEGG SSDB tool was used to perform a sequence similarity search on BT0461. The results were visualised by producing a dendrogram of best orthologues to BT0458^{GH2} using a Smith-Waterman similarity score was used to identify the best hits.

Using the dendrogram displayed in **Figure 4.61**, β-mannosidases from *B. vulgatus* (BVU_4138), *B. fragilis* (BF_1728), *B.ovatus* (Bovatus_00311) and *B. xylanisolvens* (BXY_09330) were selected for the comparison due to their close taxonomic relationship to BT0458^{GH2}. The sequences were aligned using the Clustal Omega Multiple Sequence Alignment tool and visualised using the ESPript 3.0 tool. Based on the fully conserved amino acid residues (highlighted in red), the results show that BT0458^{GH2} amino acid sequence is closely related to these β-mannosidases (**Figure 4.62**). The similarity matrix obtained from Clustal Omega alignment shows that BT0458^{GH2} is 80% identical to Bovatus_00311 and BXY_09330; 75% similar to BF_1728 and 64% related to BVU_4138. These results suggest that these mannosidases could have similar enzymatic activities and may target N-glycan structures. This theory is reinforced by the fact that all four enzymes share the same active site residues, that have been previously identified in BT0458^{GH2} by Dr. Louise Tailford (Tailford *et al.,* 2007). These are Trp395, Asn461, Glu462, Try537, Glu555 and Trp645 in BT0458^{GH2}.

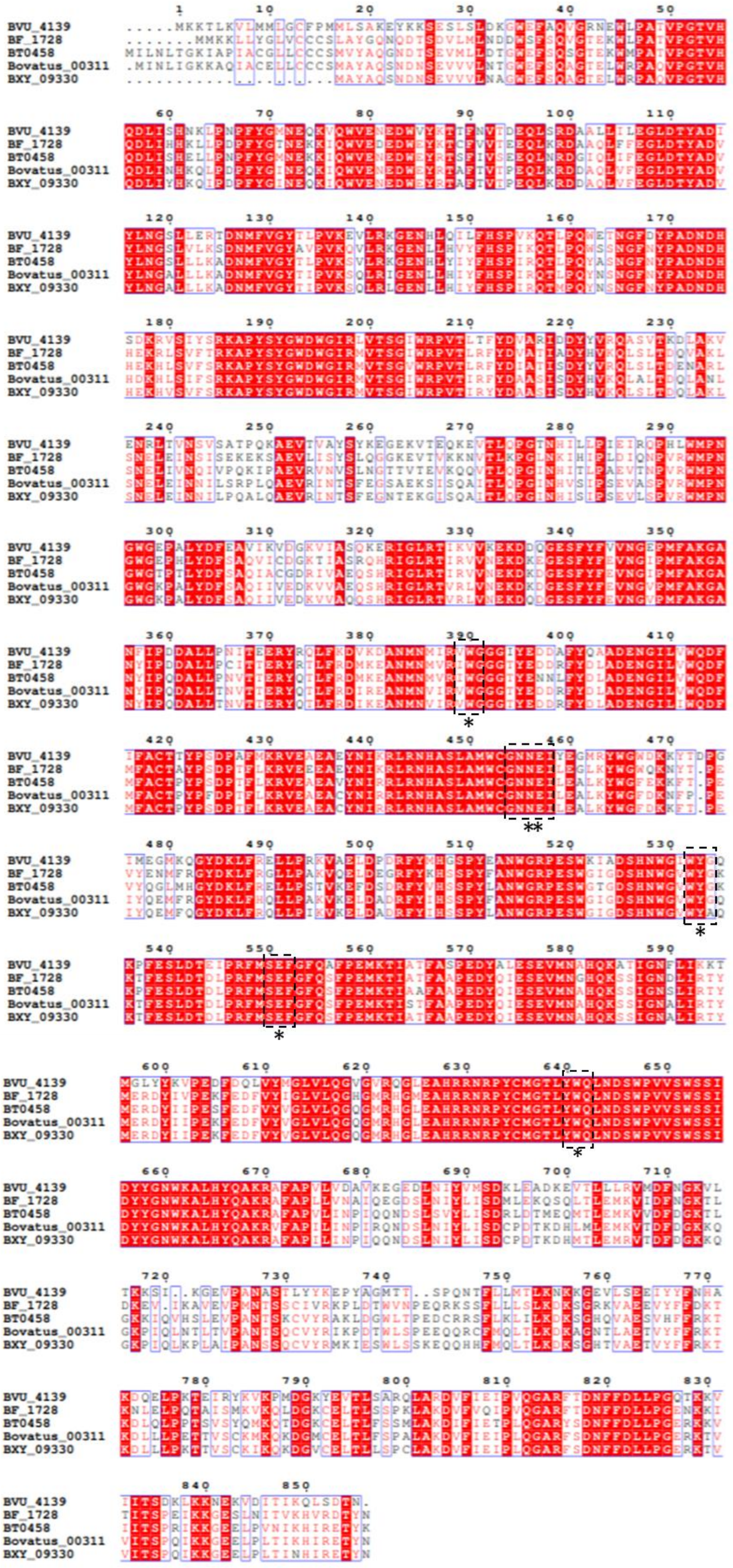


Figure 4.62: Amino sequence alignments of *B. thetaiotaomicron* β-mannosidase BT0458^{GH2} to the orthologous *Bacteroides spp*. β-mannosidases. Amino acids highlighted in red are fully conserved. Conserved active site residues are highlighted in black boxes and labelled with an asterisk (*).

4.3.5.3. Confirming BT0458^{GH2} substrate specificity

To initially screen BT0458^{GH2} enzyme activity and confirm the results obtained by Dr. Louise Tailford (Tailford *et al.,* 2007), colorimetric assays against the para-Nitrophenol (pNP)-linked substrates were set up as described in Section 2.3.5.1 using 1 mM of pNP-substrate and incubation for 30 min at 37 °C . Consistent with the results obtained by Tailford *et al.* (2007), the BT0458^{GH2} activity was only detected against p-nitrophenyl-β-mannopyranoside (pNP-β-Man). No activity was detected against pNP-α-Man nor pNP-β-Gal, confirming BT0458^{GH2} is a β-mannosidase. To investigate specific glycan structures β-mannosidase BT0458^{GH2} targets and investigate the potential differences compared to the activity of BT1033^{GH130}, a β-manno-phosphorylase that was shown to be upregulated in RNA-seq data when *B. thetaiotaomicron* was grown on N-glycoproteins (**Figure 3.7.1**) , enzyme assays against various disaccharide substrates were set up following a protocol described in Section 2.3.5. The assays were incubated in sodium phosphate 20mM (pH 7) for 16 hours at 37 °C, aerobically and visualised using the TLC following a protocol described in Section 2.3.2. Based on the appearance of the bands representative of monosaccharide products, the results showed that β-mannosidase BT0458^{GH2} and β-manno-phosphorylase BT1033^{GH130} are active against β1,4-mannobiose and Man-β1,4-GlcNAc (**Figure 4.63**). Neither BT0458^{GH2} neither BT1033^{GH130} showed any activity against Man-β1,2-GlcNAc, suggesting they both have a β1,4-linkage specificity. BT1033^{GH130} catalyses the phosphorolysis of mannose disaccharides, thus it produces mannosyl-phosphate products (*, Dr. Lucy Crouch, unpublished data).

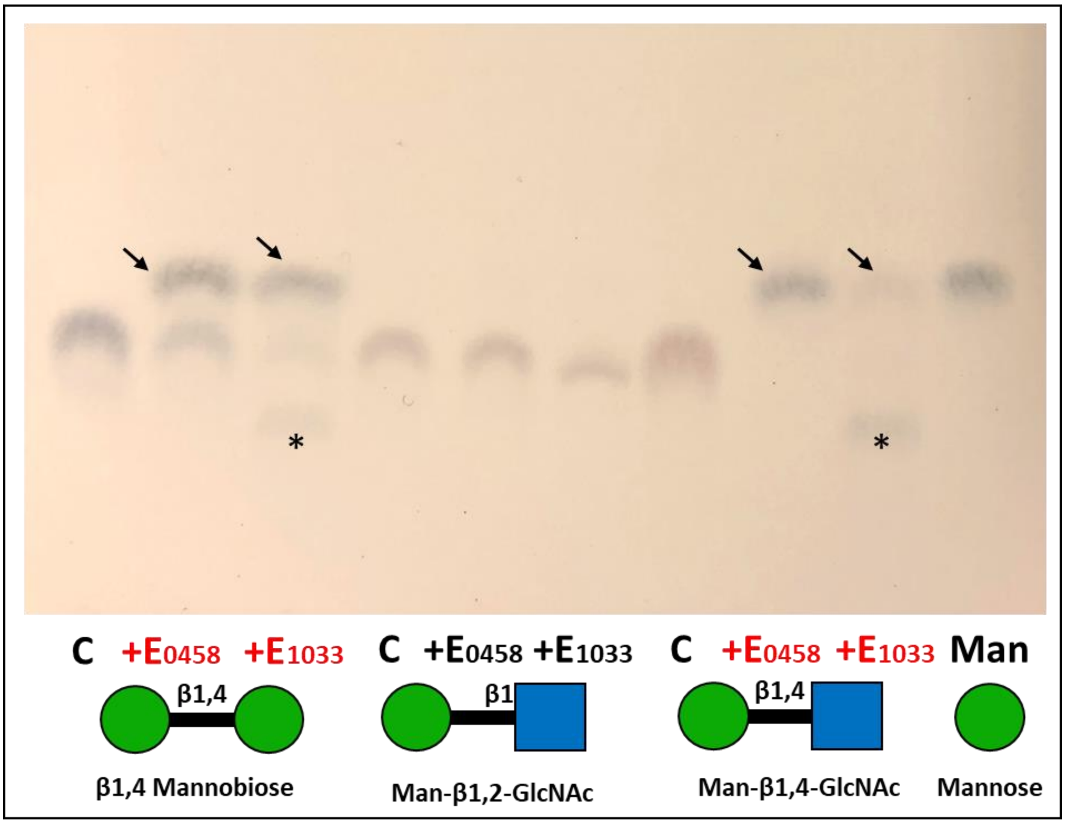
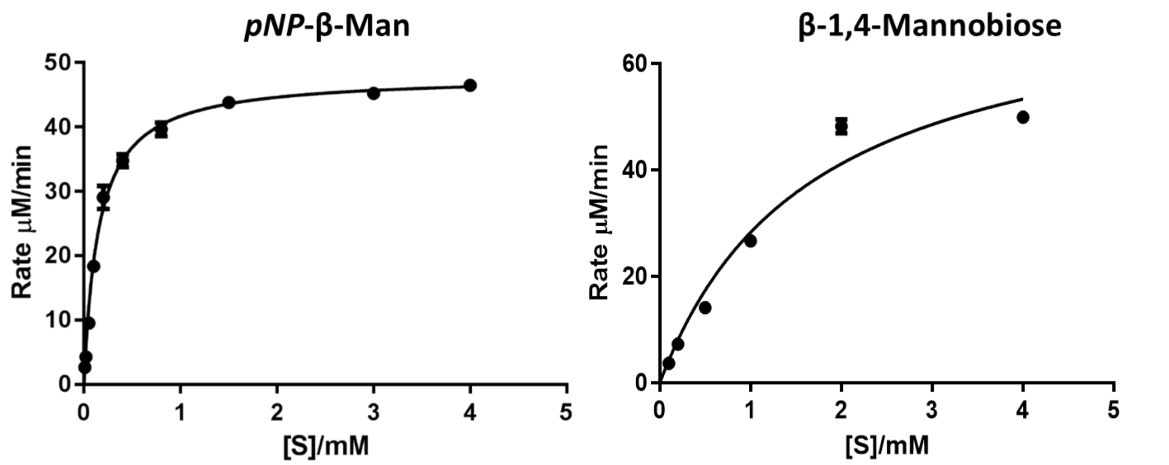


Figure 4.63: Screening for the enzyme activity of BT0458^{GH2} and BT1033^{GH130} against simple disaccharides. 1µM of BT0458^{GH2} and BT1033^{GH130} were incubated separately with 1mM of disaccharides. C – disaccharide control; +E$_{0458}$ – substrate incubated with BT0458^{GH2}. +E$_{1033}$ – substrate incubated with BT1033^{GH130}. Active enzymes are highlighted in red. 1mM of mannose standard was used. The bands labelled with an arrow show mannose release whereas the asterisks (*) show mannosyl-phosphate products. TLC plate was stained using the DPA stain.

To investigate the catalytic efficiency of the β-mannosidase BT0458^{GH2} and compare to the results obtained by Dr. Louise Tailford (Tailford *et al.,* 2007), the kinetic parameters of the enzyme against β1,4-Mannobiose was measured at 340nm using D-Mannose detection kit (Megazyme) following the manufacturer's instructions (Section 2.3.5.2). Meanwhile, the BT0458^{GH2} activity against pNP-β-Man was measured at 320nm following a protocol described in Section 2.3.5.1. The kinetic parameters and respective kinetic curves for each disaccharide are shown in **Figure 4.64**. The difference in catalytic preference is displayed in variation in the k_{cat} and K_M values. Based on the k_{cat} and K_M values the results suggest that β1,4-mannobiose is not a preferred substrate for the BT0458^{GH2} β-mannosidase. Unfortunately, the enzyme kinetics experiments against Man-β1,4-GlcNAc could not be performed due to commercial unavailability of large quantities of this substrate. However, the kinetic properties obtained by Tailford *et al.* (2007) show that β-mannosidase BT0458^{GH2} has a preference for Man-β1,4-GlcNAc over β1,4-Mannobiose. Man-β1,4-GlcNAc is a core structure found in all N-glycans, suggesting that this enzyme may have preference for it.



	Substrate	k_{cat} (min^{-1})	K_m (mM)	K_{cat}/K_m
BT0458	pNP-β-Man	4801	0.15±0.0072	32000
	Man-β1,4-Man	13	1.7±0.3	8

Figure 4.64: Kinetic properties of BT0458^{GH2}. The enzyme reactions were performed at 37 °C with 20 mM NaH_2PO_4 (pH 7.5). Rates were measured at 320 nm for pNP-Gal and 340 nm for the β1,4-Galactobiose. The results were plotted in GraphPad Prism 5.0 software using non-regression analysis (see Section 2.3.5). The standard errors were generated from biological triplicates.

4.4. Discussion

In this chapter, the enzymes involved in complex N-glycan utilization by *B. thetaiotaomicron* were characterised and their biological roles in N-glycan degradation were identified.

Seven enzymes encoded by the B.theta$^{0455\text{-}0461}$ locus and BT0506^{GH20} were upregulated when the wild type *B. thetaiotaomicron* was grown on N-glycoproteins as sole carbon sources, compared to the glucose control (**Figure 3.7.1**). These genes include the sialidase BT0455^{GH33}, four β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}, β-mannosidase BT0458^{GH2} and β-galactosidase BT0461^{GH2}. Very high basal expression levels of these genes suggests they are important for the N-glycan metabolism and proliferation of the *B. thetaiotaomicron*.

The BT0455^{GH33} is the only known sialidase produced by *B. thetaiotaomicron.* Investigation of its activities in the context of N-glycan degradation *in vitro* has identified it to be the key enzyme required for the removal of terminal sialic acid caps in order for other enzymes to access their carbohydrate targets. It is a broad-acting sialidase that can catalyse the hydrolysis of any terminal sialic acid linkages, such as α2-3/6/8-, present on both human and bovine N-glycoproteins (NeuAc and NeuGc) (**Figure 4.6; Figure 4.7; Figure 4.8**). These results suggested that the sialylated N-glycan degradation by *B. thetaiotaomicron* could be a sequential process. Interestingly, *B. thetaiotaomicron* cleaves the sialic acid caps but does not possess a genetic machinery to utilize this monosaccharide, releasing it into the environment where it can be digested by other specialised bacteria.

The BT0457^{CE} is a sialic acid esterase that works in conjunction with the sialidase BT0455^{GH33} (**Figure 4.14**). It was found to be 60% homologous to the fully characterized sialic acid-specific 9-O-acetylesterase from *T. forsythia* (**Figure 4.10**). Because *B. thetaiotaomicron* does not utilise the sialic acid as a substrate, the precise role of BT0457^{CE} remains to determined. However, it may be required to boost the efficiency of the sialic acid removal from the N-glycans to uncap the underlying carbohydrate structures.

The BT0461^{GH2} β-galactosidase activity has been screened against a range of disaccharides, oligosaccharides and de-sialylated N-glycan structures (**Figures 4.17-4.22**). It has a Gal-β1,4- linkage specificity and shows a kinetic preference for Gal-β1,4-GlcNAc disaccharide, a structural component of complex N-glycan structures (**Figure 4.18**). Although BT0461^{GH2} could catalyse the hydrolysis of all β1,4-linked substrates tested, combined, these results suggest that Gal-β1,4-GlcNAc-linkages found on N-glycans are its preferential target. The requirement for the de-sialylation of N-glycan substrates by the sialidase BT0455^{GH33} before BT0461^{GH2} could cleave the galactose caps confirmed that BT0461^{GH2} is an exo-glycosidase and further reinforced the claim that N-glycan degradation by *B. thetaiotaomicron* is a sequential process.

Four N-acetyl-β-hexosaminidases were characterised in this chapter: BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. It was a challenging task to investigate their substrate specificities. Despite having quite different amino acid sequences (**Figure 4.29; Figure 4.30**), the four enzymes appeared to display similar enzymatic activities against simple disaccharides (**Figure 4.31**) and chito-oligosaccharides (**Figure 4.32**) tested. During the comparison, the overlapping specificities of the glycosidases were noted. The only significant difference that was observed at this point was the superior enzymatic efficiency of BT0459^{GH20} whereas the BT0456^{GH20} and BT0506^{GH20} displayed the lowest catalytic efficiency against these substrates, suggesting an importance of a +2 subsite of these glycoside hydrolases. These results suggested that BT0459^{GH20} prefers a GlcNAc in -1 subsite and can accommodate any other aldohexose in +1 subsite. The same pattern was observed in the enzyme kinetics data, however all four β-hexosaminidases showed a preference for GlcNAc-β1,2-Man disaccharide commonly found on complex N-glycan structures in nature (**Figure 4.34**). The active site of the four GH20 glycosidases appears to be conserved, thus they likely possess different structural features responsible for these differences in activities. Furthermore, the BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} were confirmed to be exo-glycosidases (**Figure 4.33**). They require galactose caps to be removed by the galactosidase, such as BT0461^{GH2}, before they can access their carbohydrate targets, further supporting the claim that N-glycan degradation is a sequential process. It was also found that BT0459^{GH20} is the only enzyme out of the four GH20s that can efficiently digest all chito-oligosaccharides, including chitobiose (GlcNAc-β1,4-GlcNAc), suggesting it may be required to cleave the terminal GlcNAc-β1,4-GlcNAc linkage of the scavenged PNGaseF-cleaved N-glycan structure.

The four GH20 enzymes displayed differential substrate specificities when screened for activity against various complex N-glycan structures (**Figures 4.35-42**). These results also confirmed that BT0459^{GH20} is a broad-acting enzyme. Interestingly, BT0506^{GH20} displayed almost identical N-glycan GlcNAc linkage degrading efficiency as BT0459^{GH20} (**Figure 4.40**), suggesting it is specialized in targeting complex N-glycans. This claim is supported by the RNA seq data analysis results, where BT0506^{GH20} is the highest upregulated β-hexosaminidase (**Figure 3.7.1**). Considering the minimal activity displayed against the disaccharide substrates tested (**Figure 4.31**), these results also support the hypothesis for the importance of the +2 subsite in BT0506^{GH20}. Meanwhile, BT0456^{GH20} remained the least efficient enzyme of the four. It was found to be unable to degrade GlcNAc-β1,2 linkages, but could cleave GlcNAc-β1,3/6 linkages (**Figures 4.38 and 4.42**), suggesting a specificity for the bisecting, tri-antennary and tetra-antennary complex N-glycan structures. It would also explain such a low-fold upregulation of BT0456^{GH20} seen on the RNA seq data of *B. thetaiotaomicron* grown on primarily bi-antennary bovine $\alpha_{1\text{-}}$AGP complex N-glycans (**Figure 3.7.1**). On the other hand, BT0460^{GH20} showed preference for one of the GlcNAc-β1,2 linkages of the bi-antennary complex N-glycan structure (**Figure 4.41**), suggesting the presence of GlcNAc caps on both of the bi-antennary N-glycan arms is required for its enzymatic efficiency or for binding.

Because N-glycoproteins can have an incredibly diverse glycosylation pattern, the degradation pathways of the fluorescently-labelled known complex N-glycan structures were analysed. This kind of assay is highly sensitive, thus allowing for an accurate identification of specific N-glycan linkages the four GH20s target (**Figure 4.42**). Based on these results, BT0460^{GH20} appears to have a bisecting GlcNAc-β1,4-Man linkage specificity. Similar to BT0459^{GH20} activity, BT0506^{GH20} could degrade all

GlcNAc linkages present on these complex N-glycan structures, with the exception of the bisecting GlcNAc. BT0456^{GH20} and BT0459^{GH20} showed a trace bisecting GlcNAc activity, but only when the galactose has been removed from both of the bi-antennary N-glycan arms, suggesting bisecting GlcNAc is not their preferential target (**Figure 4.42-D**). Enzymatic efficiency of BT0459^{GH20} on bisecting GlcNAc is also impaired by the presence of the core fucose on PNGaseF-cleaved IgG N-glycan structures (**Figure 4.42-D**). Considering *B. thetaiotaomicron* does not normally see PNGaseF-cleaved complex N-glycans, it is not surprising that BT0459^{GH20} struggles to target these core-fucosylated N-glycan structures. Although it is active on all complex N-glycan structures tested, BT0459^{GH20} shows a slight preference for GH18-cleaved bi-antennary complex N-glycan structures over the PNGaseF-cleaved ones whereas BT0506^{GH20} and BT0460^{GH20} do not display any preference. Based on these results, when the four GH20 glycosidases are combined, they should be capable of cleaving off every GlcNAc linkage found on typical complex N-glycan structures.

In the structure of BT0459^{GH20}, the CBM32-like domain appears to be positioned above the active site like an 'arm', suggesting it may be involved in the binding of complex N-glycan structures (**Figure 4.48**). Overlay of the BT0459^{GH20} structure with the StrepGH20b domain of N-acetyl-β-hexosaminidase (StrH) from *S. pneumonia* bound to the NGA2b N-glycan structure identified a lack of a platform loop in BT0459^{GH20} that StrepGH20b possesses (**Figure 4.49**). This difference and the low sequence identity between StrH and BT0459^{GH20} (21%) suggests that BT0459^{GH20} could employ a different method of N-glycan binding. A proposed binding model involves a novel CBM domain (F5/8_typeC) closing down over the active site of GH20 domain, binding and anchoring the potential N-glycan substrate in place (**Figure 4.51**). This kind of domain movement has already been observed by the CBM domain of a GH5_4 family enzyme from *B. licheniformis* (Liberato *et al.,* 2016). This hypothesis is further reinforced by the confirmation of the involvement of the CBM domains of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH10} in the binding of the complex biantennary N-glycan structures (**Figure 4.54**). Most of the CBMs bind their ligands with relatively low affinities compared to protein-protein binding interactions and this was observed in this experiment, where low binding affinity was displayed by the CBM domains of BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. More complex N-glycan structures (e.g. tetraantennary) should be tested to investigate whether these GH20 CBMs may be specialised in binding certain types of complex N-glycans.

The truncated derivative of BT0459^{GH20} lacking its C-terminal CBM (BT0459ΔCBM) domain displayed no significant difference in activity compared to the wild type BT0459^{GH20} when assayed against bovine α_1-AGP N-glycans, suggesting the CBM is not required for the binding and degradation of simple biantennary N-glycan structures. Instead its role may be to effectively bind and degrade more complex tri- and tetra-antennary N-glycan structures, although this remains to be investigated.

The work presented in this chapter indicates a model of N-glycan degradation where the activity of *B. thetaiotaomicron* enzymes is sequential. The BT0455^{GH33}, BT0456^{GH20}, BT0457^{CE}, BT0459^{GH20}, BT0460^{GH20}, BT0461^{GH20} and BT0506^{GH20} enzymes characterised and discussed in this chapter can degrade the typical complex N-glycan structure down to the core N-glycan tetrasaccharide (Manα1-3(Manα1-6)Manβ1-4GlcNAc) or pentasaccharide (Manα1-3(Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAc), depending on what enzyme was used to cleave it off the protein backbone (BT1044^{GH18} or PNGaseF). For the thermo-stable exo- β1,4-mannosidase BT0458^{GH2} to access its target Man-β1,4-GlcNAc, the α-mannoses capping it must be removed. The RNA seq data analysis suggests that N-glycan degradation requires the cooperative activity of several binding proteins and degrading enzymes encoded by several other PULs (**Figure 3.7.1; Figure 3.7.2**). For example, the full degradation of the plant rhamnogalacturonan-II (RG-II) by *B. thetaiotaomicron* requires three different PULs (Ndeh *et al.,* 2017). In order to establish this model of degradation, the other key enzymes required for the full N-glycan degradation must be characterised.

Chapter 5: Proposed model of N-glycan utilization by *B. thetaiotaomicron*

5.1. Introduction

All N-glycans share a common core structure (Man-α1,6(Man-α1,3)Man-β1,4GlcNAc-β1,4GlcNAc-β1-Asn-X-) and are grouped into high-mannose, complex and hybrid structures (Stanley *et al.,* 2009). Complex N-glycans are the most commonly found structures in the human gut in a form of dietary and host-derived N-glycoproteins. These structures can be further classified into biantennary, triantennary or tetraantennary structures and are commonly composed of N-acetylglucosamine (GlcNAc), mannose, galactose and sialic acid residues (**Figure 5.1**). In some cases, they can also contain a bisecting N-acetylglucosamine (GlcNAc) residue attached to the core mannose. For example, N-glycosylation patterns in immunoglobulin A (IgA), a most abundant secreted antibody in the human gut, are extremely diverse and change in response to the human health state (e.g. in disease, during pregnancy) (Bondt *et al.,* 2017). In healthy individuals, IgA predominantly contains bisecting sialylated biantennary and triantennary complex N-glycan structures. In some rare cases, complex N-glycan structures can also contain N-acetylgalactosamine (GalNAc) residues (Goettig *et al.,* 2016).

The human gut microbiota is exposed to a range of dietary and host-derived sources of N-glycans. The N-glycan degradation pathways have been investigated in a number of pathogenic bacteria, such as an opportunistic pathogen *B. fragilis*, where it was found to be important for the survival in the extra-intestinal niche (Cao *et al.,* 2014). However, the knowledge of N-glycan utilisation pathways by the symbiotic gut microbiota was lacking.

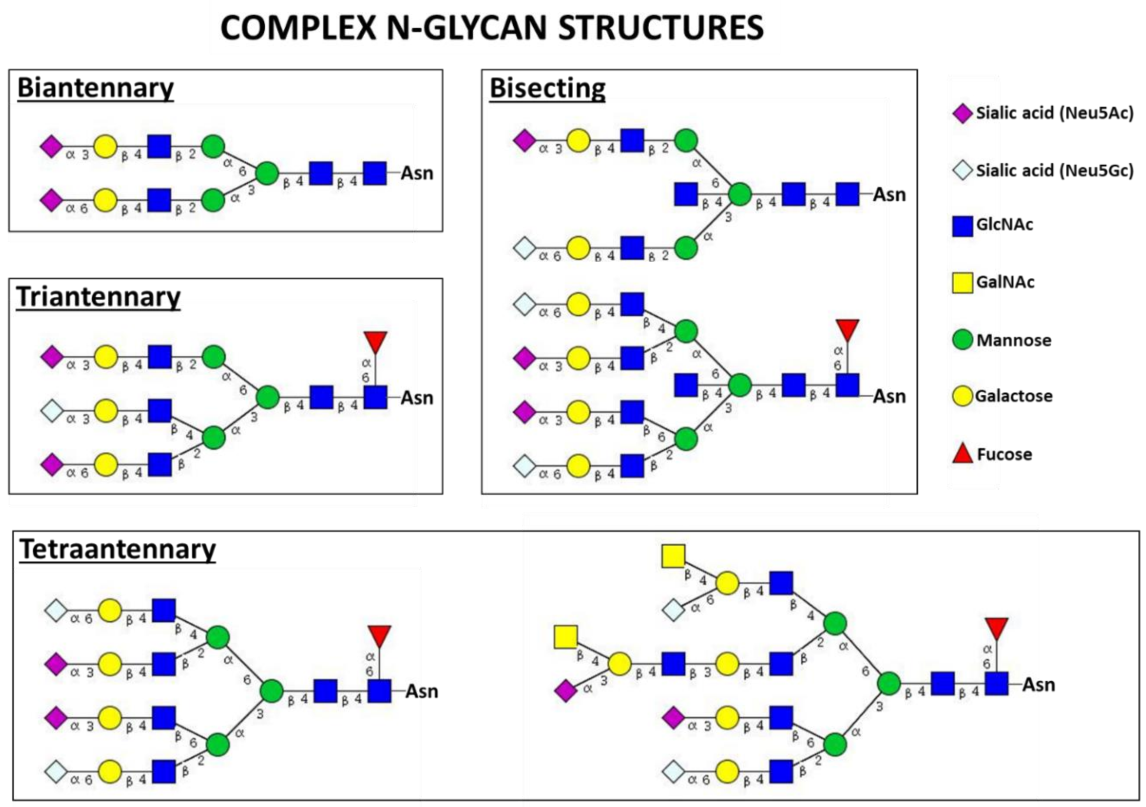


Figure 5.1: Schematic representation of complex N-glycan structures found on mammalian N-glycoproteins. N-glycan structures can be biantennary, triantennary and tetraantennary. The N-glycan structures were built using the Glycan Builder tool hosted on Expasy.org.

In **Chapter 3** it was shown that a number of human gut *Bacteroides* can grow on the glycan component of N-linked glycoproteins and *B. thetaiotaomicron* was used as a model to understand the mechanism of the breakdown process (**Figure 3.3**). To degrade such complex N-glycan structures, *B. thetaiotaomicron* requires a cooperative activity of numerous enzymes encoded by multiple polysaccharide utilization loci (PULs). These PULs have been identified using the RNA seq data analysis (**Figure 3.7.1**). Previous studies outlined in **Chapter 4** provided initial insights into how *B. thetaiotaomicron* degrades complex N-glycans and points to an exo-model of liberated N-glycan degradation where the N-glycan structures are first cleaved off protein backbone by GH18 endoglycosidase. No full complex N-glycan degradation pathway has been described in a commensal gut microbe previously. For example, the N-glycan utilisation that was observed in an opportunistic pathogen *B. fragilis* in vitro also appears to rely on the activity of the GH18 endoglycosidase, a member of an outer member protein complex termed Don (Cao *et al.,* 2014). However, the full degradation pathway was not investigated and this PUL does not contain all types of enzymes required for a full deglycosylation of complex N-glycans, suggesting similar multi-PUL apparatus as described in this thesis could be employed by this bacterium.

However, to establish a model of degradation, the importance of all key N-glycan degrading enzymes identified in **Chapters 3 and 4** to the proliferation of *B. thetaiotaomicron* must be clarified. The aim of this chapter was to review the enzyme activities, investigate their importance to the bacterium and ultimately, to build a model of N-glycan degradation and investigate how this apparatus is conserved in other members of gut microbiota.

5.2. Research objectives

I. Review and clarify the activities of enzymes involved in complex N-glycan degradation
II. Confirm the cellular localization of the enzymes involved in complex N-glycan degradation
III. Build a model of complex N-glycan degradation
IV. Investigate the conservation of the key enzymes of N-glycan degradation apparatus in other members of gut microbiota

5.3. Results

5.3.1. Overview of enzymes involved in N-glycan degradation

To degrade complex N-glycan structures*, B. thetaiotaomicron* requires a cooperative activity of multiple N-glycan binding proteins and N-glycan degrading enzymes. In **Chapter 4**, enzymes encoded by B.theta$^{0455\text{-}0461}$ locus and BT0506^{GH20} were characterised. These included a sialidase BT0455^{GH33}, sialic acid esterase BT0457^{CE}, β1,4-galactosidase BT0461^{GH2}, four β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20}, BT0506^{GH20} and β1,4-mannosidase BT0458^{GH2} (**Figure 5.2**). Combined activity of these glycosidases displayed an exo-model of liberated complex N-glycan structures. Due to the complexity of the N-glycan structures, it was observed that full N-glycan degradation also requires α-mannosidases and endo-β-N-acetylglucosaminidases not present in B.theta$^{0455\text{-}0461}$ locus. RNA sequencing data analysis identified BT0455-0461, BT0506-0507, BT1032-1051 and BT4404-4407 as key loci upregulated by *B. thetaiotaomicron* in response to complex N-glycans (**Figure 3.7.1**). An α-fucosidase BT1625^{GH29} was also upregulated. This enzyme and enzymes encoded by the BT1032-1051 locus have been investigated and characterised by Dr. Lucy Crouch (unpublished data). This locus encodes several key enzymes essential for N-glycan degradation, such as endo-β-N-acetylglucosaminidase BT1044^{GH18}; and α-mannosidase BT1032^{GH92}. The endo-β-N-acetylglucosaminidase BT1044^{GH18} was routinely used during the biochemical characterisation of BT0455-0461 and BT0506^{GH20} enzymes in Chapter 4. The overall information collected about these glycosidases is displayed in **Table 5.1**. Based on the data collected in Chapter 3, Chapter 4 and by Dr. Lucy Crouch, schematic representation of N-glycan linkages the key enzymes are predicted to target is displayed in **Figure 5.3**. It is important to note that the α1,3-mannosidase BT3991 that targets α1,3-Man linkages of N-glycans (Dr. Fiona Cuskin, unpublished data) was included in this model, but it was not used otherwise in the work presented in this thesis.

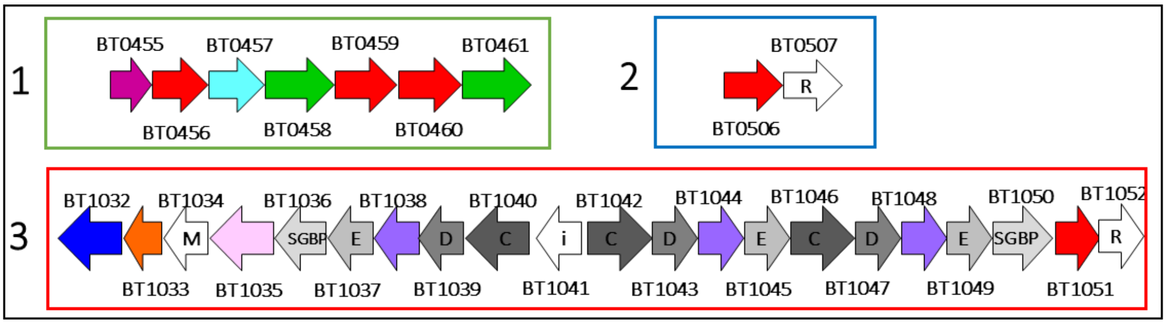


Figure 5.2: Composition of the key loci required for complex N-glycan utilisation by *B. thetaiotaomicron*. The combined information obtained from the Uniprot and KEGG databases was used to make this figure. The loci are color-coded for easier identification in Table 5.1.

Table 5.1: Details of the characterized enzymes.

Enzyme	Classification	Linkage specificity	CAZy family	Predicted size (kDa)	Predicted signal peptide
BT0455	α-sialidase	Sial-α2/3/6-	**GH33**	61	SP I
BT0456	β-hexosaminidase	bisecting GlcNAc-β1,4-	**GH20**	78	SP I
BT0457	sialic-acid-specific 9-O-acetylesterase	Neu9Ac-	**CE**	79	SP I
BT0458	β-mannosidase	Man-β1,4-	**GH2**	99	SP I
BT0459	β-hexosaminidase	GlcNAc-β1,2/3/4/6-	**GH20**	87	SP II
BT0460	β-hexosaminidase	GlcNAc-β-1,2/4/6-	**GH20**	78	SP I
BT0461	β-galactosidase	Gal-β1,4-	**GH2**	95	SP I
BT0506	β-hexosaminidase	GlcNAc-β1,2/3/4/6-	**GH20**	86	SP II
BT1032*	α-mannosidase	Man-α1,6/-	**GH92**	87	SP I
BT1033*	mannosyl-phosphorylase	Man-β1,4-	**GH130**	37	SP II
BT1035*	LacNAc'ase	Gal-β1,4-GlcNAc-β1,4-	**New**	84	SP I
BT1038*	endo-β-N-acetylglucosaminidase	GlcNAc-β1,4-GlcNAc	**GH18**	38	SP II
BT1044*	endo-β-N-acetylglucosaminidase	GlcNAc-β1,4-GlcNAc	**GH18**	42	SP II
BT1048*	endo-β-N-acetylglucosaminidase	GlcNAc-β1,4-GlcNAc	**GH18**	42	SP II
BT1051*	β-hexosaminidase	GlcNAc-β1,4-	**GH20**	87	SP II
BT1625*	α-fucosidase	Fuc-α1,3-	**GH29**	69	SP II

Enzymes labelled with * were characterised by Dr. Lucy Crouch. SP I stands for signal peptide recognised by signal peptidase I and SP II stands for signal peptide recognised by signal peptidase II done using LipoP 1.0 server. GH - glycoside hydrolase. CE – carbohydrate esterase.

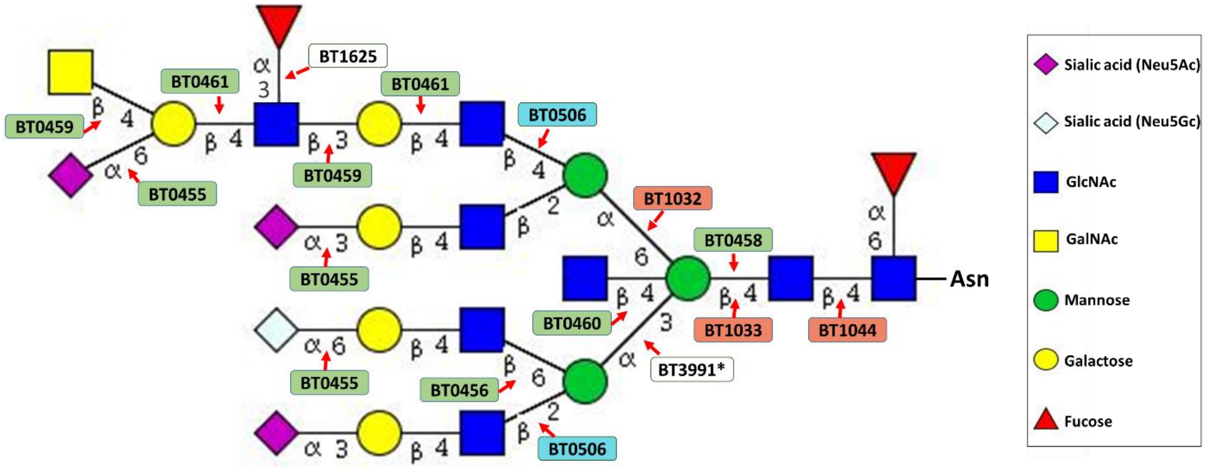


Figure 5.3: Schematic representation of the complex N-glycan structure characterised enzymes would target. The linkages cleaved by each enzyme are shown by arrows. The GH20s BT0456, BT0459, BT0460 and BT0506 have overlapping linkage specificities, thus the linkages displayed in this figure are just an example of what structures they can target. Glycoside hydrolases are colour-coded based on the PULs they are found in: green (BT0455-0461), blue (BT0506-0507) and red (BT1032-1051).

5.3.2. Cellular localization

In order to fully understand the pathway of glycan degradation, it is important to know where the key enzymes are located in the context of the cell. The LipoP 1.0 Bioinformatics tool can be used to predict the cellular localization of the proteins based on their signal peptide (SP). Proteins with a SP type I (SP I) signal peptide are usually found in the periplasm whereas SP type II (SP II) signal peptide is characteristic of lipoproteins, that are usually found on the cell surface in Bacteroidetes (Zuckert *et al.,* 2014). The predictions based on the signal peptides of the key Cazymes were obtained using the LipoP 1.0 tool and are listed in a **Table 5.1**. The results showed that BT0459^{GH20}, BT0506^{GH20} and BT1044^{GH18} have SP II signal peptide, suggesting they may be localized on the cell surface. Interestingly, the only sialidase encoded by *B. thetaiotaomicron,* BT0455^{GH33}, was predicted to have an ambigious SP I cleaving site which indicates it may be located in the periplasm.

To investigate the cellular localization experimentally, rabbit polyclonal antibodies were raised against BT0455^{GH33}, BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20}, BT0506^{GH20} and BT1044^{GH18} (Eurogentec). Cellular localization assays were done following a protocol described in Section 2.3.9.1. The generated antibodies were tested against the purified recombinant proteins and were found to be very specific with relatively low background contamination. Based on the disappearance of the band corresponding to the recombinant protein in the proteinase K-treated sample (+), the results displayed in **Figure 5.4** show that BT1044^{GH18}, as predicted by the SP II signal peptide, is localised outside of the cell. Despite having a predicted SP II signal peptide, the data suggests that BT0459^{GH20} and BT0506^{GH20} are located inside of the cell, likely facing the periplasm to enable for efficient deglycosylation of the N-glycan structures. BT0455^{GH33}, BT0456^{GH20} and BT0460^{GH20}, as predicted by the SP I signal peptide, were shown to be inside of the cell. It was also investigated whether these proteins are secreted outside of the cell, but no secreted proteins were detected by Western Blot of concentrated spent media (data not shown).

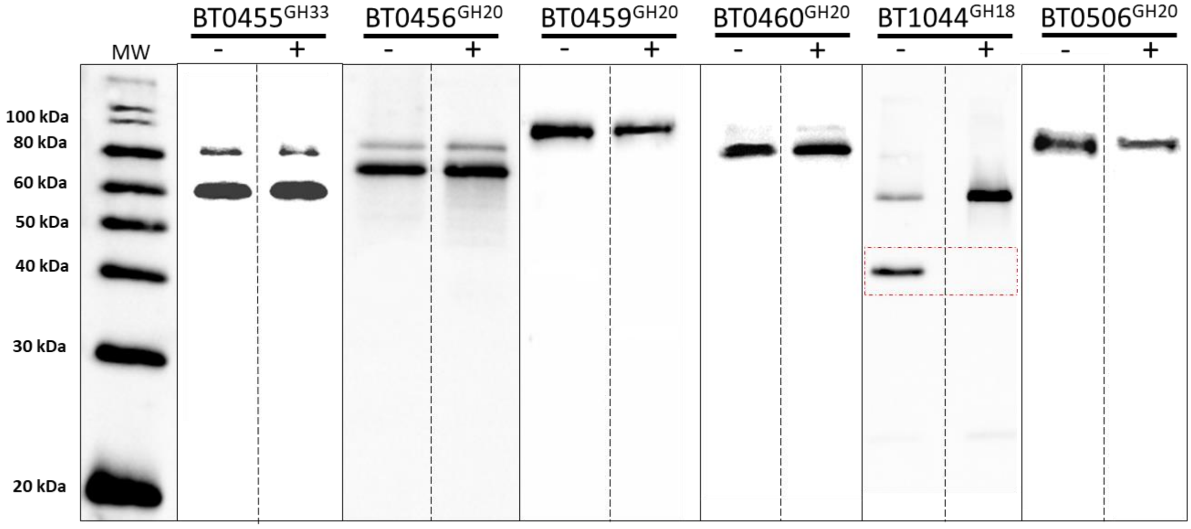


Figure 5.4: Cellular localization of *Bt* enzymes involved in N-glycan breakdown. The + stands for Proteinase K treated cells and the – stands for untreated cells. The presence of a band in samples treated with Proteinase K (+) shows that the protein has not been degraded and is likely located inside of the cell. The absence of the band shows that the protein is accessible to the protease and is therefore likely located outside of the cell. BT1044^{GH18} (41kDa) results (red box) suggest it is located outside of the cell. The ~60kDa band present in BT1044^{GH18} sample is suspected to be a protein that cross reacts with the polyclonal antibody. MW – MagicMark^{TM} molecular weight marker (ThermoFisher).

5.3.3. Whole cell assays

The cellular localization data indicated that N-glycans are cleaved off the protein backbone by BT1044^{GH18} externally prior to the import and further processing. To further confirm this activity was localised to the cell surface, wild type *B. thetaiotaomicron* was grown on α_1-AGP until mid-exponential phase (OD_{600nm}=0.4-0.6) and whole cell assay was set up as outlined in Section 2.3.10. *B. thetaiotaomicron* is an obligate anaerobe, thus the whole cells harvested in this way are intact but metabolically inactive due to the aerobic conditions and cannot actively import glycans because SusC-like transporters are TonB-dependent, therefore require energy to import the glycans across the outer membrane. Whole cells, supernatant and sonically-lysed cells were incubated with 2% α_1-AGP in PBS at 37 °C up to 24 hours. Samples were collected over the duration of time and analysed by TLC (**Figure 5.5**). Based on the appearance of the bands corresponding to different sugars, N-glycan degradation was observed in whole cells and lysed cells but not in supernatant or Proteinase K (PK) treated cells, suggesting the N-glycans are cleaved off protein backbone and sialic acid is released at the cell surface. The sialidase activity was also observed in both whole cells and proteinase K-treated whole cells, suggesting the BT0455^{GH33} enzyme could be located both at the cell surface and in the periplasm. Although N-glycan degradation products are visible in the supernatant sample (**B**), no additional N-glycan degradation activity was observed, suggesting the required enzymes are not secreted.

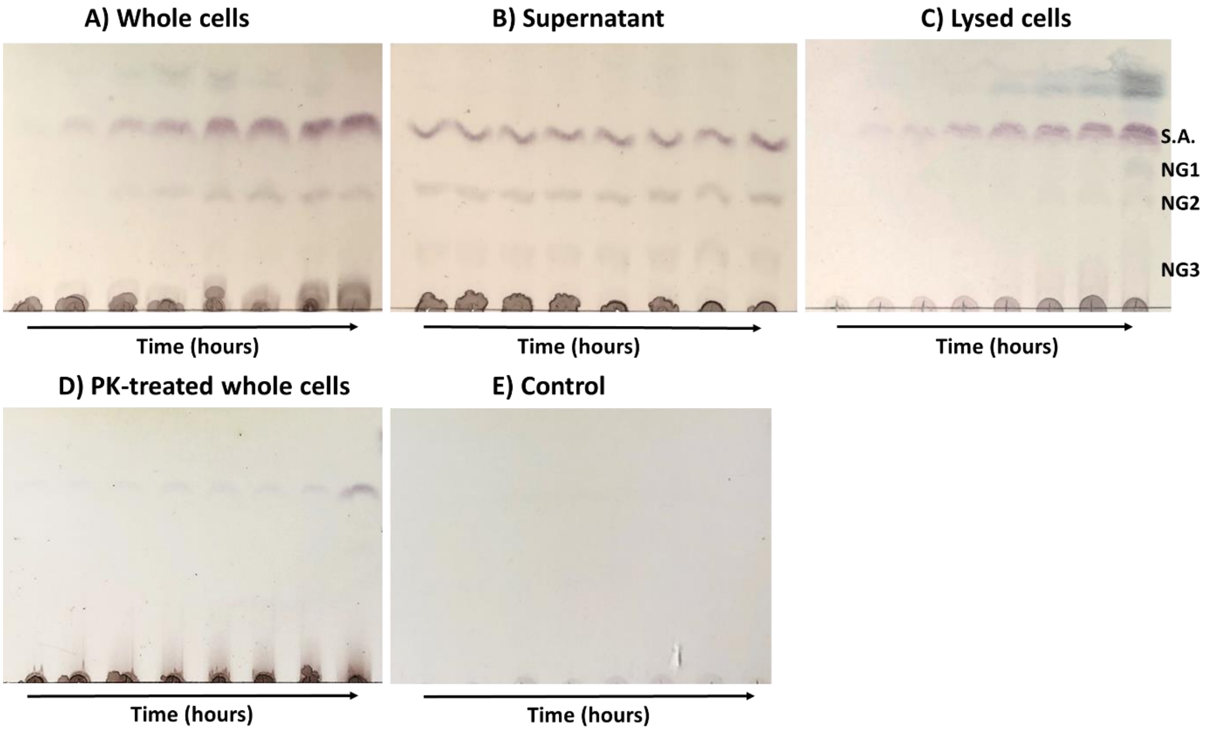


Figure 5.5: Whole cell assays show surface whole N-glycan release activity by the enzymes encoded by *B. thetaiotaomicron*. Assays were performed in the presence of 2% α_1-acid glycoprotein with harvested whole cells (**A**), growth media supernatant (**B**), sonicated cell lysate (**C**) and proteinase K-treated whole cells (**D**). Control (**E**) contains whole cells incubated in PBS without a substrate. Reactions were incubated for up to 24 hours, with sampling at 1, 2, 4, 8 and 24 hours. The release of the sialic acid (S.A.) and predicted N-glycan structures (NG1-3) was observed. TLC plates were stained with DPA.

5.3.4. Gene deletion analysis

B. thetaiotaomicron was shown to be capable of utilising $\alpha_{1\text{-}}$AGP *in-vitro* as a sole carbon source (**Figure 3.3**). In this study, we wanted to test the role of the key enzymes from BT0455-0461 and BT1032-1051 loci in N-glycan utilization by *B. thetaiotaomicron*. To do this, it was attempted to generate the deletion mutants ΔBT0455^{GH33}, ΔBT0455^{GH33}+ΔBT1035, ΔBT0458^{GH2}, ΔBT0458^{GH2}+ΔBT1033^{GH130}, ΔBT0459^{GH20}, ΔBT0460^{GH20}, ΔBT0461^{GH2}, ΔBT0506^{GH20} and a deletion mutant missing an entire ΔBT0455-0461 loci using the methods described in Sections 2.1.13 and 2.1.18. The primers were designed to amplify 1kb of upstream and downstream of genomic region flanking the genes of interest and are listed in Supplemental table 3.2 (**Figure 5.6**). For1 primer was engineered to contain *SalI* site and Rev1 primer was engineered to contain the *XbaI* site that allows for cloning of the sewing PCR product into the pExchange-*tdk* vector.

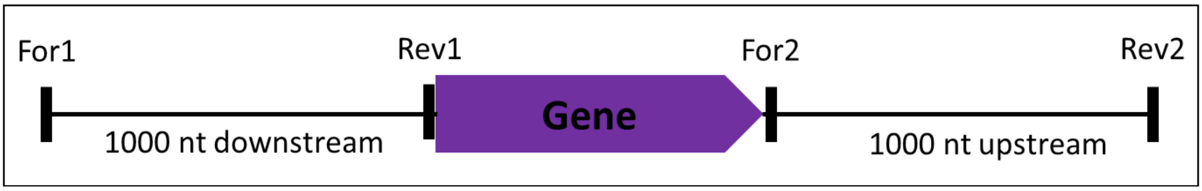


Figure 5.6: Deletion mutant design. 1000nt downstream and 1000 nt upstream of the gene of interest were amplified and stitched together.

Following the conjugation, recombinant mutants were screened for desired deletions. Clones carrying a desired deletion showed a ~1000 bp band amplified by using specially designed PCR primers (Supplemental table 3.2). Wild type *B. thetaiotaomicron* would produce a fragment of ~4000 bp in size. An example of a successful conjugation is displayed in **Figure 5.7**. The successfully generated deletion mutants were: ΔBT0458^{GH2}, ΔBT0459^{GH20}, ΔBT0460^{GH20}, ΔBT0461^{GH2} and ΔBT0506^{GH20}. The ΔBT1044^{GH18} mutant was generated by Dr. Lucy Crouch.

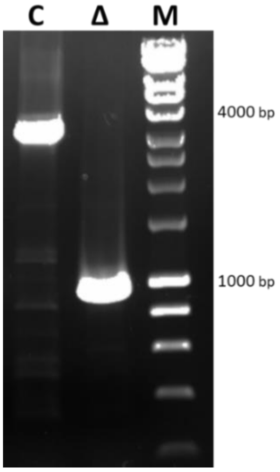
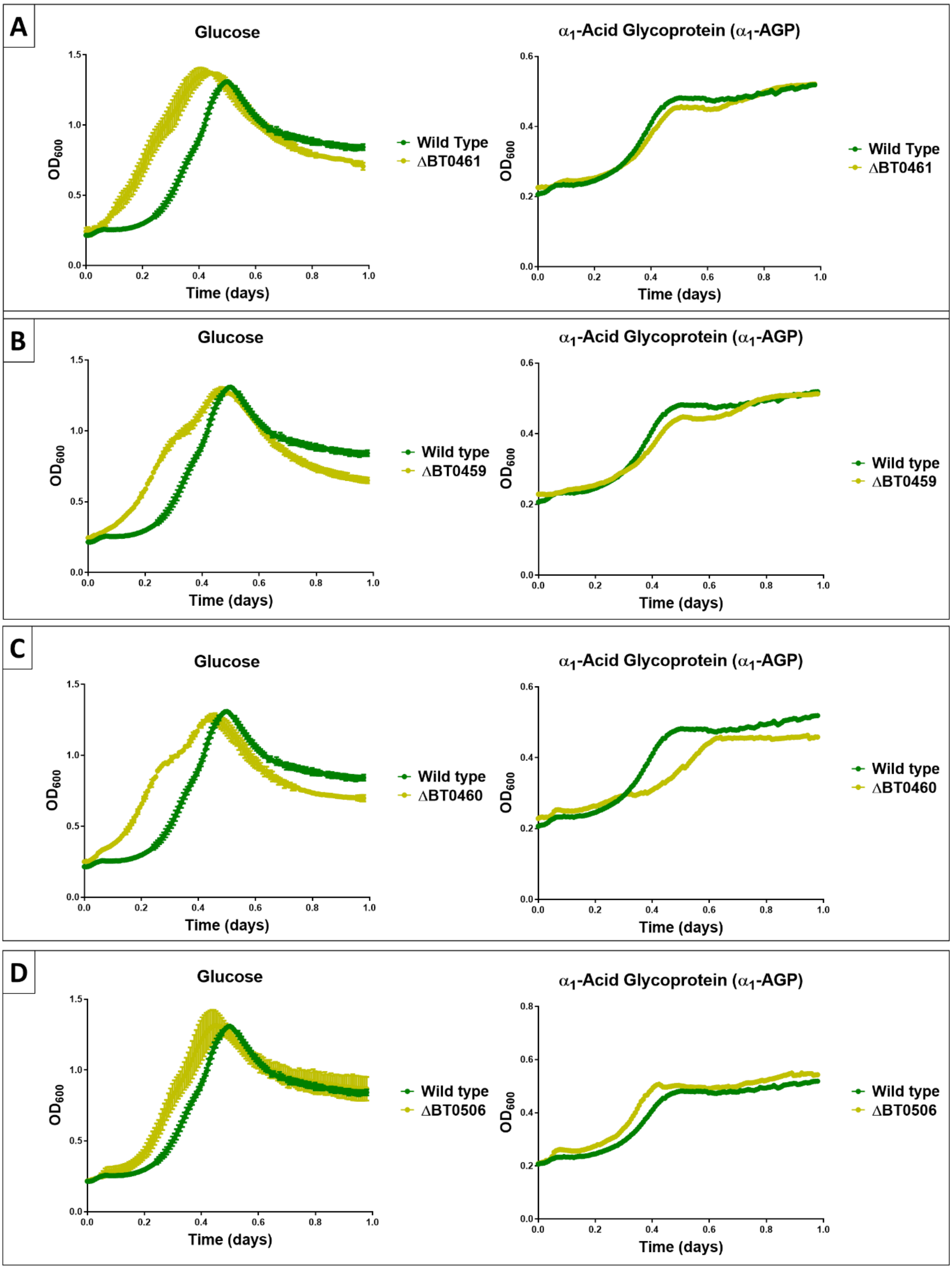


Figure 5.7: Screening for deletion mutants. PCR results showing the amplification of the *B. thetaiotaomicron* genomic DNA of ΔBT0461^{GH2} mutant. C – wild type *B. thetaiotaomicron* control. Δ - deletion mutant. M- High fidelity DNA marker (Sigma).

To investigate the effect of the deletion mutants on the ability of *B. thetaiotaomicron* to utilise N-glycoproteins, wild type and deletion strains were grown separately in minimal media supplemented with either 1% glucose or 2% α_1-AGP. Growths were monitored automatically by measuring the optical density of the cultures at 600nm every 15 minutes for 24 hours using the Epoch^{TM} Microplate Spectrophotometer (see Section 2.1.19). The results are displayed in **Figure 5.8**. Compared to the wild type *B. thetaiotaomicron*, the growth of all mutants on glucose had a much shorter lag phase. No defect was detected on ΔBT0461^{GH2} mutant growth on α_1-AGP, suggesting *B. thetaiotaomicron* produces other GH2 β-galactosidases that have overlapping specificities with the BT0461^{GH2} (**Figure 5.8-A**). When the deletion mutants of the β-hexosaminidases BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} were compared, only the growth of ΔBT0460^{GH20} on α_1-AGP was observed to be partially retarded in comparison to the wild type (**Figure 5.8-B**). These results suggest that the activities of these β-hexosaminidases are redundant when *B. thetaiotaomicron* is grown on complex bi-antennary N-glycan structures. The deletion of β-mannosidase BT0458^{GH2} does not appear to have any effect on *B. thetaiotaomicron* growth (**Figure 5.8-E**). These results indicate that *B. thetaiotaomicron* possesses enzymes that can replace the activities of some of the BT0455-0461 loci enzymes, but not as efficiently. In comparison, the deletion of the endo-β-N-acetylglucosaminidase BT1044^{GH18} has a severe growth defect compared to the wild type *B. thetaiotaomicron* (**Figure 5.7-F**). These results indicate that BT1044^{GH18} is a key B. *thetaiotaomicron* enzyme required to initiate the N-glycan degradation by cleaving the complex biantennary N-glycans from the α$_1$-AGP backbone.



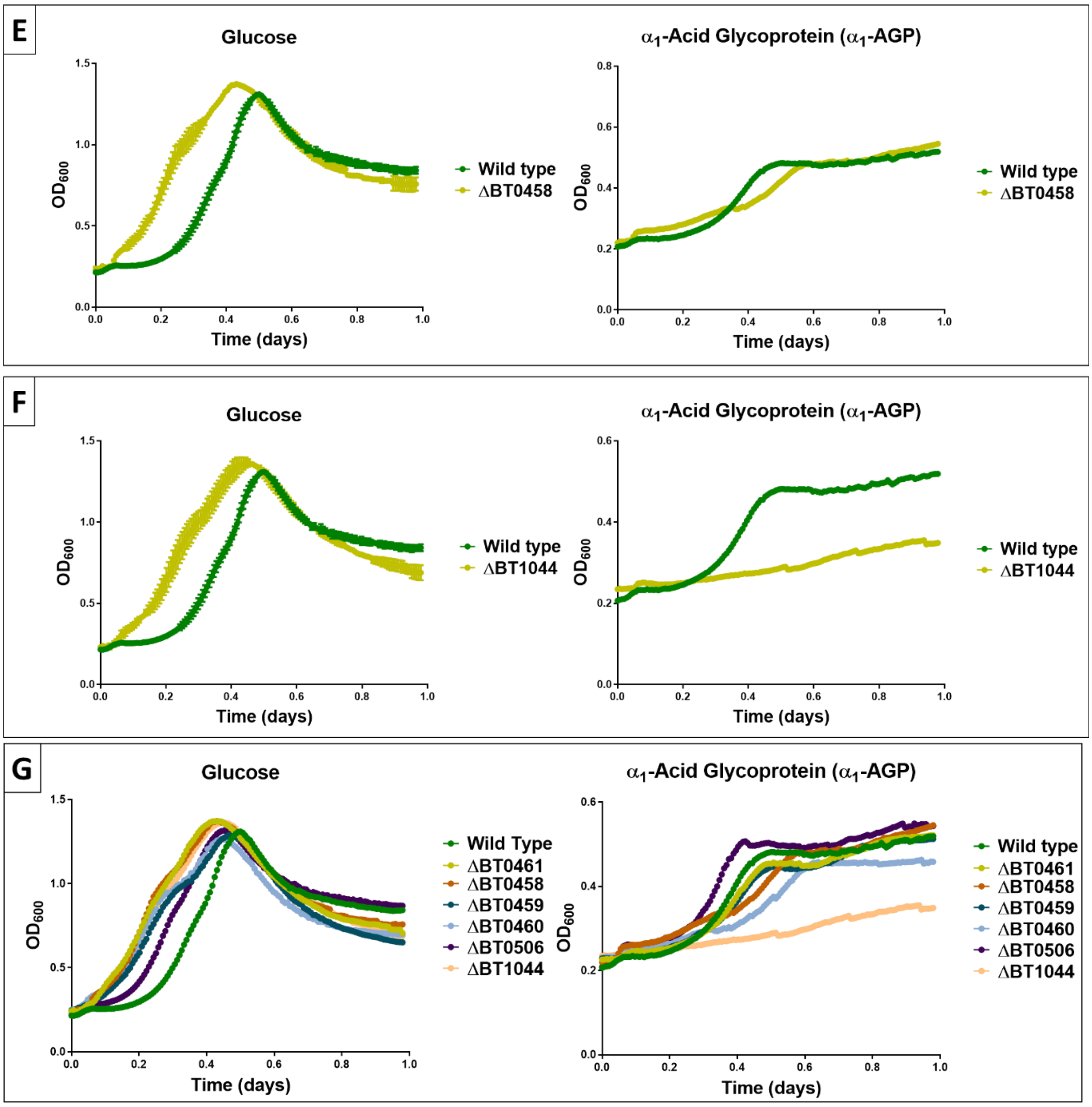
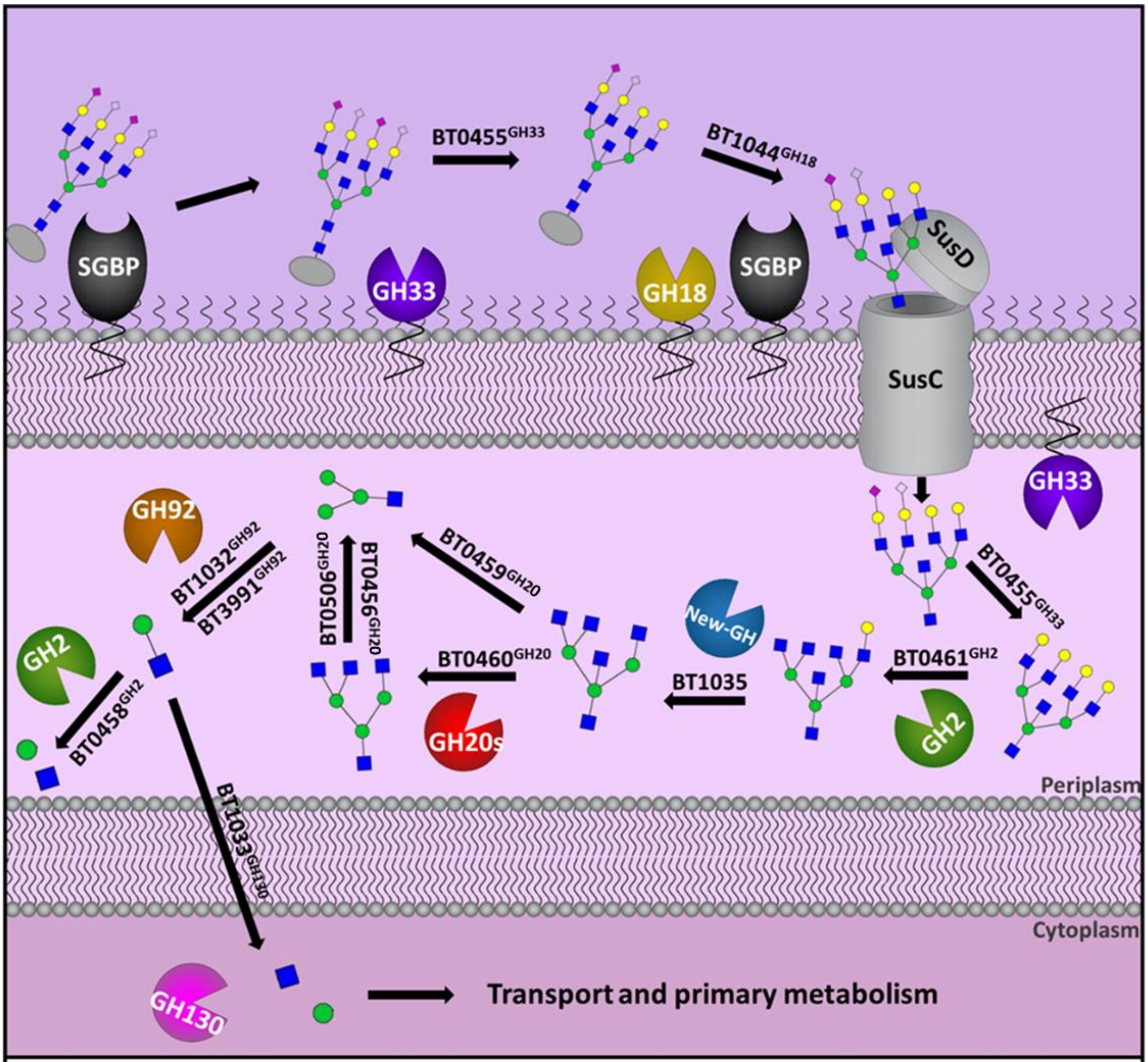


Figure 5.8: Comparing the growth of the wild type *B. thetaiotaomicron* and deletion mutants on native α_1-AGP N-glycans. A) ΔBT0461^{GH2}; **B)** ΔBT0459^{GH20}; **C)** ΔBT0460^{GH20}; **D)** ΔBT0506^{GH20}; **E)** ΔBT0458^{GH2}; **F)** ΔBT1044^{GH18}; **G)** Overlay of all mutant growth data (mean only). The strains were cultured in minimal media containing either 1% glucose or 2% α_1-AGP. Growth data for each condition was obtained from triplicate cultures. The error bars indicate the standard deviation from the mean of each triplicate data set.

5.3.5. Proposed N-glycan utilization model

The combination of the work presented in this thesis with the results of the studies conducted by Dr. Lucy Crouch (ICaMB, Newcastle University) enabled us to propose a model of N-glycan utilization by *B. thetaiotaomicron (***Figure 5.9)**. RNA sequencing data analysis followed by an extensive biochemical characterisation indicates that the complex N-glycan structure utilisation requires cooperative activity of multiple loci. Based on the upregulation data, the key enzymes required for the degradation of these structures are encoded by the BT0455-0461, BT0506-0507 and BT1032-1051 loci. However, the KO mutant growth data (**Figure 5.8**) suggests that this pathway is much more complicated where only endoglycosidase BT1044^{GH18} was shown to be critical for complex biantennary α_1-AGP N-glycan breakdown. Unfortunately, the failure to generate the ΔBT0455-0461 mutant prevented us from investigating the importance of this locus for the *B. thetaiotaomicron* proliferation.

Degradation of the complex N-glycans is likely initiated by the binding of the N-glycoproteins to the SGBPs at the cell surface, however it has not been shown yet. Here, the sialidase BT0455^{GH33} could cleave some of the sialic acid caps off the terminal N-glycan antennae. Cellular localization data (**Figure 5.4**) suggests that BT0455^{GH33} is located inside of the cell however the whole cell assay data (**Figure 5.5**) indicates that the sialic acid is cleaved off the N-glycan substrate following the 24 hours of incubation with the whole cells pre-treated with the Proteinase-K, protease that destroys all proteins localised at the cell surface. Considering BT0455^{GH33} is the only sialidase encoded by *B. thetaiotaomicron*, the observed sialic acid release suggests an BT0455^{GH33} activity. It is important to note that, based on the low LipoP prediction score of 2.59, the predicted signal peptide for BT0455^{GH33} SP-I is ambigious. Considering that the TonB-dependent outer membrane transporter complex (SusC/D-like) requires energy to import N-glycan structures into the cell for the degradation by periplasmic glycosidases, when combined these results support the proposition that the sialidase BT0455^{GH33} may employ an unknown mechanism of action to be able to be both surface and periplasm localised. This kind of activity could be required to desialylate N-glycan structures that were imported into the cell before all sialic acid caps could be removed by the sialidase on the cell surface. BT1044^{GH18} then cleaves the N-glycans from the protein backbone generating glycan structures that can be imported into the periplasm by the SusC/D-like system. It is important to note that BT1044^{GH18} prefers de-sialylated complex N-glycan structures but can cleave sialylated structures as well (Dr. Lucy Crouch, unpublished data). Once inside the periplasm, the galactose caps are removed off the N-glycan structure by the β-1,4-galactosidase BT0461^{GH2}. It was previously observed that BT0461^{GH2} can leave some galactose caps on the N-glycan structures. If that happens, a new GH-family enzyme that been identified and characterised by Dr. Lucy Crouch can cleave the Gal-β1,4-GlcNAc- disaccharide off the complex N-glycan structures. Then, various GlcNAc linkages are cleaved by four redundant β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. However, because these glycosidases were shown to have overlapping specificities, it is currently unclear which linkages they would target *in vivo*. Based on the biochemistry data obtained *in vitro* that was discussed in Chapter 4, their predicted target linkages are displayed in the model. Once GlcNAc linkages are removed, the mannose caps are cleaved off by the α-mannosidases BT1032^{GH92} and BT3991^{GH92}. The generated disaccharide (Man-β1,4-GlcNAc) is then either digested by the exo-acting β-mannosidase BT0458^{GH2} or imported into the cytoplasm to be processed by the phosphorylase BT1033^{GH130}. It is possible that both, the released monosaccharides and the generated Man-β1,4-GlcNAc disaccharide can be imported into the cytoplasm. Based on the RNA seq-derived upregulation data (**Figure 3.7.1.**), *B. thetaiotaomicron* upregulates BT1033^{GH130} more in response to the presence of complex N-glycans. However, it is not clear what the levels of the BT0458^{GH2} mannosidase are present in the cell so it is currently unknown which enzyme of these two is preferred by this bacterium.



◆ **Sialic acid (Neu5Ac)**

◇ **Sialic acid (Neu5Gc)**

🟦 **GlcNAc**

🟢 **Mannose**

🟡 **Galactose**

SGBP – Predicted Surface Glycan-Binding Proteins
GH33 – BT0455 (sialidase)
SusC/D – BT1039/BT1040; BT1042/BT1043, BT1046/BT1047
GH18 – BT1044 (endo-β-N-acetylglucosaminidase)
GH2 – BT0461 (β-galactosidase)
GH20 – BT0456, BT0459, BT0460 and BT0506 (β-hexosaminidases)
GH92 – BT1032 and BT3991* (α-mannosidases)
GH2 – BT0458 (β-mannosidase)
GH130 – BT1033 (mannosyl phosphorylase)
New GH – BT1035 (cleaves Galβ1,4-GlcNAc- disaccharide off)

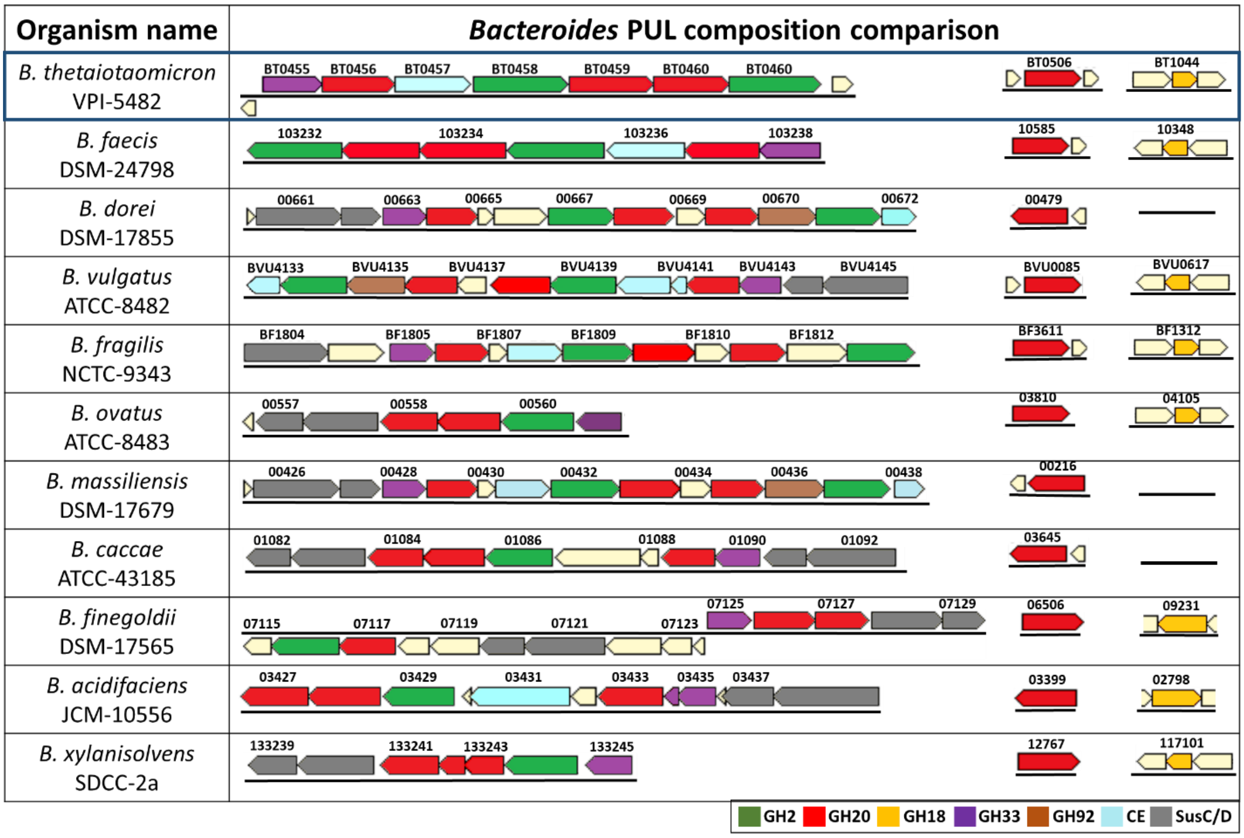
Figure 5.9: Proposed model of complex N-glycan utilization by *B. thetaiotaomicron*. The sequential degradation is represented by arrows. The list of key enzymes identified to be involved in complex N-glycan degradation is listed in the legend. For details see the text above.

5.3.6. N-glycan PULs in other *Bacteroides*

Deglycosylation of N-glycoproteins has been previously observed in pathogens such as *Capnocytophaga canimorsus*, *Streptococcus pyogenes*, *Streptococcus oralis, Streptococcus pneumoniae* and *Bacteroides fragilis* (Renzi *et al.,* 2011; Robb *et al.,* 2017). To our knowledge, the N-glycan degradation model described in this chapter is the first system devoted to foraging N-glycoproteins by a commensal, symbiotic gut microbe. The ability to utilise N-glycans could give *B. thetaiotaomicron* an advantage during the periods of starvation and could play a role in the prevention of opportunistic infections. This suggests that the ability to deglycosylate dietary and host-derived N-glycoproteins could be a favourable trait in the gut ecosystem.

The enzymes encoded by the BT0455-0461, BT0506-0507 and BT1032-1052 loci were characterised and found to be the key members of the N-glycan degradation apparatus in *B. thetaiotaomicron*. In this study, we wanted to compare and investigate the prevalence of the genes homologous to BT0455-0461, BT0506^{GH20} and BT1044^{GH18} in other members of human gut microbiota. Amino acid sequences of each of these proteins were used to identify homologous proteins on the KEGG database and Integrated Microbial Genomes (IMG) system. The BT0455-0461 locus was used as a query on the Polysaccharide Utilization Loci (PUL) database (PULDB) to predict homologous PULs in other *Bacteroidetes* species found in human gut microbiota (www.cazy.org/PULDB). The results from both searches were analysed and compared. Ten species of *Bacteroides* were found to encode similar apparatuses to *B. thetaiotaomicron* (**Table 5.2**). These include *B. faecis, B. dorei, B. vulgatus, B. fragilis, B. ovatus, B. massiliensis, B. caccae, B. finegoldii, B. acidifaciens, B. xylanisolvens*. All of these microbes contain homologues of endo-β-N-acetylglucosaminidase BT1044^{GH18}, with an exception of *B. dorei, B. massiliensis* and *B. caccae* that do not encode any GH18 enzymes, suggesting they employ a different mechanism to cleave the N-glycans off the protein backbone, potentially by utilising a PNGaseF-like enzymes. All of these bacteria encode a GH33 sialidase homologues and at least one GH2- and three GH20-family enzymes. The high similarity of the PULs identified in these microorganisms suggests they all could be specialised in N-glycan degradation.

Table 5.2: Prevalence of key enzymes of N-glycan degradation apparatus in gut *Bacteroides*.



Organism name	*Bacteroides* PUL composition comparison		
B. thetaiotaomicron VPI-5482	BT0455 BT0456 BT0457 BT0458 BT0459 BT0460 BT0460	BT0506	BT1044
B. faecis DSM-24798	103232 103234 103236 103238	10585	10348
B. dorei DSM-17855	00661 00663 00665 00667 00669 00670 00672	00479	—
B. vulgatus ATCC-8482	BVU4133 BVU4135 BVU4137 BVU4139 BVU4141 BVU4143 BVU4145	BVU0085	BVU0617
B. fragilis NCTC-9343	BF1804 BF1805 BF1807 BF1809 BF1810 BF1812	BF3611	BF1312
B. ovatus ATCC-8483	00557 00558 00560	03810	04105
B. massiliensis DSM-17679	00426 00428 00430 00432 00434 00436 00438	00216	—
B. caccae ATCC-43185	01082 01084 01086 01088 01090 01092	03645	—
B. finegoldii DSM-17565	07125 07127 07129 07115 07117 07119 07121 07123	06506	09231
B. acidifaciens JCM-10556	03427 03429 03431 03433 03435 03437	03399	02798
B. xylanisolvens SDCC-2a	133239 133241 133243 133245	12767	117101

🟩 GH2 🟥 GH20 🟨 GH18 🟪 GH33 🟫 GH92 ▢ CE ▨ SusC/D

PULs and enzymes homologous to BT0455-0461, BT0506^{GH20} and BT1044^{GH18} are displayed. IMG, KEGG and PULDB databases were used to identify these enzymes and to build the annotated composition figures. – means no homologous GH18 enzyme was found in these bacteria.

Variations of these PULs were observed in other species of bacteria. For example, *Akkermansia muciniphila* and *Akkermansia glycaniphila* do not possess any GH18 family enzymes, but both produce two homologous sialidases to BT0455^{GH33} and several homologues of a BT0461^{GH2} galactosidase. This could be explained by their preference for O-glycan mucin structures (Derrien *et al.,* 2004). A dog oral cavity commensal *Capnocytophaga canimorsus,* that causes severe infections in dog bite wounds in humans, was also shown to be capable of deglycosylating N-glycans present on human IgG by utilising a large homologous complex of proteins that consists of a sialidase SiaC and GpdCDGEF PUL proteins. This PUL contains a SusC/D pair, a GH18 endo-β-*N*-acetyl-glucosaminidase and two glycan binding proteins (Renzi *et al.,* 2011). The comparison of these results indicate that the members of *Bacteroides* spp. are more specialised in degradation of N-glycan structures.

5.3.7. Discussion

The work described in this study revealed that the complex N-glycan degradation by *B. thetaiotaomicron* requires a cooperative activity of numerous enzymes encoded by multiple loci. Compared to the previously conducted carbohydrate utilisation studies where it was shown that *Bt* usually requires one to three specialised PULs to degrade complex glycan structures, such as heparin/herapan sulphate (1 PUL_{Hep}; Cartmell *et al.,* 2017), rhamnogalacturonan-I backbone (1 RGI-PUL; Luis *et al.,* 2018), galactan (1 Gal-PUL; Luis *et al*., 2018) and RG-II (3 discrete PULS; Ndeh *et al.,* 2017), the identification of this complex multi-PUL-dependent N-glycan degradation pathway is interesting. In case of these complex glycans, it was observed that the number of PULs required to degrade these structures usually increases with the variety and complexity of the glycan. So from the initial point of view it looks like the N-glycan degradation mechanism is not that different, however it appears to have a lot more redundancy compared to the apparatuses required to degrade other glycans, such as RG-I, which suggests that there may be a lot of heterogeneity in the N-glycan structures faced by the members of human gut. The combined results of the RNA-seq data analysis (**Chapter 3**), biochemical characterisation (**Chapter 4**), cellular localization and genetic analysis results described in this chapter have identified and confirmed that the BT0455-0461, BT0506-0507 and BT1032-1051 loci are involved in complex N-glycan utilization.

The glycoside hydrolases that were characterised in **Chapter 4** are the broad-acting α2,3/6-sialidase $\text{BT0455}^{\text{GH33}}$, four β-hexosaminidases with overlapping specificities $\text{BT0456}^{\text{GH20}}$, $\text{BT0459}^{\text{GH20}}$, $\text{BT0460}^{\text{GH20}}$ and $\text{BT0506}^{\text{GH20}}$, β1,4-mannosidase $\text{BT0458}^{\text{GH2}}$ and β1,4-galactosidase $\text{BT0461}^{\text{GH2}}$. In order to fully understand N-glycan degradation pathway and the roles of these enzymes, it was important to know where they are located in the context of the cell. The N-terminal signal peptide of these enzymes can be used to predict their cellular localization. A SP I signal peptide predicts the protein is located in the cell periplasm whereas SP II predicts the protein is surface localized (Zuckert *et al.,* 2014). It is important to note that signal peptide prediction by bioinformatic tools, such as LipoP, is based on the machine-learning methods, such as training using known protein sequences as examples, and may not be as accurate when predicting signal peptides of novel proteins. So, utilizing cellular localization assays to confirm these predictions are accurate is of utmost importance. In this study, a cellular localization assay utilizing target-specific antibodies was used because it is a highly accurate and sensitive technique (Cuskin *et al*., 2015). Cellular fractionation technique could also be used to investigate where the protein of interest is located. It utilises ultracentrifugation method to separate organelles and macromolecules into separate fractions, such as whole cells, nuclei and cytoskeletons during the low-speed centrifugation, mitochondria and lysosomes during the medium-speed centrifugation and microsomes together with ribosomes during the high-speed centrifugation. The presence of the protein in isolated fractions can then be investigated using SDS-PAGE and immunoblotting (Alberts et al., 2002). Other methods, such as fluorescence microscopy-based techniques (e.g. utilising fluorescent dyes) could also be used to investigate the cellular localization of the proteins, but these methods are not as sensitive (Collings, 2013).

Although the broad-acting BT0459^{GH20} and BT0506^{GH20} β-hexosaminidases were initially predicted to be present on the cell surface (SP II signal peptide) with the rest of the proteins predicted to be periplasmic (SP I signal peptide) (**Table 5.1**), the cellular localization study revealed that all of them are located inside of the cell (**Figure 5.4**). In comparison, it was found that the endo-β-N-acetylglucosaminidase BT1044^{GH18} (SP II) is surface localized. These results suggested a major role for this enzyme in the complex N-glycan utilization pathway. Surface localization indicates that BT1044^{GH18} is cleaving the complex N-glycan structures off the protein backbone prior to their import into the cell for further processing, suggesting that this is the initial step of N-glycan utilization pathway.

Considering that majority of complex N-glycans are heavily sialylated, sialic acid removal is a major step required for their utilization. It was previously shown that BT0455^{GH33} is the only sialidase encoded by *B. thetaiotaomicron*, thus its importance to the N-glycan utilization in vitro is undeniable (Park *et al.,* 2013). However, it is important to note that it is possible that in a competitive environment in vivo, *B. thetaiotaomicron* may be able to forage on N-glycan structures that were de-sialylated by the other members of the human gut microbiota. When combined, cellular localization and the whole cell assay data presented in this chapter revealed that the sialidase BT0455^{GH33} may be both surface and periplasm localized, which is further supported by the ambiguous predicted SPI/SPII signal peptide. In support of this claim, the release of sialic acid from complex N-glycan structures was observed in both whole cell and Proteinase K-treated whole cell (after 24h incubation) assays suggesting that N-glycan structures could be de-sialylated both before and after the import into the cell.

Apart from much longer lag phases compared to the wild type, the deletion of the single genes from the BT0455-0461 locus did not result in any significant growth defects in vitro when *B. thetaiotaomicron* was grown on α_1-AGP N-glycans. However, considering how complex N-glycan structures can be, these genetic perturbations may affect the growth rate of the deletion strains on different source and structures of N-glycans. Furthermore, considering the broad activities and overlapping specificities of the BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases discussed in **Chapter 4**, these enzymes appear to be redundant since they could all be individually knocked out without drastically affecting the growth of *B. thetaiotaomicron* cells. For future work, a triple deletion mutant should be generated to investigate the requirement of all three BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} β-hexosaminidases for the growth of *B. thetaiotaomicron* on complex N-glycans. It is important to note that the activity of BT0461^{GH2} may also be redundant based on the growth assays performed in vitro because several other uncharacterised GH2-family galactosidases were identified in the RNA seq data analysis as being upregulated, albeit to a very low level in this non-competitive environment (see Supplemental Table 3.1). The work presented in this thesis was done in a non-competitive environment in vitro. However, when results are combined, they suggest that the enzymes encoded by BT0455-0461 locus could have specialised metabolically to efficiently function in N-glycan utilization in a competitive environment *in vivo*. This could be tested by setting up a competitive growth assay of various deletion mutant strains of *B. thetaiotaomicron* in the presence of other members of gut microbiota, such as *B. massiliensis*, *B. uniformis* and *B. fragilis* that were previously shown to be capable of utilising complex α_{1-}AGP N-glycans (**Chapter 3**).

A BT1044^{GH18} endo-β-N-acetylglucosaminidase was identified as a key enzyme *B. thetaiotaomicron* requires to initiate the utilisation of the complex biantennary α_1-AGP N-glycans. Interestingly, several other GH18 family enzymes are encoded in BT1032-1051 PUL, however only the deletion of BT1044^{GH18} was shown to have a profound defect (Dr. Lucy Crouch, unpublished data). These results indicate that the other GH18 enzymes could be specialised in targeting different types of N-glycan structures, such as tetraantennary.

When these results were combined, it allowed us to propose a model of N-glycan utilization by *B. thetaiotaomicron* in which N-glycan structure is degraded sequentially by a cooperative activity of 15 enzymes encoded by the BT0455-0461, BT0506-0507 and BT1032-1051 loci. This model has been tested *in vitro* where all enzymes were incubated with α_1-AGP and complete degradation of the N-glycan structures was observed (Dr. Lucy Crouch, unpublished data). The activities of these enzymes were also screened against several other sources of complex N-glycans, such as fetuin and transferrin, where a similar pattern of degradation was observed (**Chapter 4**).

The enzymes important for the complex N-glycan degradation by *B. thetaiotaomicron* were identified as the glycoside hydrolases encoded by the BT0455-0461 locus, broad-acting β-hexosaminidase BT0506^{GH20} and endo-β-N-acetylglucosaminidase BT1044^{GH18}. When these were screened against IMG and PULDB databases, homologous enzymes in ten other members of *Bacteroides* were identified. Interestingly, this system appears to be fully conserved in *B. faecis*, suggesting the same complex N-glycan specificity for these two species of bacteria. It was also observed that the homologues of the sialidase BT0455^{GH33}, β-hexosaminidases BT0459^{GH20}, BT0506^{GH20} and BT0460^{GH20} and β-galactosidase BT0461^{GH2} are conserved in all ten species of *Bacteroides*, suggesting an importance of these enzymes for the N-glycan utilization. With an exception of *B. ovatus*, the rest of the identified *Bacteroides* have four GH20-family enzymes, which indicates that the number of GH20s present in this intricate N-glycan degradation apparatus of *B. thetaiotaomicron* is not a special trait. However, with an exception of *B. faecis*, the homologous PULs identified in the rest of *Bacteroides* encode predicted SusC/D pairs. This TonB-dependent outer membrane complex is not present in the *B. thetaiotaomicron* BT0455-0461 locus, which makes it quite special considering SusC/D pair is a hallmark of a typical glycan utilization locus. Surprisingly, three members of *Bacteroides, B. dorei, B. massiliensis and B. caccae,* do not possess any GH18-family endo-glycosidases, suggesting they must employ a different strategy to cleave N-glycan structures off the protein backbones, potentially by utilising PNGaseF-like enzymes. In overall, these findings suggest that the variant of this intricate N-glycan utilization apparatus identified in *B. thetaiotaomicron* could be employed by the majority of *Bacteroides* present in the human gut. However, this remains to be tested *in vitro*.

Chapter 6: Final discussion

The human gut microbiota is a glycan-rich environment that is metabolically well-equipped to utilise a diverse range of complex dietary and host-derived glycan sources present in the human gut, such as N-glycoproteins. N-linked glycosylation is a major post-translational modification of proteins produced by mammalians, plants, bacteria, insects and fungi. The structures produced by these organisms can be incredibly complex and based on their basic composition are classified into three main groups: complex, high mannose and hybrid structures (Staudacher *et al.,* 2015). The N-glycans are prominent in human gut in a form of dietary and host-derived N-glycoproteins (e.g. IgA), and judging by their structural complexity, they would require a cooperative activity of numerous enzymes to fully degrade and utilise. Despite their prevalence in the human gut, the investigation of the N-glycan degradation pathways has been mainly focused on pathogenic bacteria, such as an opportunistic pathogen *B. fragilis*, where the capacity to utilize N-glycans has been linked to the increased rate of survival in the extra-intestinal niche (Cao *et al.,* 2014). However, the knowledge of metabolic capacity of the commensal members of human gut microbiota to utilise these highly complex N-glycan structures was lacking. This chapter will review the work presented in this thesis and focus on how it contributed to the understanding of the complex N-glycan utilisation pathways of commensal gut microbes by proposing a model of degradation by *B. thetaiotaomicron*.

Members of Bacteroidetes, such as *B. thetaiotaomicron*, have adapted to metabolise extremely complex glycan structures present in the human gut by expanding their genetic repertoire of enzymes, each specialised in targeting specific glycan structures, such as plant pectin rhamnogalacturonan-II (RG-II; Ndeh *et al.,* 2018). This ability to metabolise highly complex glycan structures was observed in **Chapter 3**, where several prominent members of human gut microbiota were found to be capable of utilising complex N-glycans as a sole carbon source in vitro. These include wild type *B. thetaiotaomicron*, *B. fragilis*, *B. massiliensis* and *B. uniformis*, all of which displayed a relatively efficient capacity to metabolise the levels of complex biantennary N-glycan structures present on the bovine α_1-Acid Glycoprotein used in this study (**Figure 3.1**). This level of metabolism could suggest that N-glycans are not their nutrient of choice and they could have evolved to utilise these structures as a last resort in times of starvation. Because α_1-AGP is not usually found in the human gut environment, it is also plausible that they prefer more complex N-glycan structures actually present in the human gut, such as tetraantennary immunoglobulin A (IgA) structures. Considering nutrient cross-feeding has been previously observed among the members of gut microbiota (Luis *et al.,* 2018; Seth *et al*., 2014), another explanation for such reduced maximum growth levels of these species would be that the members of *Bacteroides* actually work as a community to degrade these highly complex structures more efficiently in vivo. It is probable that growth on complex N-glycans is more metabolically taxing to the bacterium than growth on simple monosaccharides such as GlcNAc or glucose, because it is dependent on the production of many factors (e.g. enzymes) required to utilise these glycoconjugates. In overall, the metabolic ability to utilise complex N-glycans appears to be a favourable trait in this ecosystem.

It was also observed that *B. thetaiotaomicron* had a metabolic capacity to efficient utilise a range of different N-glycoproteins, such as transferrin, fetuin and α_1-AGP (**Figure 3.2**). These structures contain a variation of complex biantennary, triantennary and tetraantennary N-glycan structures, suggesting *B. thetaiotaomicron* must possess a number of specialised enzymes capable of accommodating and degrading such highly complex glycan structures. Transcriptomic data analysis helped to identify this complex N-glycan utilization apparatus in *B. thetaiotaomicron*. Based on these data (**Figure 3.7.1**), the main loci upregulated on α_1-AGP N-glycan were found to be BT0455-0461, BT0506-0507 and BT1032-1051.

As suspected, numerous enzymes are required to fully degrade complex N-glycan structures. In **Chapter 4**, the enzymes encoded by the BT0455-0461 and BT0506-0507 loci were characterised and their biological roles in complex N-glycan degradation were identified. The utilisation of liberated complex N-glycan structures is achieved by a sequential exo-mode of degradation where the cooperative activity of multiple linkage-specific glycoside hydrolases is required. These included a broad-acting α-sialidase BT0455^{GH33} and β1,4-galactosidase BT0461^{GH2}. *B. thetaiotaomicron* appears to require a cooperative activity of four different β-hexosaminidases for the complex N-glycan degradation, BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20}. Although these enzymes have overlapping specificities for simple disaccharides and some of the glycosidic linkages found on biantennary complex N-glycan structures (e.g. GlcNAc-β1,2-), slight linkage preferences were observed in degradation of complex bisecting, triantennary and tetraantennary N-glycan structures that contain a broader spectrum of GH20 target linkages, such as GlcNAc-β1,2/4/6 and bisecting GlcNAc-β1,4-Man (**Figure 4.42**). The requirement for such an array of GH20-family enzymes could be explained by the variability of linkages present on complex N-glycan structures found in mucosal and luminal niches of the human gut. For example, GlcNAc-β1,2/4/6 and bisecting GlcNAc-β1,4-Man linkages present on a heavily N-glycosylated host-derived IgA antibody, a major part of the innate immune system of the gut, would require a number of specialised GH20-family enzymes to degrade (**Figure 6.1**; Huang *et al.,* 2015). Four GH20s may also be needed due to the requirement for a quick, metabolically efficient degradation of these structures in vivo. It was previously observed that the initial targeting of a bisecting GlcNAc linkage by a multimodular GH20-family *exo*-β-D-N-acetylglucosaminidase StrH in *S. pneumoniae* limited the ability of this enzyme to target antennaery GlcNAc linkages of the complex N-glycan structure, reducing the N-glycan degradation efficiency (Pulvinage *et al.,* 2011).

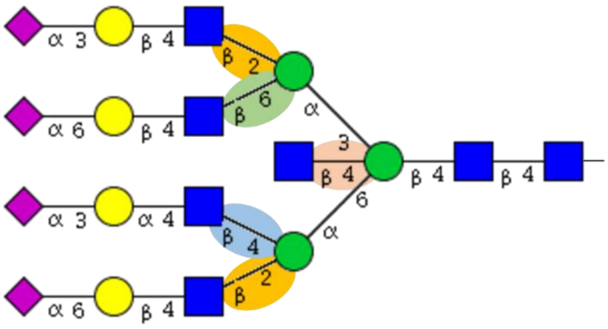


Figure 6.1: An example of a complex N-glycan structure found on human colostrum IgA.
The glycosidic linkages the four β-hexosaminidases BT0456^{GH20}, BT0460^{GH20}, BT0459^{GH20} and BT0506^{GH20} could target (once sialic acid and galactose are removed) are highlighted and color-coded. Structure was built using a GlycanBuilder tool based on the linkage information obtained from Huang *et al.,* 2015.

Carbohydrate binding modules (CBMs) are an integral part of numerous glycoside hydrolases (GHs) where they contribute to the glycan degradation by enhancing substrate-to-enzyme interactions. The classical function of the CBM domains in the endo-acting enzymes is to tether it to an insoluble target polysaccharide via a flexible linker domain whereas in exo-glycosidases the CBM domains are more closely associated with more structured, rigid linker domains, making their role less clear (Gilbert et al., 2010). Recently, a novel CBM domain was identified in a *T. forsythia* sialidase NanH, where the CBM was shown to be involved in glycosidase-ligand binding of host-derived sialoglycans (Frey *et al.,* 2018). In **Chapter 4**, the characterisation of β-hexosaminidases BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} also identified novel CBM domains that specifically bind to complex N-glycans derived from α_1-AGP (**Figure 4.54**). To our knowledge, it is the first time the F5/8 type_C domains present in BT0459^{GH20} and BT0460^{GH20}, and PA14 domain present in BT0506^{GH20} were shown to be involved in carbohydrate binding and it was observed that the overall domain structure of these GH20s is quite unique and present predominantly in *Bacteroides*. The binding to N-glycan structures rather than monosaccharides (ITC, **Figure 4.53**) and disaccharides found on these structures suggested that these CBMs target the whole complex N-glycan structure. Considering all of the upregulated GH20s have similar domain composition, it is plausible that this domain structure is specialised to target N-glycans. The structural characterization of BT0459^{GH20} allowed for a model for substrate binding to be proposed, where substrate binding to the CBM32-like F5/F8 Type C domain induces a movement of the CBM and the flexible linker domain to position the substrate in the active site, possibly enhancing the degradation efficiency or modulating substrate preference (**Figure 4.51**). Although the removal of the CBM domain appeared to decrease the catalytic efficiency of BT0459^{GH20} against simple disaccharides (**Figure 4.58**), it appeared to have no impact on BT0459^{GH20} activity against complex biantennary α_1-AGP structures. However, it may be required for efficient degradation of triantennary, tetraantennary and bisecting complex N-glycan linkages (**Figure 6.2**). This model remains to be tested by investigating the structure of a substrate-bound BT0459^{GH20}. Interestingly, majority of the B*. thetaiotaomicron* putative and characterised GH20-family enzymes, including the BT0460^{GH20}, contain the same predicted CBM domain (F5/8_type_C) as BT0459^{GH20}, suggesting they all could employ similar binding strategy.

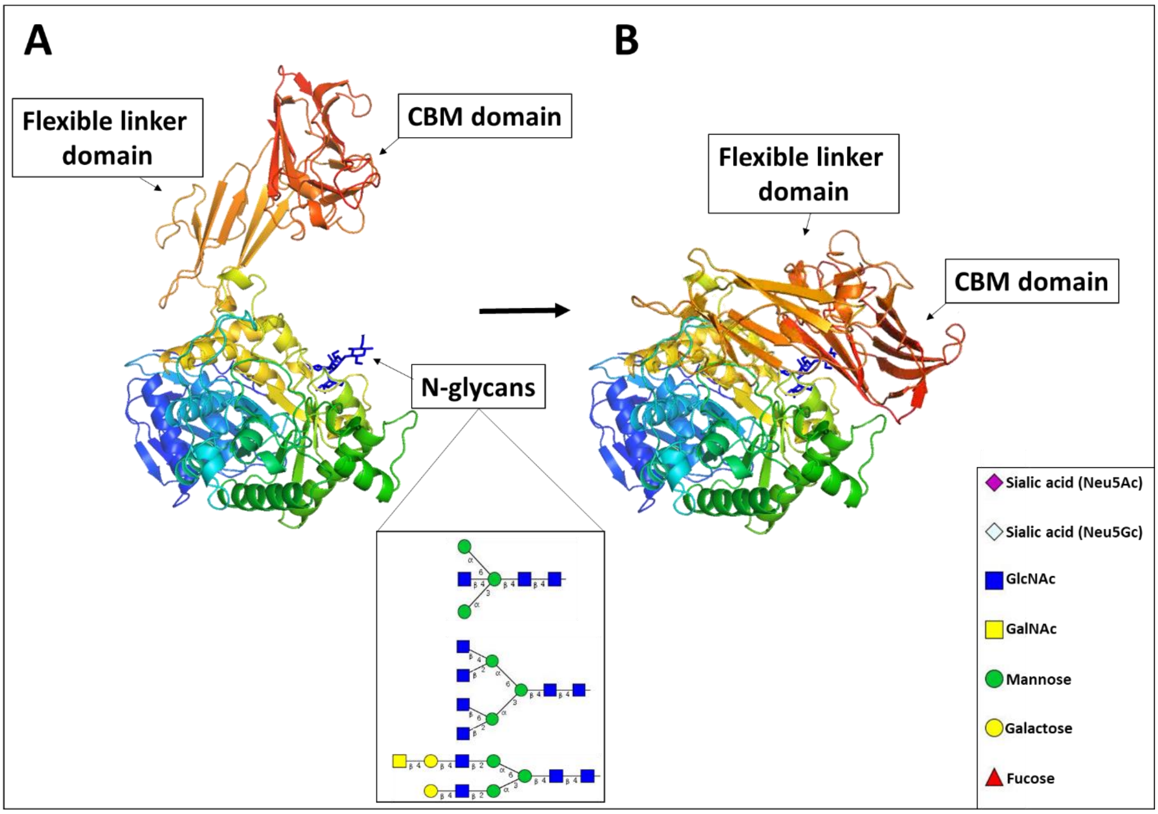


Figure 6.2: The proposed model of N-glycan binding by BT0459^{GH20}. A) Open state of BT0459^{GH20}. CBM domain and the flexible linker domain are in 'arm-like' position above the open binding site of the catalytic GH20 domain. Model N-glycan structure is colored in blue. Examples of linkages present on N-glycan structures that BT0459^{GH20} was found to be active against are displayed. **B)** Proposed closed state of an N-glycan-bound BT0459^{GH20}. X-ray structure of BT0459^{GH20} was coloured rainbow from red (C-terminus) to blue (N-terminus).

On the basis of the results presented in this thesis and in the context of our current knowledge, a model of complex N-glycan utilization by *B. thetaiotaomicron* was proposed (**Chapter 5**), where the degradation is achieved by a cooperative activity of numerous N-glycan degrading enzymes encoded by multiple loci (**Figure 5.8**). The characterised machinery required for the complete complex N-glycan degradation is encoded by three discrete loci. Combined, they encode 15 specialised enzymes, three predicted SusC/D-like pairs of outer membrane transporter systems, two predicted surface glycan-binding proteins (SGBPs) and two predicted regulator proteins. Based on the deletion mutant growth results, extracellular endo-glycosidase BT1044^{GH18} is a key enzyme required to initiate complex N-glycan degradation by releasing it from the protein backbone **(Figure 5.7**). The liberated N-glycan structure is then imported into the cell where it is sequentially degraded by a number of exo-glycosidases. Although the full N-glycan degradation pathway has not been described in any bacterium yet, this degradation pathway appears to be very different from a mode of action of the pathogenic microbe *S. pneumonia*e where the complex N-glycan structures are first depolymerised down to the $Man_3GlcNAc_2$ core by a number of extracellular glycosidases before it is cleaved off the protein backbone and imported into the cell (Burnaugh *et al.,* 2008; Robb *et al.,* 2017). The important difference observed between the *S. pneumoniae* N-glycan degradation apparatus and the *B. thetaiotaomicron* apparatus described in this thesis, is the number and the redundancy of the enzymes required. In *S. pneumoniae*, deglycosylation of complex N-glycans down to the $Man_3GlcNAc_2$ core depends on an exo-sialidase (NanA), an exo-galactosidase (BgaA) and an exo-N-acetylglucosaminidase (StrH). The deletion of any of these enzymes resulted in a severe retardation of *S. pneumoniae* growth on complex N-glycans. However, the main difference observed between these two bacteria is the requirement for an GH18 endo-β-N-acetylglucosaminidase activity. In *S. pneumoniae*, it was observed that the deletion of this endo-glycosidase (EndoD) has no substancial defect when it was grown on complex N-glycans as a sole carbon source (Robb *et al*., 2017). In comparison, the deletion of the GH18 endo-glycosidase BT1044^{GH18} displayed a severe growth defect in *B. thetaiotaomicron* (**Figure 5.8**). These differences suggest different strategies of complex N-glycan utilization could be employed by pathogenic and symbiotic bacteria.

Furthermore, the identification of genes homologous to BT0455-0461, BT0506^{GH20} and BT1044^{GH18} in ten prominent members of *Bacteroides spp.* suggests that the variant of this proposed model of the complex N-glycan degradation may also be widely utilised by thes human gut *Bacteroide*s.

Combined, the work presented in this thesis contributes to the knowledge of complex N-glycan degradation by prominent, symbiotic members of human gut microbiota by identifying and characterising the key enzymes of the N-glycan degradation apparatus in *B. thetaiotaomicron*. Understanding how complex N-glycans are utilised by the HGM helps us to better appreciate the metabolic potential of this ecosystem and contributes to our understanding of its role in human health. Classically, deglycosylation of complex N-glycan structures has been considered a virulence factor, strongly linked with an increased risk of inflammation by various human pathogens (Cao *et al.,* 2014; Robb *et al.,* 2017). However, considering a commensal generalist *B. thetaiotaomicron* is capable of foraging on these complex N-glycans structures that are prominent in the human mucosal and luminal niches in a form of dietary and host-derived glycoproteins, such as IgA and mucins, this process may also have a different role. For example, by utilising N-glycans present on the glycocalyx of sloughed epithelial cells, *B. thetaiotaomicron* could contribute to the efficiency of the gastric epithelial cell turnover. Epithelial cell turnover is a natural, balanced process required to maintain gut homeostasis. Moreover, removal of the relatively low numbers of N-glycan structures present on predominantly O-glycosylated mucins would enable for a better access to these structures for other specialised members of gut microbiota. This could potentially benefit the inter-microbial relationships of the sub-communities formed by specialised members of HGM. However, the precise role of complex N-glycan degradation by commensal human gut microbes in human health remains to be determined.

6.1. Future work

Research into N-glycan degradation by the members of human gut microbiota is still in early stages. The successful identification of complex N-glycan degradation apparatus in *B. thetaiotaomicron* has opened a new avenue of studying N-glycan degradation by commensal gut microbes. The successful cloning, expression and purification of N-glycan binding proteins and degrading enzymes *B. thetaiotaomicron* requires to utilise dietary and host-derived heavily N-glycosylated glycoproteins present in the human gut allowed us to propose a model for complex N-glycan degradation. Questions which remain include, what role do the rest of the upregulated enzymes play? Can this model of degradation be applied to more complex N-glycan structure degradation? Biochemical characterisation of all upregulated enzymes would provide a clearer picture of N-glycan degradation by this bacterium.

Regarding the four GH20 glycosidases; how exactly do the mechanisms of β-hexosaminidases BT0456^{GH20}, BT0459^{GH20}, BT0460^{GH20} and BT0506^{GH20} N-glycan binding and degradation differ? Do the CBM domains contribute to the N-glycan degradation efficiency? What binding residues are necessary for these enzymes to recognise the substrate? Crystal structures of ligand-bound catalytic mutants of these enzymes would be invaluable in answering these questions.

Furthermore, what is the mode of action of the broad-acting sialidase BT0455^{GH33}? The cellular localization and whole cell assay data suggests it may be both surface and periplasm localised. Considering BT0455^{GH33} is the only sialidase encoded by *B. thetaiotaomicron*, investigating its mechanism of action is of utmost importance to fully understand N-glycan degradation apparatus of this bacterium.

Finally, does the mechanism of complex N-glycan degradation observed in *B. thetaiotaomicron* apply to other members of *Bacteroides* present in the human gut? The PUL composition comparison revealed homologous PULs in numerous other *Bacteroides*. Dissecting these PULs to investigate whether the encoded homologous apparatus is also utilised to forage on complex N-glycans would provide further insight into N-glycan degradation pathways of the dominant members of HGM.

The human gut microbiota composition heavily depends on the availability of different nutrient sources in the gut, in turn influencing the host-microbiota symbiosis. The ability to utilise N-glycans by the members of human gut microbiota could be important in times of starvation and may play an important role in human health. The work presented in this thesis addressed some of the questions on how commensal gut microbes utilise complex N-glycans but many questions, such as whether and how it contributes to the host-microbe symbiosis, remain to be answered.

References

Abed, J., Maalouf, N., Parhi, L., Chaushu, S., Mandelboim, O., Bachrach, G. 2017. Tumor targetting by Fusobacterium nucleatum: A pilot study and future perspectives. *Front. Cell. Infect. Microbiol.* 7: 295.

Abou Hachem, M., E. Nordberg Karlsson, E. Bartonek-Roxa, S. Raghothama, P. J. Simpson, H. J. Gilbert, M. P. Williamson, and O. Holst. 2000. 'Carbohydrate-binding modules from a thermostable Rhodothermus marinus xylanase: cloning, expression and binding studies', *Biochem J*, 345 Pt 1: 53-60.

Adesioye, F. A., T. P. Makhalanyane, P. Biely, and D. A. Cowan. 2016. 'Phylogeny, classification and metagenomic bioprospecting of microbial acetyl xylan esterases', *Enzyme Microb Technol*, 93-94: 79-91.

Agus, A., J. Denizot, J. Thevenot, M. Martinez-Medina, S. Massier, P. Sauvanet, A. Bernalier-Donadille, S. Denis, P. Hofman, R. Bonnet, E. Billard, and N. Barnich. 2016. 'Western diet induces a shift in microbiota composition enhancing susceptibility to Adherent-Invasive E. coli infection and intestinal inflammation', *Sci Rep*, 6: 19032.

Alberts, B., Johnson A., Lewis J., *et al*. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Fractionation of Cells.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res*, 25: 3389-402.

Anderson, K. L., and A. A. Salyers. 1989. 'Biochemical evidence that starch breakdown by Bacteroides thetaiotaomicron involves outer membrane starch-binding sites and periplasmic starch-degrading enzymes', *J Bacteriol*, 171: 3192-8.

Arnal, G., D. W. Cockburn, H. Brumer and N. M. Koropatkin 2018. "Structural basis for the flexible recognition of α-glucan substrates by Bacteroides thetaiotaomicron SusG." Protein Science 27(6): 1093-1101.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, H. I. T. Consortium Meta, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. Melo Minardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, and P. Bork. 2011. 'Enterotypes of the human gut microbiome', *Nature*, 473: 174-80.

Backhed, F., R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon. 2005. 'Host-bacterial mutualism in the human intestine', *Science*, 307: 1915-20.

Backhed, F., J. Roswall, Y. Peng, Q. Feng, H. Jia, P. Kovatcheva-Datchary, Y. Li, Y. Xia, H. Xie, H. Zhong, M. T. Khan, J. Zhang, J. Li, L. Xiao, J. Al-Aama, D. Zhang, Y. S. Lee, D. Kotowska, C. Colding, V. Tremaroli, Y. Yin, S. Bergman, X. Xu, L. Madsen, K. Kristiansen, J. Dahlgren, and J. Wang. 2015. 'Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life', *Cell Host Microbe*, 17: 852.

Bagenholm, V., S. K. Reddy, H. Bouraoui, J. Morrill, E. Kulcinskaja, C. M. Bahr, O. Aurelius, T. Rogers, Y. Xiao, D. T. Logan, E. C. Martens, N. M. Koropatkin, and H. Stalbrand. 2017. 'Galactomannan Catabolism Conferred by a Polysaccharide Utilization Locus of Bacteroides ovatus: ENZYME SYNERGY AND CRYSTAL STRUCTURE OF A beta-MANNANASE', *J Biol Chem*, 292: 229-43.

Bakir, M. A., Kitahara, M., Sakamoto, M., Matsumoto, M., Benno, Y. 2006. *Bacteroides finegoldii* sp. nov., isolated from human faeces. *Int.J.Syst.Evol.Microbiol.* 56 : 931-935 .

Barabote, R. D., Xie, G., Leu, D. H., Normand, P., Necsulea, A., Daubin, V., … Berry, A. M. 2009. Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations. *Genome Research*, *19*(6), 1033–1043.

Barboza, M., J. Pinzon, S. Wickramasinghe, J. W. Froehlich, I. Moeller, J. T. Smilowitz, L. R. Ruhaak, J. Huang, B. Lonnerdal, J. B. German, J. F. Medrano, B. C. Weimer, and C. B. Lebrilla. 2012. 'Glycosylation of human milk lactoferrin exhibits dynamic changes during early lactation enhancing its role in pathogenic bacteria-host interactions', *Mol Cell Proteomics*, 11: M111 015248.

Berlemont, R., and A. C. Martiny. 2016. 'Glycoside Hydrolases across Environmental Microbial Communities', *PLoS Comput Biol*, 12: e1005300.

Bermingham, M. L., M. Colombo, S. J. McGurnaghan, L. A. K. Blackbourn, F. Vuckovic, M. Pucic Bakovic, I. Trbojevic-Akmacic, G. Lauc, F. Agakov, A. S. Agakova, C. Hayward, L. Klaric, C. N. A. Palmer, J. R. Petrie, J. Chalmers, A. Collier, F. Green, R. S. Lindsay, S. Macrury, J. A. McKnight, A. W. Patrick, S. Thekkepat, O. Gornik, P. M. McKeigue, H. M. Colhoun, and Sdrn Type 1 Bioresource Investigators. 2018. 'N-Glycan Profile and Kidney Disease in Type 1 Diabetes', *Diabetes Care*, 41: 79-87.

Bessman, N. J., and G. F. Sonnenberg. 2016. 'Emerging roles for antigen presentation in establishing host-microbiome symbiosis', *Immunol Rev*, 272: 139-50.

Bhattacharya, T., T. S. Ghosh, and S. S. Mande. 2015. 'Global Profiling of Carbohydrate Active Enzymes in Human Gut Microbiome', *PLoS One*, 10: e0142038.

Bhavanandan, V. P., N. J. Ringler, and D. C. Gowda. 1998. 'Identification of the glycosidically bound sialic acid in mucin glycoproteins that reacts as "free sialic acid" in the Warren assay', *Glycobiology*, 8: 1077-86.

Bieberich, E. 2014. 'Synthesis, Processing, and Function of N-glycans in N-glycoproteins', *Adv Neurobiol*, 9: 47-70.

Bissaro, B., P. Monsan, R. Faure, and M. J. O'Donohue. 2015. 'Glycosynthesis in a waterworld: new insight into the molecular basis of transglycosylation in retaining glycoside hydrolases', *Biochem J*, 467: 17-35.

Bjursell, M. K., E. C. Martens, and J. I. Gordon. 2006. 'Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period', *J Biol Chem*, 281: 36269-79.

Blacher, E., M. Levy, E. Tatirovsky, and E. Elinav. 2017. 'Microbiome-Modulated Metabolites at the Interface of Host Immunity', *J Immunol*, 198: 572-80.

Bolam, D. N., and N. M. Koropatkin. 2012. 'Glycan recognition by the Bacteroidetes Sus-like systems', *Curr Opin Struct Biol*, 22: 563-9.

Bondt, A., Nicolardi, S., Jansen, B. C., Kuijper, T. M., Hazes, J. M. W., van der Burgt, Y. E. M., … Dolhain, R. J. E. M. 2017. IgA *N*- and *O*-glycosylation profiling reveals no association with the pregnancy-related improvement in rheumatoid arthritis. *Arthritis Research & Therapy*, *19*, 160.

Burnaugh, A. M., L. J. Frantz, and S. J. King. 2008. 'Growth of Streptococcus pneumoniae on human glycoconjugates is dependent upon the sequential activity of bacterial exoglycosidases', *J Bacteriol*, 190: 221-30.

Cameron, E. A., K. J. Kwiatkowski, B. H. Lee, B. R. Hamaker, N. M. Koropatkin, and E. C. Martens. 2014. 'Multifunctional nutrient-binding proteins adapt human symbiotic bacteria for glycan competition in the gut by separately promoting enhanced sensing and catalysis', *MBio*, 5: e01441-14.

Cameron, E. A., M. A. Maynard, C. J. Smith, T. J. Smith, N. M. Koropatkin, and E. C. Martens. 2012. 'Multidomain Carbohydrate-binding Proteins Involved in Bacteroides thetaiotaomicron Starch Metabolism', *J Biol Chem*, 287: 34614-25.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. 2009. 'The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics', *Nucleic Acids Res*, 37: D233-8.

Cao, Y., Rocha, E. R., & Smith, C. J. 2014. Efficient utilization of complex N-linked glycans is a selective advantage for Bacteroides fragilis in extraintestinal infections. Proceedings of the National Academy of Sciences of the United States of America, 111(35), 12901–12906.

Carding, S., K. Verbeke, D. T. Vipond, B. M. Corfe, and L. J. Owen. 2015. 'Dysbiosis of the gut microbiota in disease', *Microb Ecol Health Dis*, 26: 26191.

Cartmell, A., E. C. Lowe, A. Baslé, S. J. Firbank, D. A. Ndeh, H. Murray, N. Terrapon, V. Lombard, B. Henrissat, J. E. Turnbull, M. Czjzek, H. J. Gilbert and D. N. Bolam 2017. "How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans." *Proceedings of the National Academy of Sciences* 114(27): 7037-7042.

Castellani, A., and A. J. Chalmers. 1919. *Manual of tropical medicine*, 3rd ed., p. 959-960. Williams Wood & Co., New York.

Cerf-Bensussan, N., and V. Gaboriau-Routhiau. 2010. 'The immune system and the gut microbiota: friends or foes?', *Nat Rev Immunol*, 10: 735-44.

Chen, H., Z. Deng, C. Huang, H. Wu, X. Zhao, and Y. Li. 2017. 'Mass spectrometric profiling reveals association of N-glycan patterns with epithelial ovarian cancer progression', *Tumour Biol*, 39: 1010428317716249.

Cockburn, D. W., and N. M. Koropatkin. 2016. 'Polysaccharide Degradation by the Intestinal Microbiota and Its Influence on Human Health and Disease', *J Mol Biol*, 428: 3230-52.

Cockburn, D. W., C. Suh, K. P. Medina, R. M. Duvall, Z. Wawrzak, B. Henrissat, and N. M. Koropatkin. 2018. 'Novel carbohydrate binding modules in the surface anchored alpha-amylase of Eubacterium rectale provide a molecular rationale for the range of starches used by this organism in the human gut', *Mol Microbiol*, 107: 249-64.

Collings, D.A. 2013. 'Subcellular localization of transciently expressed fluorescent fusion proteins'. Methods Mol Biol 1069:227-58.

Corfield, A.P., Donapaty, S.R., Carrington, S.D., *et al.* 2005. 'Identification of 9-O-acetyl-N-acetylneuraminic acid in normal canine pre-ocular tear film secreted mucins and its depletion in *Keratoconjunctivitis sicca'* Glycoconjugate Journal 22, 409-416.

Cuskin, F., E. C. Lowe, M. J. Temple, Y. Zhu, E. Cameron, N. A. Pudlo, N. T. Porter, K. Urs, A. J. Thompson, A. Cartmell, A. Rogowski, B. S. Hamilton, R. Chen, T. J. Tolbert, K. Piens, D. Bracke, W. Vervecken, Z. Hakki, G. Speciale, J. L. Munoz-Munoz, A. Day, M. J. Pena, R. McLean, M. D. Suits, A. B. Boraston, T. Atherly, C. J. Ziemer, S. J. Williams, G. J. Davies, D. W. Abbott, E. C. Martens, and H. J. Gilbert. 2015. 'Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism', *Nature*, 517: 165-69.

Dall'Acqua, W. and P. Carter 2000. "Substrate-assisted catalysis: molecular basis and biological significance." Protein Sci 9(1): 1-9.

Davies, G. and B. Henrissat 1995. "Structures and mechanisms of glycosyl hydrolases." Structure 3(9): 853-859.

Dell, A., A. Galadari, F. Sastre, and P. Hitchen. 2010. 'Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes', *Int J Microbiol*, 2010: 148178.

Derrien, M., E. E. Vaughan, C. M. Plugge and W. M. de Vos 2004. 'Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium.' International Journal of Systematic and Evolutionary Microbiology 54(5): 1469-1476.

Desai, M. S., A. M. Seekatz, N. M. Koropatkin, N. Kamada, C. A. Hickey, M. Wolter, N. A. Pudlo, S. Kitamoto, N. Terrapon, A. Muller, V. B. Young, B. Henrissat, P. Wilmes, T. S. Stappenbeck, G.

Nunez, and E. C. Martens. 2016. 'A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility', *Cell*, 167: 1339-53 e21.

Distaso, A. 1912. Contribution à l'étude sur l'intoxication intestinale. Zentralbl.Bakteriol.Hyg.I Abt. 62 : 433-469 .

Donaldson, D. S., and N. A. Mabbott. 2016. 'The influence of the commensal and pathogenic gut microbiota on prion disease pathogenesis', *J Gen Virol*, 97: 1725-38.

Donaldson, G. P., M. S. Ladinsky, K. B. Yu, J. G. Sanders, B. B. Yoo, W. C. Chou, M. E. Conner, A. M. Earl, R. Knight, P. J. Bjorkman, and S. K. Mazmanian. 2018. 'Gut microbiota utilize immunoglobulin A for mucosal colonization', *Science*.

Earle, K. A., G. Billings, M. Sigal, J. S. Lichtman, G. C. Hansson, J. E. Elias, M. R. Amieva, K. C. Huang, and J. L. Sonnenburg. 2015. 'Quantitative Imaging of Gut Microbiota Spatial Organization', *Cell Host Microbe*, 18: 478-88.

Eggerth, A. H., and B. H. Gagnon. 1933. 'The Bacteroides of human feces', *J. Bacteriol*., 25, 389-413.

El Kaoutari, A., F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat. 2013. 'The abundance and variety of carbohydrate-active enzymes in the human gut microbiota', *Nat Rev Microbiol*, 11: 497-504.

Fenner, L., V. Roux, M. N. Mallet, and D. Raoult. 2005. Bacteroides massiliensis sp. nov., isolated from blood culture of a newborn. *Int. J. Syst. Evol.*

Microbiol. 55:1335–1337

Ficko-Blean, E., and A. B. Boraston. 2006. 'The interaction of a carbohydrate-binding module from a Clostridium perfringens N-acetyl-beta-hexosaminidase with its carbohydrate receptor', *J Biol Chem*, 281: 37748-57.

Ficko-Blean, E., Gregg, K. J., Adams, J. J., Hehemann, J.-H., Czjzek, M., Smith, S. P., & Boraston, A. B. 2009. Portrait of an Enzyme, a Complete Structural Analysis of a Multimodular β-N-Acetylglucosaminidase from Clostridium perfringens. The Journal of Biological Chemistry, 284(15), 9876–9884.

Flint, H. J., K. P. Scott, P. Louis, and S. H. Duncan. 2012. 'The role of the gut microbiota in nutrition and health', *Nat Rev Gastroenterol Hepatol*, 9: 577-89.

Foley, M. H., D. W. Cockburn, and N. M. Koropatkin. 2016. 'The Sus operon: a model system for starch uptake by the human gut Bacteroidetes', *Cell Mol Life Sci*, 73: 2603-17.

Foley, M. H., E. C. Martens, and N. M. Koropatkin. 2018. 'SusE facilitates starch uptake independent of starch binding in B. thetaiotaomicron', *Mol Microbiol*.

Franzosa EA, Morgan XC, Segata N, et al. 2014. 'Relating the metatranscriptome and metagenome of the human gut'. Proc Natl Acad Sci U S A. 111(22):E2329-38.

Freeze, H. H., J. X. Chong, M. J. Bamshad, and B. G. Ng. 2014. 'Solving glycosylation disorders: fundamental approaches reveal complicated pathways', *Am J Hum Genet*, 94: 161-75.

Frey, A. M., M. J. Satur, C. Phansopa, J. L. Parker, D. Bradshaw, J. Pratten and G. P. Stafford 2018. 'Evidence for a carbohydrate-binding module (CBM) of *Tannerella forsythia* NanH sialidase, key to interactions at the host–pathogen interface.' Biochemical Journal 475(6): 1159-1176.

Gilbert, H. J., J. P. Knox, and A. B. Boraston. 2013. 'Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules', *Curr Opin Struct Biol*, 23: 669-77.

Gilkes, N. R., R. A. Warren, R. C. Miller, Jr., and D. G. Kilburn. 1988. 'Precise excision of the cellulose binding domains from two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis', *J Biol Chem*, 263: 10401-7.

Glenwright AJ, Pothula KR, Bhamidimarri SP et al. 2017. 'Structural basis for nutrient acquisition by dominant members of the human gut microbiota'. Nature;541:407–11.

Glenwright AJ. 2017. Understanding nutrient transport across the outer membrane by members of the human gut microbiota (Doctoral thesis). ICAMB, Newcastle University.

Goettig, P. 2016. 'Effects of Glycosylation on the Enzymatic Activity and Mechanisms of Proteases', *Int J Mol Sci*, 17.

Ghosh, S., Bandyopadhyay, S., Mukherjee, K. *et al.* Glycoconj J (2007) 24: 17.

Gregg, K. J., R. Finn, D. W. Abbott, and A. B. Boraston. 2008. 'Divergent modes of glycan recognition by a new family of carbohydrate-binding modules', *J Biol Chem*, 283: 12604-13.

Grondin, J. M., K. Tamura, G. Dejean, D. W. Abbott, and H. Brumer. 2017. 'Polysaccharide Utilization Loci: Fueling Microbial Communities', *J Bacteriol*, 199.

Hedemann, M. S., P. K. Theil, and K. E. Bach Knudsen. 2009. 'The thickness of the intestinal mucous layer in the colon of rats fed various sources of non-digestible carbohydrates is positively correlated with the pool of SCFA but negatively correlated with the proportion of butyric acid in digesta', *Br J Nutr*, 102: 117-25.

Henrissat, B., and G. Davies. 1997. 'Structural and sequence-based classification of glycoside hydrolases', *Curr Opin Struct Biol*, 7: 637-44.

Henrissat, B., S. E. Heffron, M. D. Yoder, S. E. Lietzke, and F. Jurnak. 1995. 'Functional implications of structure-based sequence alignment of proteins in the extracellular pectate lyase superfamily', *Plant Physiol*, 107: 963-76.

Herrero, M., V. de Lorenzo, and K. N. Timmis. 1990. 'Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria', *J Bacteriol*, 172: 6557-67.

Herve, C., A. Rogowski, A. W. Blake, S. E. Marcus, H. J. Gilbert, and J. P. Knox. 2010. 'Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects', *Proc Natl Acad Sci U S A*, 107: 15293-8.

Huang, J., Guerrero, A., Parker, E., Strum, J. S., Smilowitz, J. T., German, J. B., & Lebrilla, C. B. 2015. 'Site-specific Glycosylation of Secretory Immunoglobulin A from Human Colostrum'. Journal of Proteome Research, 14(3), 1335–1349.

Huang, Y., and R. Orlando. 2017. 'Kinetics of N-Glycan Release from Human Immunoglobulin G (IgG) by PNGase F: All Glycans Are Not Created Equal', *J Biomol Tech*, 28: 150-57.

Jensen, P. H., D. Kolarich, and N. H. Packer. 2010. 'Mucin-type O-glycosylation--putting the pieces together', *FEBS J*, 277: 81-94.

Jeong, Y. R., S. Y. Kim, Y. S. Park, and G. M. Lee. 2018. 'Simple and Robust N-Glycan Analysis Based on Improved 2-Aminobenzoic Acid Labeling for Recombinant Therapeutic Glycoproteins', *J Pharm Sci*.

Jerabek-Willemsen, M., C. J. Wienken, D. Braun, P. Baaske, and S. Duhr. 2011. 'Molecular interaction studies using microscale thermophoresis', *Assay Drug Dev Technol*, 9: 342-53.

Johansson, M. E., J. K. Gustafsson, K. E. Sjoberg, J. Petersson, L. Holm, H. Sjovall, and G. C. Hansson. 2010. 'Bacteria penetrate the inner mucus layer before inflammation in the dextran sulfate colitis model', *PLoS One*, 5: e12238.

Johansson, M. E., J. M. Larsson, and G. C. Hansson. 2011. 'The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions', *Proc Natl Acad Sci U S A*, 108 Suppl 1: 4659-65.

Johnson J, Moore W, Moore L. 1986. ' *Bacteroides caccae sp.* nov., *Bacteroides merdae* sp. nov., and B*acteroides stercoris* sp. nov. Isolated from Human Feces Int J Syst Evol Microbiol 36(4):499-501

Juge, N., L. Tailford, and C. D. Owen. 2016. 'Sialidases from gut bacteria: a mini-review', *Biochem Soc Trans*, 44: 166-75.

Kahya, H. F., P. W. Andrew, and H. Yesilkaya. 2017. 'Deacetylation of sialic acid by esterases potentiates pneumococcal neuraminidase activity for mucin utilization, colonization and virulence', *PLoS Pathog*, 13: e1006263.

Kelly, W. J., S. C. Leahy, E. Altermann, C. J. Yeoman, J. C. Dunne, Z. Kong, D. M. Pacheco, D. Li, S. J. Noel, C. D. Moon, A. L. Cookson, and G. T. Attwood. 2010. 'The glycobiome of the rumen bacterium Butyrivibrio proteoclasticus B316(T) highlights adaptation to a polysaccharide-rich environment', *PLoS One*, 5: e11942.

King, S. J., K. R. Hippe, and J. N. Weiser. 2006. 'Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by Streptococcus pneumoniae', *Mol Microbiol*, 59: 961-74.

Kobata, A. 2013. 'Exo- and endoglycosidases revisited', Proc Jpn Acad Ser B Phys Biol Sci. 89(3): 97–117.

Koenig, J. E., A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley. 2011. 'Succession of microbial consortia in the developing infant gut microbiome', *Proc Natl Acad Sci U S A*, 108 Suppl 1: 4578-85.

Koropatkin, N. M., E. A. Cameron, and E. C. Martens. 2012. 'How glycan metabolism shapes the human gut microbiota', *Nat Rev Microbiol*, 10: 323-35.

Koropatkin, N. M., and E. C. Martens. 2017. 'Meds Modify Microbiome, Mediating Their Effects', *Cell Metab*, 26: 456-57.

Koropatkin, N. M., E. C. Martens, J. I. Gordon, and T. J. Smith. 2008. 'Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices', *Structure*, 16: 1105-15.

Koropatkin, N., E. C. Martens, J. I. Gordon, and T. J. Smith. 2009. 'Structure of a SusD homologue, BT1043, involved in mucin O-glycan utilization in a prominent human gut symbiont', *Biochemistry*, 48: 1532-42.

Koshland, D. E., Jr., and E. Clarke. 1953. 'Mechanism of hydrolysis of adenosinetriphosphate catalyzed by lobster muscle', *J Biol Chem*, 205: 917-24.

Laemmli, U. K. 1970. 'Cleavage of structural proteins during the assembly of the head of bacteriophage T4', *Nature*, 227: 680-5.

LeBlanc, J. G., C. Milani, G. S. de Giori, F. Sesma, D. van Sinderen, and M. Ventura. 2013. 'Bacteria as vitamin suppliers to their host: a gut microbiota perspective', *Curr Opin Biotechnol*, 24: 160-8.

Leroy, J. G. 2006. 'Congenital disorders of N-glycosylation including diseases associated with O- as well as N-glycosylation defects', *Pediatr Res*, 60: 643-56.

Ley, R. E., F. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight, and J. I. Gordon. 2005. 'Obesity alters gut microbial ecology', *Proc Natl Acad Sci U S A*, 102: 11070-5.

Ley, R. E., C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon. 2008. 'Worlds within worlds: evolution of the vertebrate gut microbiota', *Nat Rev Microbiol*, 6: 776-88.

Li, H., J. P. Limenitakis, T. Fuhrer, M. B. Geuking, M. A. Lawson, M. Wyss, S. Brugiroux, I. Keller, J. A. Macpherson, S. Rupp, B. Stolp, J. V. Stein, B. Stecher, U. Sauer, K. D. McCoy, and A. J. Macpherson. 2015. 'The outer mucus layer hosts a distinct intestinal microbial niche', *Nat Commun*, 6: 8292.

Li, H., J. P. Limenitakis, S. C. Ganal, and A. J. Macpherson. 2015. 'Penetrability of the inner mucus layer: who is out there?', *EMBO Rep*, 16: 127-9.

Liberato, M. V., R. L. Silveira, E. T. Prates, E. A. de Araujo, V. O. Pellegrini, C. M. Camilo, M. A. Kadowaki, O. Neto Mde, A. Popov, M. S. Skaf, and I. Polikarpov. 2016. 'Molecular characterization of a family 5 glycoside hydrolase suggests an induced-fit enzymatic mechanism', *Sci Rep*, 6: 23473.

Liu, T., J. Yan, and Q. Yang. 2012. 'Comparative biochemistry of GH3, GH20 and GH84 beta-N-acetyl-Dhexosaminidases and recent progress in selective inhibitor discovery', *Curr Drug Targets*, 13: 512-25.

Luis, A. S., Briggs, J., Zhang, X., Farnell, B., Ndeh, D., Labourel, A., … Gilbert, H. J. 2018. 'Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic *Bacteroides*.' *Nature Microbiology*, *3*(2), 210–219.

Luis AS, Briggs J, Zhang X et al. 2018. 'Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides.' Nature Microbiology ;3:210–9.

Makki, K., E. C. Deehan, J. Walter, and F. Backhed. 2018. 'The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease', *Cell Host Microbe*, 23: 705-15.

Marchesi, J. R., D. H. Adams, F. Fava, G. D. Hermes, G. M. Hirschfield, G. Hold, M. N. Quraishi, J. Kinross, H. Smidt, K. M. Tuohy, L. V. Thomas, E. G. Zoetendal, and A. Hart. 2016. 'The gut microbiota and host health: a new clinical frontier', *Gut*, 65: 330-9.

Mark, B. L., D. J. Vocadlo, S. Knapp, B. L. Triggs-Raine, S. G. Withers and M. N. James (2001). "Crystallographic evidence for substrate-assisted catalysis in a bacterial beta-hexosaminidase." J Biol Chem 276(13): 10330-10337.

Martens, E. C., H. C. Chiang, and J. I. Gordon. 2008. 'Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont', *Cell Host Microbe*, 4: 447-57.

Martens, E. C., N. M. Koropatkin, T. J. Smith, and J. I. Gordon. 2009. 'Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm', *J Biol Chem*, 284: 24673-7.

Martens, E. C., E. C. Lowe, H. Chiang, N. A. Pudlo, M. Wu, N. P. McNulty, D. W. Abbott, B. Henrissat, H. J. Gilbert, D. N. Bolam, and J. I. Gordon. 2011. 'Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts', *PLoS Biol*, 9: e1001221.

Martens, E. C., R. Roth, J. E. Heuser, and J. I. Gordon. 2009. 'Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont', *J Biol Chem*, 284: 18445-57.

McGuckin, M. A., S. K. Linden, P. Sutton, and T. H. Florin. 2011. 'Mucin dynamics and enteric pathogens', *Nat Rev Microbiol*, 9: 265-78.

McNulty, N. P., M. Wu, A. R. Erickson, C. Pan, B. K. Erickson, E. C. Martens, N. A. Pudlo, B. D. Muegge, B. Henrissat, R. L. Hettich, and J. I. Gordon. 2013. 'Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome', *PLoS Biol*, 11: e1001637.

Mizutani, K., M. Sakka, T. Kimura, and K. Sakka. 2014. 'Essential role of a family-32 carbohydrate-binding module in substrate recognition by Clostridium thermocellum mannanase CtMan5A', *FEBS Lett*, 588: 1726-30.

Moran, A. P., A. Gupta, and L. Joshi. 2011. 'Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract', *Gut*, 60: 1412-25.

Mullis, K. B., and F. A. Faloona. 1987. 'Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction', *Methods Enzymol*, 155: 335-50.

Munkley, J., and D. J. Elliott. 2016. 'Hallmarks of glycosylation in cancer', *Oncotarget*, 7: 35478-89.

Munkley, J., I. G. Mills, and D. J. Elliott. 2016. 'The role of glycans in the development and progression of prostate cancer', *Nat Rev Urol*, 13: 324-33.

Nakano, M., K. Kakehi, M. H. Tsai, and Y. C. Lee. 2004. 'Detailed structural features of glycan chains derived from alpha1-acid glycoproteins of several different animals: the presence of hypersialylated, O-acetylated sialic acids but not disialyl residues', *Glycobiology*, 14: 431-41.

Nakayama-Imaohji, H., M. Ichimura, T. Iwasa, N. Okada, Y. Ohnishi, and T. Kuwahara. 2012. 'Characterization of a gene cluster for sialoglycoconjugate utilization in Bacteroides fragilis', *J Med Invest*, 59: 79-94.

Ndeh, D., Rogowski, A., Cartmell, A., Luis, A. S., Baslé, A., Gray, J., … Gilbert, H. J. 2017. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*, *544*(7648), 65–70.

Ndeh, D. and Gilbert, H.J. 2018. Biochemistry of complex glycan depolymerisation by the human gut microbiota. FEMS Microbiology Reviews, 42(2), 146–164.

Ng, K. M., J. A. Ferreyra, S. K. Higginbottom, J. B. Lynch, P. C. Kashyap, S. Gopinath, N. Naidu, B. Choudhury, B. C. Weimer, D. M. Monack, and J. L. Sonnenburg. 2013. 'Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens', *Nature*, 502: 96-9.

Noinaj N, Guillier M, Barnard TJ, Buchanan SK. 2010. 'TonB-dependent transporters: regulation, structure, and function'. Annu Rev Microbiol. 64:43-60.

Ottman, N., H. Smidt, W. M. de Vos, and C. Belzer. 2012. 'The function of our microbiota: who is out there and what do they do?', *Front Cell Infect Microbiol*, 2: 104.

Ouwerkerk, J. P., W. M. de Vos, and C. Belzer. 2013. 'Glycobiome: bacteria and mucus at the epithelial interface', *Best Pract Res Clin Gastroenterol*, 27: 25-38.

Pacheco, A. R., D. Barile, M. A. Underwood, and D. A. Mills. 2015. 'The impact of the milk glycobiome on the neonate gut microbiota', *Annu Rev Anim Biosci*, 3: 419-45.

Panesar, P. S., S. Kumari, and R. Panesar. 2010. 'Potential Applications of Immobilized beta-Galactosidase in Food Processing Industries', *Enzyme Res*, 2010: 473137.

Park, K. H., M. G. Kim, H. J. Ahn, D. H. Lee, J. H. Kim, Y. W. Kim, and E. J. Woo. 2013. 'Structural and biochemical characterization of the broad substrate specificity of Bacteroides thetaiotaomicron commensal sialidase', *Biochim Biophys Acta*, 1834: 1510-9.

Phansopa, C., R. P. Kozak, L. P. Liew, A. M. Frey, T. Farmilo, J. L. Parker, D. J. Kelly, R. J. Emery, R. I. Thomson, L. Royle, R. A. Gardner, D. I. Spencer, and G. P. Stafford. 2015. 'Characterization of a sialate-O-acetylesterase (NanS) from the oral pathogen Tannerella forsythia that enhances sialic acid release by NanH, its cognate sialidase', *Biochem J*, 472: 157-67.

Pluvinage, B., M. A. Higgins, D. W. Abbott, C. Robb, A. B. Dalia, L. Deng, J. N. Weiser, T. B. Parsons, A. J. Fairbanks, D. J. Vocadlo, and A. B. Boraston. 2011. 'Inhibition of the pneumococcal virulence factor StrH and molecular insights into N-glycan recognition and hydrolysis', *Structure*, 19: 1603-14.

Pudlo, N. A., Urs, K., Kumar, S. S., German, J. B., Mills, D. A., & Martens, E. C. 2015. Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *mBio*, *6*(6), e01282–15.

Rangarajan, E. S., K. M. Ruane, A. Proteau, J. D. Schrag, R. Valladares, C. F. Gonzalez, M. Gilbert, A. F. Yakunin, and M. Cygler. 2011. 'Structural and enzymatic characterization of NanS (YjhS), a 9-O-Acetyl N-acetylneuraminic acid esterase from Escherichia coli O157:H7', *Protein Sci*, 20: 1208-19.

Rangel, A., S. M. Steenbergen, and E. R. Vimr. 2016. 'Unexpected Diversity of Escherichia coli Sialate O-Acetyl Esterase NanS', *J Bacteriol*, 198: 2803-9.

Reddy, S. K., V. Bagenholm, N. A. Pudlo, H. Bouraoui, N. M. Koropatkin, E. C. Martens, and H. Stalbrand. 2016. 'A beta-mannan utilization locus in Bacteroides ovatus involves a GH36 alpha-galactosidase active on galactomannans', *FEBS Lett*, 590: 2106-18.

Renzi, F., P. Manfredi, M. Mally, S. Moes, P. Jeno, and G. R. Cornelis. 2011. 'The N-glycan glycoprotein deglycosylation complex (Gpd) from Capnocytophaga canimorsus deglycosylates human IgG', *PLoS Pathog*, 7: e1002118.

Reuter, G. (1963). Vergleichende Untersuchungen über die Bifidus-Flora des Säuglings- und Erwachsenenstuhl. Zentralbl.Bakteriol.Parasitenkd.Orig.Abt.I 191 : 486-507 .

Robb, M., J. K. Hobbs, S. A. Woodiga, S. Shapiro-Ward, M. D. Suits, N. McGregor, H. Brumer, H. Yesilkaya, S. J. King, and A. B. Boraston. 2017. 'Molecular Characterization of N-glycan Degradation and Transport in Streptococcus pneumoniae and Its Contribution to Virulence', *PLoS Pathog*, 13: e1006090.

Robert, C., Chassard, C., Lawson, P. A., Bernalier-Donadille, A. 2007. Bacteroides cellulosilyticus sp. nov., a cellulolytic bacterium from the human gut microbial community. *Int.J.Syst.Evol.Microbiol.* 57 : 1516-1520

Robinson, L. S., W. G. Lewis, and A. L. Lewis. 2017. 'The sialate O-acetylesterase EstA from gut Bacteroidetes species enables sialidase-mediated cross-species foraging of 9-O-acetylated sialoglycans', *J Biol Chem*, 292: 11861-72.

Rodriguez, J. M., K. Murphy, C. Stanton, R. P. Ross, O. I. Kober, N. Juge, E. Avershina, K. Rudi, A. Narbad, M. C. Jenmalm, J. R. Marchesi, and M. C. Collado. 2015. 'The composition of the gut microbiota throughout life, with an emphasis on early life', *Microb Ecol Health Dis*, 26: 26050.

Rogers, T. E., N. A. Pudlo, N. M. Koropatkin, J. S. Bell, M. Moya Balasch, K. Jasker, and E. C. Martens. 2013. 'Dynamic responses of Bacteroides thetaiotaomicron during growth on glycan mixtures', *Mol Microbiol*, 88: 876-90.

Rogowski, A., J. A. Briggs, J. C. Mortimer, T. Tryfona, N. Terrapon, E. C. Lowe, A. Basle, C. Morland, A. M. Day, H. Zheng, T. E. Rogers, P. Thompson, A. R. Hawkins, M. P. Yadav, B. Henrissat, E. C. Martens, P. Dupree, H. J. Gilbert, and D. N. Bolam. 2016. 'Corrigendum: Glycan complexity dictates microbial resource allocation in the large intestine', *Nat Commun*, 7: 10705.

Rothschild, D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, D. Zeevi, P. I. Costea, A. Godneva, I. N. Kalka, N. Bar, S. Shilo, D. Lador, A. V. Vila, N. Zmora, M. Pevsner-Fischer, D. Israeli, N. Kosower, G. Malka, B. C. Wolf, T. Avnit-Sagi, M. Lotan-Pompan, A. Weinberger, Z. Halpern, S. Carmi, J. Fu, C. Wijmenga, A. Zhernakova, E. Elinav and E. Segal (2018). "Environment dominates over host genetics in shaping human gut microbiota." Nature 555: 210.

Roy, S., K. Honma, C. W. Douglas, A. Sharma, and G. P. Stafford. 2011. 'Role of sialidase in glycoprotein utilization by Tannerella forsythia', *Microbiology*, 157: 3195-202.

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. 2013. 'Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin'. *Cell host & microbe*, 14(2), 195-206.

Sender, R., S. Fuchs, and R. Milo. 2016. 'Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans', *Cell*, 164: 337-40.

Sethi, M. K., and S. Fanayan. 2015. 'Mass Spectrometry-Based N-Glycomics of Colorectal Cancer', *Int J Mol Sci*, 16: 29278-304.

Sethi, M. K., W. S. Hancock, and S. Fanayan. 2016. 'Identifying N-Glycan Biomarkers in Colorectal Cancer by Mass Spectrometry', *Acc Chem Res*, 49: 2099-106.

Sharma, V., M. Ichikawa, and H. H. Freeze. 2014. 'Mannose metabolism: more than meets the eye', *Biochem Biophys Res Commun*, 453: 220-8.

Sheng, Y. H., R. Lourie, S. K. Linden, P. L. Jeffery, D. Roche, T. V. Tran, C. W. Png, N. Waterhouse, P. Sutton, T. H. Florin, and M. A. McGuckin. 2011. 'The MUC13 cell-surface mucin protects against intestinal inflammation by inhibiting epithelial cell apoptosis', *Gut*, 60: 1661-70.

Shipman, J. A., K. H. Cho, H. A. Siegel, and A. A. Salyers. 1999. 'Physiological characterization of SusG, an outer membrane protein essential for starch utilization by Bacteroides thetaiotaomicron', *J Bacteriol*, 181: 7206-11.

Shoseyov, O., Z. Shani, and I. Levy. 2006. 'Carbohydrate binding modules: biochemical properties and novel applications', *Microbiol Mol Biol Rev*, 70: 283-95.

Simurina, M., N. de Haan, F. Vuckovic, N. A. Kennedy, J. Stambuk, D. Falck, I. Trbojevic-Akmacic, F. Clerc, G. Razdorov, A. Khon, A. Latiano, R. D'Inca, S. Danese, S. Targan, C. Landers, M. Dubinsky, Consortium Inflammatory Bowel Disease Biomarkers, D. P. B. McGovern, V. Annese, M. Wuhrer, and G. Lauc. 2018. 'Glycosylation of Immunoglobulin G Associates With Clinical Features of Inflammatory Bowel Diseases', *Gastroenterology*, 154: 1320-33 e10.

Skorupski, K., and R. K. Taylor. 1996. 'Positive selection vectors for allelic exchange', *Gene*, 169: 47-52.

Song, Y. L., Liu, C. X., McTeague, M., Finegold, S. M. 2004. "*Bacteroides nordii*" sp. nov. and "*Bacteroides salyersae*" sp. nov. isolated from clinical specimens of human intestinal origin. *J.Clin.Microbiol.* 42 : 5565-5570

Sonnenburg, J. L., C. T. Chen, and J. I. Gordon. 2006. 'Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host', *PLoS Biol*, 4: e413.

Sonnenburg, J. L., J. Xu, D. D. Leip, C. H. Chen, B. P. Westover, J. Weatherford, J. D. Buhler, and J. I. Gordon. 2005. 'Glycan foraging in vivo by an intestine-adapted bacterial symbiont', *Science*, 307: 1955-9.

Sriswasdi, S., Yang, C., & Iwasaki, W. 2017. Generalist species drive microbial dispersion and evolution. *Nature Communications*, *8*, 1162.

Sriwilaijaroen, N., S. I. Nakakita, S. Kondo, H. Yagi, K. Kato, T. Murata, H. Hiramatsu, T. Kawahara, Y. Watanabe, Y. Kanai, T. Ono, J. Hirabayashi, K. Matsumoto, and Y. Suzuki. 2018. 'N-glycan structures of human alveoli provide insight into influenza A virus infection and pathogenesis', *FEBS J*, 285: 1611-34.

Staudacher, E. 2015. 'Mucin-Type O-Glycosylation in Invertebrates', *Molecules*, 20: 10622-40.

Strasser, R. 2016. 'Plant protein glycosylation', *Glycobiology*, 26: 926-39.

Studier, F. W., and B. A. Moffatt. 1986. 'Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes', *J Mol Biol*, 189: 113-30.

Stowel, S.R., Ju, T., and Cummings R.D. 2015. 'Protein Glycosylation in Cancer'. Annu Rev Pathol. 10: 473-510.

Suganuma, M., T. Nomura, Y. Higa, Y. Kataoka, S. Funaguma, H. Okazaki, T. Suzuki, K. Fujiyama, H. Sezutsu, K. I. Tatematsu, and T. Tamura. 2018. 'N-glycan sialylation in a silkworm-baculovirus expression system', *J Biosci Bioeng*.

Sun, L., L. Ma, Y. Ma, F. Zhang, C. Zhao, and Y. Nie. 2018. 'Insights into the role of gut microbiota in obesity: pathogenesis, mechanisms, and therapeutic perspectives', *Protein Cell*.

Szekrenyes, A., S. S. Park, E. Cosgrave, A. Jones, T. Haxo, M. Kimzey, S. Pourkaveh, Z. Szabo, Z. Sosic, P. Feng, P. Sejwal, K. Dent, D. Michels, G. Freckleton, J. Qian, C. Lancaster, T. Duffy, M. Schwartz, J. K. Luo, J. van Dyck, P. K. Leung, M. Olajos, R. Kowle, K. Gao, W. Wang, J. Wegstein, S. Tep, A. Domokos, C. Varadi, and A. Guttman. 2018. 'Multi-site N-Glycan mapping study 2: UHPLC', *Electrophoresis*, 39: 998-1005.

Tailford, L. E., E. H. Crost, D. Kavanaugh, and N. Juge. 2015. 'Mucin glycan foraging in the human gut microbiome', *Front Genet*, 6: 81.

Tailford, L. E., V. A. Money, N. L. Smith, C. Dumon, G. J. Davies, and H. J. Gilbert. 2007. 'Mannose foraging by Bacteroides thetaiotaomicron: structure and specificity of the beta-mannosidase, BtMan2A', *J Biol Chem*, 282: 11291-9.

Talens-Perales, D., A. Gorska, D. H. Huson, J. Polaina, and J. Marin-Navarro. 2016. 'Analysis of Domain Architecture and Phylogenetics of Family 2 Glycoside Hydrolases (GH2)', *PLoS One*, 11: e0168035.

Tengeler, A. C., T. Kozicz, and A. J. Kiliaan. 2018. 'Relationship between diet, the gut microbiota, and brain function', *Nutr Rev*.

Theodoratou, E., H. Campbell, N. T. Ventham, D. Kolarich, M. Pucic-Bakovic, V. Zoldos, D. Fernandes, I. K. Pemberton, I. Rudan, N. A. Kennedy, M. Wuhrer, E. Nimmo, V. Annese, D. P. McGovern, J. Satsangi, and G. Lauc. 2014. 'The role of glycosylation in IBD', *Nat Rev Gastroenterol Hepatol*, 11: 588-600.

Thursby, E., and N. Juge. 2017. 'Introduction to the human gut microbiota', *Biochem J*, 474: 1823-36.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. 2007. 'The human microbiome project', *Nature*, 449: 804-10.

Tuson, H. H., M. H. Foley, N. M. Koropatkin, and J. S. Biteen. 2018. 'The Starch Utilization System Assembles around Stationary Starch-Binding Proteins', *Biophys J*.

Val-Cid, C., X. Biarnes, M. Faijes, and A. Planas. 2015. 'Structural-Functional Analysis Reveals a Specific Domain Organization in Family GH20 Hexosaminidases', *PLoS One*, 10: e0128075.

Valdes, A. M., J. Walter, E. Segal, and T. D. Spector. 2018. 'Role of the gut microbiota in nutrition and health', *BMJ*, 361: k2179.

Varki, A., and R. Schauer. 2009. 'Sialic Acids.' in nd, A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart and M. E. Etzler (eds.), *Essentials of Glycobiology* (Cold Spring Harbor (NY)).

Veillon, A., and A. Zuber. 1898. 'Recherches sur quelques microbes strictment anakrobies et leur role en pathologie.' *Arch. Med. Exp. Anat. Pathol.* 10:517-545.

Vimr, E. R., K. A. Kalivoda, E. L. Deszo, and S. M. Steenbergen. 2004. 'Diversity of microbial sialic acid metabolism', *Microbiol Mol Biol Rev*, 68: 132-53.

Vuong, T. V., and D. B. Wilson. 2010. 'Glycoside hydrolases: catalytic base/nucleophile diversity', *Biotechnol Bioeng*, 107: 195-205.

Wang, P., H. Wang, J. Gai, X. Tian, X. Zhang, Y. Lv, and Y. Jian. 2017. 'Evolution of protein N-glycosylation process in Golgi apparatus which shapes diversity of protein N-glycan structures in plants, animals and fungi', *Sci Rep*, 7: 40301.

Wang, X., Z. Deng, C. Huang, T. Zhu, J. Lou, L. Wang, and Y. Li. 2018. 'Differential N-glycan patterns identified in lung adenocarcinoma by N-glycan profiling of formalin-fixed paraffin-embedded (FFPE) tissue sections', *J Proteomics*, 172: 1-10.

Watanabe, Y., Nagai, F., Morotomi, M., Sakon, H., Tanaka, R. 2010. *Bacteroides clarus* sp. nov., *Bacteroides fluxus* sp. nov. and *Bacteroides oleiciplenus* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 60 (Pt8): 1864-1869 .

Wefers, D., J. J. V. Cavalcante, R. R. Schendel, J. Deveryshetty, K. Wang, Z. Wawrzak, R. I. Mackie, N. M. Koropatkin, and I. Cann. 2017. 'Biochemical and Structural Analyses of Two Cryptic Esterases in Bacteroides intestinalis and their Synergistic Activities with Cognate Xylanases', *J Mol Biol*, 429: 2509-27.

Wegmann U., Goesmann A., Carding S.R. Complete Genome Sequence of *Bacteroides ovatus* V975. 2016. *Genome Announcements*.4(6):e01335-16.

Wu, H., E. Esteve, V. Tremaroli, M. T. Khan, R. Caesar, L. Manneras-Holm, M. Stahlman, L. M. Olsson, M. Serino, M. Planas-Felix, G. Xifra, J. M. Mercader, D. Torrents, R. Burcelin, W. Ricart, R. Perkins, J. M. Fernandez-Real, and F. Backhed. 2017. 'Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug', *Nat Med*, 23: 850-58.

Wu, H. J., and E. Wu. 2012. 'The role of gut microbiota in immune homeostasis and autoimmunity', *Gut Microbes*, 3: 4-14.

Wu, Z. L., X. Huang, A. J. Burton, and K. A. Swift. 2016. 'Probing sialoglycans on fetal bovine fetuin with azido-sugars using glycosyltransferases', *Glycobiology*, 26: 329-34.

Yan, Z., X. J. Gao, T. Li, B. Wei, P. Wang, Y. Yang, and R. Yan. 2018. 'Fecal microbiota transplantation in experimental ulcerative colitis reveals associated gut microbial and host metabolic reprogramming', *Appl Environ Microbiol*.

Zauner, G., C. A. Koeleman, A. M. Deelder, and M. Wuhrer. 2012. 'Mass spectrometric O-glycan analysis after combined O-glycan release by beta-elimination and 1-phenyl-3-methyl-5-pyrazolone labeling', *Biochim Biophys Acta*, 1820: 1420-8.

Zhu, H., H. C. Chan, Z. Zhou, J. Li, H. Zhu, L. Yin, M. Xu, L. Cheng, and J. Sha. 2004. 'A Gene Encoding Sialic-Acid-Specific 9-O-Acetylesterase Found in Human Adult Testis', *J Biomed Biotechnol*, 2004: 130-36.

Zimmermann, K., A. Haas, and A. Oxenius. 2012. 'Systemic antibody responses to gut microbes in health and disease', *Gut Microbes*, 3: 42-7.

Zivkovic, A. M., J. B. German, C. B. Lebrilla, and D. A. Mills. 2011. 'Human milk glycobiome and its impact on the infant gastrointestinal microbiota', *Proc Natl Acad Sci U S A*, 108 Suppl 1: 4653-8.

Zuckert, W. R. 2014. 'Secretion of bacterial lipoproteins: through the cytoplasmic membrane, the periplasm and beyond', *Biochim Biophys Acta*, 1843: 1509-16.

Appendix A

Supplemental table 3.1: RNA-seq data table showing differentially upregulated genes in wild type *B. thetaiotaomicron* grown on bovine α_1-AGP as a carbon source versus glucose as control.

Gene_id	BaseMean (Glucose)	BaseMean (α_1-AGP)	log2 Fold Change (α_1-AGP vs Gluc)	Fold Change (α_1-AGP vs Gluc)	lfcSE	stat	pvalue	padj
BT_0081	26.17	54.78	1.09	2.13	0.31	3.53	0.00042	0.001027
BT_0082	1314.75	2890.29	1.14	2.20	0.07	16.94	2.34E-64	4.4E-63
BT_0124	486.55	1104.46	1.18	2.27	0.09	12.89	5.04E-38	6.15E-37
BT_0224	899.92	4204.66	2.22	4.67	0.07	33.17	3E-241	2E-239
BT_0225	1965.92	8951.05	2.19	4.55	0.06	38.41	0	0

BT_0226	1147.48	5560.14	2.28	4.85	0.07	34.62	1.2E-262	8.9E-261
BT_0227	12884.04	61245.14	2.25	4.75	0.06	39.63	0	0
BT_0297	740.89	1863.69	1.33	2.52	0.08	17.25	1.12E-66	2.21E-65
BT_0298	546.05	1168.13	1.10	2.14	0.08	13.82	2.07E-43	2.81E-42
BT_0300	1650.79	3991.53	1.27	2.42	0.07	17.18	3.54E-66	6.85E-65
BT_0301	720.29	1610.60	1.16	2.23	0.08	14.56	4.79E-48	7.16E-47
BT_0302	616.56	1379.54	1.16	2.24	0.08	14.31	1.9E-46	2.72E-45
BT_0315	2018.39	10574.13	2.39	5.23	0.07	33.02	3.9E-239	2.5E-237
BT_0316	4414.40	27214.19	2.62	6.16	0.05	49.90	0	0
BT_0320	1969.17	8280.90	2.07	4.20	0.06	36.09	2.9E-285	2.2E-283
BT_0321	500.47	1804.33	1.85	3.60	0.11	16.49	4.23E-61	7.69E-60
BT_0348	150.88	324.63	1.11	2.15	0.14	8.08	6.3E-16	3.95E-15
BT_0459	10974.70	40235.79	1.87	3.67	0.07	25.55	5.1E-144	2.2E-142
BT_0461	4797.80	10997.96	1.21	2.32	0.08	15.89	7E-57	1.19E-55
BT_0498	47.45	118.16	1.31	2.49	0.22	5.89	3.83E-09	1.56E-08
BT_0504	1212.03	2565.40	1.08	2.12	0.07	15.61	6.15E-55	1.03E-53
BT_0506	400.45	34658.25	6.43	86.31	0.07	86.09	0	0
BT_0507	35.29	7849.35	7.82	225.66	0.17	46.99	0	0
BT_0508	9.12	158.10	4.09	17.01	0.35	11.74	7.84E-32	8.18E-31
BT_0509	29.63	348.50	3.55	11.67	0.21	17.06	3.18E-65	6.05E-64
BT_0510	731.04	2142.49	1.55	2.93	0.08	20.57	5.32E-94	1.42E-92
BT_0511	967.07	2280.96	1.24	2.36	0.07	17.32	3.13E-67	6.23E-66
BT_0521	532.58	2937.96	2.46	5.51	0.08	30.08	1E-198	5.4E-197
BT_0522	249.61	1160.58	2.21	4.64	0.10	23.19	6.4E-119	2.2E-117
BT_0523	1279.25	5654.94	2.14	4.42	0.07	30.40	6.2E-203	3.4E-201
BT_0524	1827.23	5232.86	1.52	2.86	0.06	23.42	2.9E-121	9.8E-120
BT_0525	2842.06	12715.09	2.16	4.47	0.06	38.98	0	0
BT_0537	199.78	686.25	1.78	3.44	0.11	15.71	1.25E-55	2.11E-54
BT_0586	1508.60	3571.36	1.24	2.37	0.07	17.34	2.33E-67	4.65E-66
BT_0669	101.65	309.47	1.61	3.05	0.16	10.36	3.61E-25	3.16E-24
BT_0670	150.17	442.67	1.57	2.97	0.14	11.16	6.43E-29	6.24E-28
BT_0671	38.24	118.77	1.63	3.09	0.23	7.03	2.07E-12	1.06E-11
BT_0683	72.59	4387.71	5.91	60.05	0.13	46.38	0	0
BT_0724	40.17	175.09	2.11	4.31	0.21	9.94	2.72E-23	2.26E-22
BT_0784	5515.36	13419.44	1.28	2.43	0.07	17.52	1.07E-68	2.2E-67
BT_0817	3270.64	6841.59	1.06	2.09	0.06	16.40	1.86E-60	3.34E-59
BT_0839	236.23	723.74	1.62	3.07	0.11	14.96	1.31E-50	2.03E-49
BT_0840	1075.65	3212.04	1.58	2.99	0.08	20.98	9.33E-98	2.54E-96
BT_0841	745.05	2141.55	1.53	2.88	0.08	19.06	5.67E-81	1.34E-79
BT_0842	246.73	707.62	1.52	2.87	0.10	14.64	1.66E-48	2.5E-47
BT_0843	547.47	1393.32	1.35	2.54	0.09	15.56	1.29E-54	2.15E-53
BT_0844	530.37	1318.82	1.32	2.49	0.09	15.12	1.2E-51	1.91E-50
BT_0845	603.13	1768.41	1.55	2.94	0.08	18.93	6.23E-80	1.46E-78
BT_0846	876.45	1853.30	1.08	2.12	0.08	13.20	8.31E-40	1.06E-38

BT_0865	429.25	981.60	1.19	2.29	0.09	13.43	4.19E-41	5.49E-40
BT_0866	283.70	706.41	1.32	2.49	0.11	12.07	1.57E-33	1.72E-32
BT_0867	470.54	1443.63	1.62	3.07	0.10	16.68	1.76E-62	3.24E-61
BT_0881	376.69	838.59	1.16	2.23	0.11	10.30	6.99E-25	6.01E-24
BT_0902	8804.15	22065.29	1.33	2.51	0.08	16.15	1.13E-58	1.97E-57
BT_0922	3078.50	7769.06	1.34	2.52	0.06	23.61	2.7E-123	9.8E-122
BT_0923	1601.01	4197.01	1.39	2.62	0.07	20.48	3.44E-93	9.05E-92
BT_0933	46.38	99.70	1.11	2.16	0.22	5.04	4.74E-07	1.62E-06
BT_0934	55.16	111.36	1.02	2.03	0.21	4.90	9.53E-07	3.17E-06
BT_0935	45.16	92.20	1.02	2.03	0.23	4.54	5.65E-06	1.74E-05
BT_0944	561.06	1185.07	1.08	2.11	0.09	12.45	1.45E-35	1.68E-34
BT_0945	957.53	1957.09	1.03	2.04	0.07	14.43	3.52E-47	5.14E-46
BT_0947	1168.65	2629.95	1.17	2.25	0.10	12.14	6.14E-34	6.81E-33
BT_0998	9392.75	21806.18	1.22	2.32	0.06	20.00	5.8E-89	1.43E-87
BT_0999	840.27	1867.03	1.15	2.22	0.08	14.69	7.06E-49	1.07E-47
BT_1031	8.75	25.14	1.50	2.83	0.46	3.25	0.001147	0.002613
BT_1032	3039.46	105062.36	5.11	34.57	0.06	90.72	0	0
BT_1033	1521.15	52217.67	5.10	34.34	0.06	83.59	0	0
BT_1034	145.31	20386.18	7.14	140.78	0.10	74.65	0	0
BT_1035	205.03	23939.00	6.86	116.34	0.09	79.34	0	0
BT_1036	265.81	96057.41	8.49	360.66	0.09	98.43	0	0
BT_1037	159.46	48381.46	8.24	303.21	0.11	76.18	0	0
BT_1038	135.87	30719.73	7.82	226.59	0.12	64.79	0	0
BT_1039	188.52	44143.54	7.87	234.27	0.12	65.08	0	0
BT_1040	305.24	111397.78	8.51	365.71	0.11	74.17	0	0
BT_1042	529.92	230144.95	8.76	434.88	0.10	85.13	0	0
BT_1043	206.47	65115.88	8.31	316.50	0.12	69.52	0	0
BT_1044	171.48	41685.28	7.93	243.83	0.11	70.18	0	0
BT_1045	199.38	61808.95	8.28	310.54	0.10	85.61	0	0
BT_1046	75.62	936.01	3.62	12.34	0.16	22.39	4.8E-111	1.5E-109
BT_1047	51.30	402.17	2.97	7.83	0.21	14.41	4.56E-47	6.63E-46
BT_1048	33.26	276.70	3.06	8.33	0.23	13.33	1.54E-40	2E-39
BT_1049	42.01	314.76	2.90	7.48	0.20	14.81	1.18E-49	1.81E-48
BT_1050	38.92	416.97	3.43	10.81	0.20	17.39	9.03E-68	1.81E-66
BT_1052	2284.82	13916.52	2.61	6.09	0.08	32.04	2.8E-225	1.7E-223
BT_1142	226.66	1736.60	2.94	7.65	0.09	31.03	2.2E-211	1.2E-209
BT_1163	579.30	1776.02	1.62	3.06	0.08	21.52	1E-102	2.9E-101
BT_1164	487.24	3714.05	2.93	7.64	0.08	37.17	2.3E-302	1.9E-300
BT_1165	306.19	2198.75	2.84	7.17	0.09	31.67	4.5E-220	2.7E-218
BT_1166	106.09	975.12	3.21	9.26	0.13	25.29	4.6E-141	1.9E-139
BT_1167	88.34	842.82	3.26	9.59	0.13	25.36	6.9E-142	2.9E-140
BT_1168	13.27	31.71	1.25	2.37	0.39	3.17	0.001544	0.003447
BT_1177	271.18	612.98	1.18	2.26	0.10	11.25	2.29E-29	2.26E-28
BT_1178	978.93	2427.48	1.31	2.48	0.07	19.29	5.93E-83	1.41E-81

BT_1179	333.51	1321.39	1.98	3.96	0.09	21.84	9.9E-106	3E-104
BT_1180	331.30	1452.04	2.13	4.39	0.09	24.35	5.9E-131	2.3E-129
BT_1181	173.96	723.57	2.05	4.15	0.12	17.69	5.15E-70	1.09E-68
BT_1182	278.66	1196.94	2.10	4.29	0.09	22.16	7.5E-109	2.4E-107
BT_1183	2303.92	5954.90	1.37	2.59	0.06	23.69	4.4E-124	1.6E-122
BT_1184	407.80	1320.29	1.70	3.25	0.09	18.22	3.86E-74	8.53E-73
BT_1226	8255.53	20089.59	1.28	2.43	0.05	23.63	2.1E-123	7.6E-122
BT_1230	91.38	4257.93	5.54	46.45	0.12	44.96	0	0
BT_1231	193.17	428.56	1.15	2.21	0.13	9.01	2.1E-19	1.53E-18
BT_1233	856.54	1966.52	1.20	2.30	0.07	16.40	2.01E-60	3.59E-59
BT_1234	795.93	1774.18	1.16	2.23	0.07	15.60	6.78E-55	1.13E-53
BT_1236	949.74	1989.09	1.07	2.10	0.08	13.85	1.34E-43	1.83E-42
BT_1277	499.88	1024.56	1.04	2.05	0.09	11.20	3.94E-29	3.86E-28
BT_1280	333.66	752.10	1.17	2.25	0.09	12.61	1.87E-36	2.22E-35
BT_1282	74.60	157.29	1.08	2.12	0.19	5.72	1.06E-08	4.13E-08
BT_1283	96.21	251.43	1.39	2.63	0.16	8.77	1.8E-18	1.26E-17
BT_1284	137.97	323.62	1.22	2.33	0.15	7.90	2.88E-15	1.75E-14
BT_1285	324.23	711.34	1.14	2.20	0.10	10.90	1.21E-27	1.14E-26
BT_1286	36.79	173.32	2.25	4.75	0.21	10.48	1.04E-25	9.26E-25
BT_1414	1710.28	3716.58	1.12	2.17	0.07	16.20	5.33E-59	9.37E-58
BT_1416	2660.32	5336.70	1.00	2.01	0.07	14.55	5.5E-48	8.2E-47
BT_1417	4652.05	10606.16	1.19	2.28	0.06	18.51	1.75E-76	3.99E-75
BT_1418	1317.36	3073.80	1.22	2.33	0.07	18.51	1.71E-76	3.92E-75
BT_1427	27.48	56.35	1.03	2.04	0.29	3.53	0.00042	0.001027
BT_1442	3463.42	7291.85	1.07	2.11	0.06	17.49	1.83E-68	3.71E-67
BT_1443	1012.58	2493.83	1.30	2.46	0.08	16.23	3.09E-59	5.47E-58
BT_1444	531.73	1441.83	1.44	2.71	0.09	16.53	2.26E-61	4.12E-60
BT_1445	593.35	1598.24	1.43	2.70	0.08	17.23	1.67E-66	3.27E-65
BT_1446	1167.75	2997.12	1.36	2.57	0.08	17.40	8.34E-68	1.68E-66
BT_1486	568.30	68261.60	6.91	120.32	0.07	93.01	0	0
BT_1487	231.89	14993.26	6.01	64.64	0.09	65.19	0	0
BT_1488	345.26	33270.82	6.59	96.65	0.08	81.73	0	0
BT_1489	1056.23	11070.63	3.39	10.49	0.07	51.87	0	0
BT_1490	1034.03	9918.54	3.26	9.60	0.06	52.06	0	0
BT_1491	3641.18	33329.63	3.19	9.16	0.05	60.40	0	0
BT_1492	545.72	1253.15	1.20	2.30	0.09	14.08	4.87E-45	6.77E-44
BT_1591	243.53	507.99	1.06	2.09	0.11	9.76	1.68E-22	1.35E-21
BT_1618	229.11	564.43	1.30	2.47	0.11	11.97	5.21E-33	5.6E-32
BT_1620	273.24	569.44	1.06	2.08	0.11	9.83	7.99E-23	6.52E-22
BT_1624	3002.27	13250.18	2.14	4.41	0.06	38.44	0	0
BT_1625	1710.03	7313.67	2.10	4.28	0.06	32.27	1.7E-228	1E-226
BT_1626	588.31	1698.68	1.53	2.88	0.08	19.69	2.88E-86	7.05E-85
BT_1636	4193.62	11271.39	1.43	2.69	0.06	23.79	3.9E-125	1.5E-123
BT_1752	1767.09	5054.87	1.52	2.86	0.06	24.26	5.3E-130	2.1E-128

BT_1863	8.59	25.14	1.57	2.96	0.46	3.37	0.000746	0.001747
BT_1915	2099.14	30398.50	3.86	14.48	0.06	59.36	0	0
BT_1916	527.70	7698.81	3.87	14.61	0.08	49.34	0	0
BT_1917	3001.62	52511.73	4.13	17.50	0.05	76.62	0	0
BT_1928	17.26	43.63	1.36	2.56	0.35	3.90	9.53E-05	0.000252
BT_1948	0.86	19.98	4.62	24.54	1.10	4.18	2.92E-05	8.26E-05
BT_1949	43.67	2327.82	5.73	52.91	0.16	36.65	4.2E-294	3.3E-292
BT_1950	49.56	2572.40	5.67	51.04	0.15	38.67	0	0
BT_1951	55.38	1921.41	5.13	34.91	0.14	35.43	5.7E-275	4.2E-273
BT_1952	126.47	6085.21	5.59	48.23	0.10	55.24	0	0
BT_1953	376.66	15466.96	5.36	40.96	0.07	73.36	0	0
BT_1954	254.14	12941.59	5.67	50.97	0.08	66.97	0	0
BT_1955	804.12	44373.39	5.79	55.22	0.07	78.00	0	0
BT_1956	741.55	44539.76	5.91	60.13	0.07	80.72	0	0
BT_1957	879.35	57719.87	6.04	65.76	0.07	83.31	0	0
BT_1994	97.41	212.30	1.13	2.18	0.16	7.20	6.09E-13	3.2E-12
BT_1995	240.47	602.24	1.32	2.50	0.11	12.26	1.43E-34	1.62E-33
BT_1997	27.57	101.39	1.87	3.66	0.25	7.41	1.27E-13	7.03E-13
BT_1998	12138.92	53602.63	2.14	4.42	0.06	33.71	4.6E-249	3.1E-247
BT_2094	151.54	2585.74	4.08	16.95	0.11	37.41	2.3E-306	1.9E-304
BT_2095	188.03	2909.04	3.95	15.48	0.10	37.88	0	0
BT_2131	468.50	2334.10	2.31	4.98	0.09	25.93	2.9E-148	1.3E-146
BT_2167	1933.93	16979.49	3.13	8.78	0.06	48.71	0	0
BT_2170	648.59	3845.74	2.57	5.93	0.07	35.21	1.7E-271	1.2E-269
BT_2171	1545.91	8728.32	2.50	5.65	0.06	40.68	0	0
BT_2172	791.92	4787.47	2.59	6.04	0.07	36.19	1E-286	7.9E-285
BT_2173	182.63	1102.33	2.59	6.04	0.11	24.54	5.1E-133	2E-131
BT_2424	23.11	49.18	1.10	2.14	0.31	3.52	0.000432	0.001052
BT_2426	17.01	68.15	2.01	4.02	0.33	6.17	6.79E-10	2.92E-09
BT_2500	445.93	926.84	1.06	2.08	0.10	10.66	1.53E-26	1.4E-25
BT_2502	1356.86	6276.02	2.21	4.63	0.06	37.14	6.6E-302	5.4E-300
BT_2503	680.53	3432.55	2.34	5.05	0.07	33.58	3.1E-247	2.1E-245
BT_2504	631.33	3066.68	2.28	4.86	0.07	32.47	3.1E-231	1.9E-229
BT_2540	1490.53	3581.78	1.26	2.40	0.06	20.22	6.82E-91	1.75E-89
BT_2570	8194.91	22165.68	1.44	2.70	0.09	15.47	5.74E-54	9.47E-53
BT_2571	5198.91	12046.54	1.21	2.32	0.09	13.62	3E-42	4.03E-41
BT_2574	389.44	829.87	1.09	2.13	0.09	11.97	5.41E-33	5.8E-32
BT_2672	15.97	62.42	1.94	3.83	0.35	5.61	2.05E-08	7.84E-08
BT_2828	783.07	2322.44	1.57	2.96	0.07	21.65	6.1E-104	1.8E-102
BT_2928	21.28	74.57	1.81	3.50	0.29	6.24	4.37E-10	1.91E-09
BT_3011	128.69	261.48	1.03	2.04	0.15	6.81	9.66E-12	4.76E-11
BT_3012	432.40	922.85	1.09	2.13	0.09	11.75	6.87E-32	7.17E-31
BT_3013	235.23	514.02	1.12	2.18	0.12	9.32	1.16E-20	8.79E-20
BT_3014	101.96	218.07	1.09	2.13	0.16	6.96	3.42E-12	1.72E-11

BT_3015	395.56	1016.53	1.36	2.57	0.09	15.47	5.84E-54	9.6E-53
BT_3057	909.32	2850.62	1.65	3.13	0.07	22.91	3.9E-116	1.3E-114
BT_3059	57.57	243.88	2.07	4.21	0.18	11.79	4.65E-32	4.87E-31
BT_3169	5956.56	14705.21	1.30	2.47	0.06	22.94	1.7E-116	5.7E-115
BT_3197	34.33	69.28	1.02	2.02	0.27	3.80	0.000142	0.000368
BT_3198	28.80	112.62	1.96	3.89	0.25	7.89	2.91E-15	1.76E-14
BT_3199	98.15	224.68	1.19	2.27	0.16	7.59	3.19E-14	1.81E-13
BT_3200	13.28	100.82	2.89	7.39	0.32	8.95	3.48E-19	2.51E-18
BT_3201	35.23	113.48	1.68	3.20	0.23	7.35	2.04E-13	1.11E-12
BT_3259	154.61	358.66	1.22	2.32	0.14	8.45	2.83E-17	1.88E-16
BT_3322	391.00	1006.62	1.36	2.57	0.10	14.26	3.78E-46	5.39E-45
BT_3390	4585.05	10611.70	1.21	2.31	0.10	12.08	1.28E-33	1.41E-32
BT_3418	1629.20	4381.21	1.43	2.69	0.07	19.88	5.51E-88	1.35E-86
BT_3436	164.37	328.10	1.00	2.00	0.13	7.81	5.9E-15	3.51E-14
BT_3437	341.23	833.90	1.29	2.45	0.09	13.81	2.17E-43	2.95E-42
BT_3438	386.89	789.66	1.03	2.04	0.10	10.28	9.13E-25	7.82E-24
BT_3536	822.71	1874.91	1.19	2.28	0.07	16.03	8.03E-58	1.38E-56
BT_3537	41.50	123.62	1.58	2.99	0.24	6.65	2.84E-11	1.35E-10
BT_3737	3288.82	7387.98	1.17	2.25	0.07	16.56	1.37E-61	2.51E-60
BT_3793	338.03	789.03	1.22	2.33	0.10	12.34	5.59E-35	6.39E-34
BT_3825	1.68	10.21	2.64	6.23	0.92	2.88	0.003997	0.008331
BT_3831	10.57	30.39	1.53	2.89	0.43	3.58	0.000344	0.00085
BT_3868	1554.49	3600.87	1.21	2.32	0.07	17.31	3.9E-67	7.72E-66
BT_3909	1845.66	4112.29	1.16	2.23	0.06	18.30	8.27E-75	1.83E-73
BT_3910	840.07	1850.40	1.14	2.20	0.08	14.84	8.37E-50	1.29E-48
BT_3968	26.69	53.31	1.00	2.01	0.29	3.47	0.000528	0.001268
BT_3990	3931.90	9701.43	1.30	2.47	0.06	22.68	7.1E-114	2.3E-112
BT_3992	1546.75	3817.02	1.30	2.47	0.06	20.35	4.62E-92	1.2E-90
BT_4050	5170.69	12108.71	1.23	2.34	0.06	21.51	1.2E-102	3.6E-101
BT_4069	2555.46	5113.84	1.00	2.00	0.08	12.99	1.36E-38	1.68E-37
BT_4223	7.97	26.58	1.73	3.31	0.47	3.65	0.000262	0.000657
BT_4224	68.31	206.54	1.60	3.03	0.18	8.77	1.81E-18	1.27E-17
BT_4225	8.77	68.11	3.00	7.97	0.39	7.69	1.52E-14	8.82E-14
BT_4226	9.78	67.30	2.79	6.90	0.38	7.36	1.83E-13	1E-12
BT_4227	58.52	704.66	3.60	12.16	0.15	23.49	5.3E-122	1.8E-120
BT_4233	57.80	242.68	2.06	4.18	0.17	11.84	2.52E-32	2.66E-31
BT_4234	24.00	129.73	2.44	5.44	0.25	9.71	2.66E-22	2.14E-21
BT_4235	2684.53	7567.65	1.49	2.82	0.10	14.95	1.45E-50	2.24E-49
BT_4283	62.94	129.21	1.05	2.07	0.20	5.19	2.08E-07	7.37E-07
BT_4284	76.83	164.05	1.10	2.14	0.17	6.30	2.96E-10	1.31E-09
BT_4301	731.85	1655.01	1.18	2.26	0.08	15.53	2.04E-54	3.37E-53
BT_4311	2241.07	5834.47	1.38	2.60	0.07	20.64	1.13E-94	3.01E-93
BT_4384	45.89	99.78	1.12	2.17	0.22	5.05	4.52E-07	1.55E-06
BT_4393	60.02	191.96	1.67	3.18	0.19	9.03	1.79E-19	1.31E-18

BT_4394	951.42	1909.22	1.00	2.00	0.09	11.28	1.68E-29	1.66E-28
BT_4395	6525.25	13837.24	1.08	2.12	0.06	17.90	1.12E-71	2.41E-70
BT_4403	1243.72	8051.32	2.69	6.47	0.06	44.74	0	0
BT_4404	889.42	29120.83	5.03	32.74	0.09	57.23	0	0
BT_4405	265.67	7946.93	4.90	29.92	0.12	41.55	0	0
BT_4406	242.33	6091.07	4.65	25.09	0.12	37.83	0	0
BT_4407	603.78	15545.11	4.69	25.79	0.08	61.47	0	0
BT_4408	236.84	746.71	1.66	3.15	0.10	16.14	1.24E-58	2.16E-57
BT_4575	436.86	968.36	1.15	2.21	0.09	12.83	1.15E-37	1.39E-36
BT_4613	322.19	1016.49	1.66	3.15	0.09	18.61	2.72E-77	6.3E-76
BT_4681	2442.35	6838.19	1.49	2.80	0.07	20.49	2.68E-93	7.1E-92
BT_4683	119.73	380.03	1.67	3.18	0.13	12.66	1.04E-36	1.24E-35
BT_4684	209.50	560.49	1.42	2.67	0.11	12.91	3.87E-38	4.73E-37
BT_4685	35.76	71.32	1.00	2.00	0.25	4.03	5.48E-05	0.000149
BT_4688	80.83	315.65	1.95	3.88	0.17	11.61	3.69E-31	3.78E-30
BT_4689	110.10	424.96	1.95	3.86	0.15	13.27	3.64E-40	4.7E-39

Supplemental table 3.2: List of primers used in this study.

Primer name	Sequence	Use
0455_KO_For1	attataGTCGACATGTCCGTATTTTTTTAAGAATTCACCACCCTGAAC	KO
0455_KO_Rev1	TCGAAGACTTTTCATGGGGGTATTAGTTAATTTAACGAAGGCAAAGATAAG	KO
0455_KO_For1	CTAATACCCCCATGAAAAGTCTTCGAATCTTTTTGGGGATTGCATTT	KO
0455_KO_Rev2	attataTCTAGAACTTTGGAGTTATCGTGGCAAAGCCCAACA	KO
0458_KO_For1	attataGTCGACCATACTATTTTGGTAAGATGCTAAGAGATA	KO
0458_KO_Rev1	GAGTTGTTTCATAGTGGGGGCATTAGATATTAACGG	KO
0458_KO_For2	CTAATGCCCCCACTATGAAACAACTCCTAAAATTAACAGGATGTGTG	KO
0458_KO_Rev2	attataTCTAGATTTCAATCAATTCGCCATATACATCTTCCA	KO
0459_KO_For1	attataGTCGACATGGGGGACAGGAGACAGTCATAATTGGGGAGTCTGGT	KO
0459_KO_Rev1	TAGTGTATCTCTTTCAAAATAAAAATTAGGTGTTAGTAGTTAATA	KO
0459_KO_For2	GAAAGAGATACACTAAAGTTTGTCTTTTCTTATGGG	KO
0459_KO_Rev2	attataTCTAGATCCAAGCCGCCTTACCAGCCTCATCGCCTCCTACAT	KO
0460_KO_For1	attataGTCGACATGAACCTTTCGGTATCGGAGGTTATTTGCCGATGGAA	KO
0460_KO_Rev1	TCATAAATTATACATCTGGTGATTTCGTCTACAAATAAAAAG	KO
0460_KO_For2	CAGATGTATAATTTATGACTATAAATATTAGATATC	KO
0460_KO_Rev2	attataTCTAGACATACGCCTTTGAATGAAGTTCTCTTTCCATTTAAG	KO
0506_KO_For1	attataGTCGACCGGGATTCCAACTATATCTACGGTCCTTCGCTTCCG	KO
0506_KO_Rev1	CTTCAGGAGATCTGTAAGGTTAACTGTTTAATAAGTTGATTACTT	KO
0506_KO_For2	ACAGATCTCCTGAAGATTTTTTATTAGGTATTAATATAGAT	KO
0506_KO_Rev2	attataTCTAGATCCAACTCGCTTATCTTGCTGCCATGCAACAAAAGC	KO
1033_KO_For1	attataGTCGACATGGCAGAACTGAATAAAGAGGATCAG	KO
1033_KO_Rev1	GATGACCTATTTTATACCTTTTACGTATGATGATTCATGAGACGTATATT	KO
1033_KO_For2	CGTATGATGATTCATGAGACGTATATTATTAACTTGCACACTGGCT	KO
1033_KO_Rev2	attataTCTAGATACAGTGATACGTGACAAGTCATCGTT	KO
0455-61PULKO_For1	attataGTCGACATGTCCGTATTTTTTTAAGAATTCACCACCCTGAAC	KO
0455-61PULKO_Rev1	GCAATGAAACACTTAGAGGGGGGTATTAGTTAATTTAACGAA	KO
0455-61PULKO_For2	AACTAATACCCCCCTCTAAGTGTTTCATTGCGATACACTTAGAGG	KO
0455-61PULKO_Rev2	attataTCTAGAGTACCTGAACCACCGGCTAAAACAATTCCTTTCAT	KO
0459D->N_For	atccatgtaggtggtAATgaatgtccg	Cat mut
0459D->N_Rev	cggacattcATTaccacctacatggat	Cat mut
0459E->A_For	atgtaggtggtgacGCCtgtccgaagg	Cat mut
0459E->A_Rev	ccttcggacaGGCgtcaccacctacat	Cat mut

0460D->N_For	attcatgtaggaggcAATGAGgctggt	Cat mut
0460D->N_Rev	accagcCTCATTgcctcctacatgaat	Cat mut
0460E->A_For	attcatgtaggaggcGATGCCgctggt	Cat mut
0460E->A_Rev	accagcGGCATCgcctcctacatgaat	Cat mut
0506D->N_For	tatgtacacatggggggtAATgaagtcgag	Cat mut
0506D->N_Rev	ctcgacttcATTacccccccatgtgtacata	Cat mut
0459_CBMMut_For	attataGCTAGCtgtggctcggtacaagaagcc	Δ mut
0459_CBMMut_Rev	attataCTCGAGTGAatcagccggattcggagtaaattc	Δ mut
CBM_0459_For	attataGCTAGCatgaagccgattgttgctaatcagccg	Δ mut
CBM_0459_Rev	attataCTCGAGTGAattaattgtgatttcgtctacaaataa	Δ mut
CBM_0460_For	attataGCTAGCaatgctccgtatagtcctcac	Δ mut
CBM_0460_Rev	attataCTCGAGTGAtttcactactatttcatctgtaaa	Δ mut
CBM_0506_For	attataGCTAGCATGgtcaaagctccttatgcg	Δ mut
CBM_0506_Rev	attataCTCGAGTGAttcatgtttcagccattcttt	Δ mut
28_BT0506_For	attataGCTAGCtgcacgcctacggtgaagcaggag	cloning
28_BT0506_Rev	attataCTCGAGTGAaggaagcacttctttctcacccttttcattgac	cloning
28_BT0683_FOR	attataGCTAGCtgtagtagcccgaaaacagaggtccag	cloning
28_BT0683_REV	attataCTCGAGTGAaacgaaacagcctgcccgcctccg	cloning
BT0452_FOR	ATGTGGCATCTCTGGATGAA	qPCR
BT0452_REV	TTAACAGTTCGTCGCCTTTG	qPCR
BT0455_FOR	CCGAAAGACATCATTCATCG	qPCR
BT0455_REV	CGGACATCATATACGCCAAG	qPCR
BT0456_FOR	TATGTCCGTGACAAAGGGAA	qPCR
BT0456_REV	GGGAGTCAATAGCGGGAATA	qPCR
BT0457_FOR	ATAGTTGGCGGAAGAATTGG	qPCR
BT0457_REV	CCAGAGAATCACCATGATCG	qPCR
BT0458_FOR	TCTCAACGGCTCATTACTGC	qPCR
BT0458_REV	ATTGGAAGCATATTGTGGCA	qPCR
BT0459_FOR	GAATCCGGCTGATGGTACTT	qPCR
BT0459_REV	CTTCAAAGCGCCATCATAAA	qPCR
BT0460_FOR	GTAAGTTGCTGGGTTGGGAT	qPCR
BT0460_REV	CAAATATCCTCCGATGGCTT	qPCR
BT0461_FOR	CGTGCAGGAGCTAATGGTAA	qPCR
BT0461_REV	CGGTTGCTCTGTCGAAGATA	qPCR
BT1032_FOR	TGCACACTGGCTTTCACCCT	qPCR
BT1032_REV	GGACATCATTCCATTCGGGCAAA	qPCR
BT1038_FOR	ACGGTTGAGCAACAGGCTGA	qPCR
BT1038_REV	CGCTTCCTTCACTGCATCATCG	qPCR
BT1044_FOR	CCAGACAATCAGTCCTAGCGGTAA	qPCR
BT1044_REV	GAGATCCGGCTGACCATCCA	qPCR
BT1048_FOR	TGCTGCTATGGATGGCGGAT	qPCR
BT1048_REV	ACTCCGCCTTTACGGAAACCA	qPCR
BT1051_FOR	AGAAGATGTGCTCTGTGCAGGA	qPCR