Investigating the mechanisms of dietary fibre breakdown by the human and animal gut microbiota

Cosette Elizabeth Hinkley

A thesis submitted for the degree of Doctor of Philosophy

2020-2024

Biosciences Institute, Faculty of Medical Sciences
Framlington Place
Newcastle University

Abstract

The intestinal tracts of mammalian species are inhabited by a complex community of microbes, known collectively as the gut microbiota. Many bacterial members of the gut microbiota play an important role in the degradation of complex dietary carbohydrates.

Hemicelluloses are plant polysaccharides, often degraded by the gut microbiota; however these are generally tightly associated with cellulose, which is more recalcitrant to breakdown. Xylans are amongst the most abundant hemicelluloses within the human and animal diet.

Here, we initially investigated the xylan-degrading capabilities of members of the Bacteroidota phylum and show that ability to utilise different types of xylan is widespread but not ubiquitous. Furthermore, we demonstrate the ability to bioinformatically predict xylan utilisation capabilities for gut Bacteroidota.

Highly decorated xylans, such as those from cereal brans, are generally recalcitrant to breakdown by endoxylanases but have previously been shown to be cleaved by a GH98 enzyme from the human gut symbiont *Bacteroides ovatus*. GH98 enzymes have also been shown to be β1,4-galactosidases targeting blood group sugars. Here we present structural and functional studies into Bo98 endoxylanase to investigate the mechanism of complex cereal xylan targeting by this enzyme and reveal how the same family can possess specificity for both xylans and blood group sugars. We also show that that other gut and non-gut derived GH98 enzymes display the same specificity as Bo98. Furthermore, these GH98 endoxylanases possess a conserved CBM35 domain, which we show is a functional carbohydrate binding protein, recognising a range of arabinose-containing hemicelluloses at a novel binding site for the family.

Finally, we show that human gut Bacillota *Ruminococcus champanellensis* is capable of growth on cellulose and a whole corn cell wall substrate. We present plans for future experiments examining gene upregulation during growth on these substrates. These data will provide insights into the plant cell wall degrading strategy of this keystone Bacillota from the human gut.

This investigation demonstrates that members of the gut microbiota have developed diverse glycan utilising strategies and preferences, allowing survival of individual species in a highly populated ecological niche.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr Dave Bolam, for the providing me with the opportunity to work in his lab for the last four years and whose support and guidance has been invaluable throughout my studies. I would also like to thank Dr Lis Lowe, who has been equally supportive, and whose advice has helped me achieve so much throughout my PhD.

I would like to thank all members of the Bolam-Lowe lab, past and present, who I have had the opportunity to work with over the years. Carl Morland, Omar Al Jourani, Diana Githwe, Joe Manion, Kaspar Garnham, Natalia Los, Dr Jean Lou Reyre and honorary lab member Dr Curtis Cottam; your friendship, support and advice has been incredible and has made my PhD experience so enjoyable. This work would not have been possible without the collaboration and expertise of many people including Dr Arnaud Basle, Dr Alan Cartmell and Dr James Connelly.

I would like to give a special mention to Dr Jenn Ross, for being an amazing friend, colleague and role model. My PhD journey would have been so much more challenging without your support!

A massive thank you my friends and family all over the country for your unwavering love and support, especially my parents and brothers for always being there for me. Shout out to Scrumpy and Cola, the best pups in the world, who have definitely become PhD therapy dogs!

Last, and by no means least, thank you to my incredible husband, James. Sorry for all the late nights, tantrums and doubts! You have been my rock, and you've never doubted me, I couldn't have done it without you.

Contents

Table of Figures

List of Tables

Abbreviations

AA	Auxillary activity
ARA	Arabinose
AX	Arabinoxylan
BCA	Bicinchoninic acid
BEX	Beechwood glucuronoxylan
BHI	Brain heart infusion
Bo98	Bacteroides ovatus GH98 enzyme
Bo98-CBM35	Bacteroides ovatus ATCC8483 GH98 CBM35 protein
BWX	Birchwood xylan
CAZyme	Carbohydrate active enzyme
CBM	Carbohydrate binding module
CD	Corn digest
CE	Carbohydrate esterase
CFE	Cell free extract
CFS	Concave face site
CFU	Colony forming unit
Cj98	Cellvibrio japonicus UEDA107 GH98 enzyme
CMC	Carboxymethyl cellulose
CpGH98	Clostridium perfringens GH98 enzyme
CX	Corn xylan
DMSO	Dimethyl sulfoxide
DUF	Domain of unknown function
EFI-EST	Enzyme function initiative - enzyme similarity tool
FUC	Fucose
GAL	Galactose
GAX	Glucuronoarabinoxylan
GH	Glycoside hydrolase
GlcA	Glucuronic Acid
GlcNAc	*N*-acetylglucosamine
GpPUL	Gram positive polysaccharide utilsation loci
GST	Glutathione-S-transferase
GT	Glycosyltransferase
GX	Glucuronoxylan
HEC	Hydroxyethyl cellulose
HGM	Human gut microbiota
HPAEC	High performance anion exchange chromatography
HTCS	Hybrid two component system
IMAC	Immobilised metal affinity chromatography
ITC	Isothermal titration calorimetry
IToL	International tree of life
LB	Luria broth
MAG	Metagenomic assembled genome
MeGlcA	Methyl glucuronic acid

MM	Minimal media
OD	Optical density
OSX	Oat spelt xylan
Pa98	Segotella albensis M384 GH98 enzyme
PAD	Pulsed amperometric detection
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PL	Polysaccharide lyase
PUL	Polysaccharide utilisation locus
RAX	Rye arabinoxylan
Rc98	Ruminococcus champanellensis 18P13 GH98 enzyme
RF	Rumen fluid
RgGH98	Ruminococcus gnavus GH98 enzyme
RNA-seq	RNA sequencing
RPM	Revolutions per minute
Sca	Scaffoldin
SCFA	Short chain fatty acid
SDM	Site directed mutagenesis
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SeMet	SelenoMethionine
SGBP	Surface glycan binding protein
SNFG	Symbol nomenclature for glycans
Sp3GH98	Streptococcus pneumoniae BS71 GH98 enzyme
Sp4GH98	Streptococcus pneumoniae TIGR4 GH98 enzyme
SSN	Sequence similarity network
SUS	Starch utilisation system
TBDT	Ton B dependent transporter
TLC	Thin layer chromatography
TYG	Tryptone yeast extract glucose
VANISH	Volatile and/or associated negatively with industrialised societies of humans
VLS	Variable loop site
WAX	Wheat arabinoxylan
XYL	Xylose
YCFA	Yeast casitone fatty acid

1 Introduction

1.1 Introduction to the Human and Animal Gut Microbiome

The gastrointestinal tract of mammalian species is inhabited by a complex, diverse and ever-adapting population of microorganisms, including bacteria, archaea, fungi and protozoa (Thursby & Juge, 2017). This gut microbiota is of critical importance in the maintenance of homeostasis in the gut, with millions of years of co-evolution resulting in mutualistic relationships between microorganisms and host (Thursby & Juge, 2017). The microbiota contributes significantly to host health and function, not only in nutrient and energy supply but also in protection against pathogens, regulation of host immunity and maintenance of gut integrity (Flint, Scott, Louis, *et al.*, 2012). In humans the hindgut harbours by far the largest population of microorganisms, with slow flow rates and a mildly acidic pH providing ideal conditions (Flint, Scott, Louis, *et al.*, 2012). In ruminants, such as cattle and sheep, most microorganisms reside in the rumen, and in horses the caecum and colon.

Perhaps the most important role of the gastrointestinal microbiota is in nutrient supply to the host organism. Mammals produce enzymes efficient in the degradation of simple starches, however they are unable to break down more complex dietary fibre including polysaccharides from plant material such as cellulose, hemicelluloses and pectins (Scott *et al.*, 2008). Members of the gut microbiota are capable of degradation and fermentation of these indigestible dietary fibres, liberating short chain fatty acids (SCFAs), acetate, propionate and butyrate. These SCFAs provide an important source of energy for intestinal mucosal cells and are also critical for the modulation of immune function (McNeil, 1984; Vinolo *et al.*, 2011). In herbivorous ruminants these SCFAs

provide 70% of the total energy supply to the host organism, whereas humans, whose diet is generally much lower in fibrous plant material, energy provision from SCFA is around 10% (McNeil, 1984). Butyrate, in particular, is an important energy source for colonocytes, epithelial cells that line the colon, and is important in colonic immune regulation and health (Mortensen & Clausen, 1996).

Herbivorous mammals such as ruminants and horses, with diets high in recalcitrant fibrous material, have a much greater reliance on the microbial degradation of plant polysaccharides to provide nutritional requirements compared to omnivorous or carnivorous species. As such specialisation in digestive anatomy and physiology to reduce passage rates and improve efficiency of microbial fermentation are evident. Differences in the diversity, composition and metabolic potential of the gut microbiota between herbivorous and carnivorous mammals have been noted (Milani *et al.*, 2020; de Jonge *et al.*, 2022). Increased diversity and glycan breakdown potential within the gut microbiome of herbivorous mammals suggests microbiome evolution in cooperation with digestive tract type to optimise efficiency of energy extraction from a fibrous diet (Milani *et al.*, 2020).

The bacterial composition of human and animal gut microbiome is hugely diverse (**Figure 1-1**) with representatives from many phyla including Actinomycetota, Bacteroidota, Bacillota, Fusobacteria, Pseudomonadota, Verrucomicrobia and others (Eckburg *et al.*, 2005; Rinninella *et al.*, 2019). However, two phyla are ubiquitously predominant across mammalian species, Bacteroidota and Bacillota, representing 90% of the human gut microbiome (HGM) (Arumugam *et al.*, 2011; Ley *et al.*, 2008). Bacteroidota of the *Bacteroides* genus has been widely studied and shown to be particularly important in glycan degradation (McKee *et al.*, 2021), whereas *Prevotellacaea* spp. are relatively understudied, but may be equally as important in

glycan degradation within the intestinal niche (Dodd *et al.* 2011). A high proportion of gut Bacillota belong to the butyrate producing Clostridial clusters XIVa and IV, which include species belonging to the genera *Clostridium, Ruminococcus, Eubacterium, Roseburia* and *Butyrivibrio* (Lopetuso *et al.*, 2013). Despite predominance of these 2 phyla, overall microbial composition can be hugely divergent, not just between species, but within species and even individuals. Variation can be attributed to many factors including diet, host genetics, age, disease states, ethnicity or breed and geography to name a few (Manor *et al.*, 2020; Dąbrowska & Witkiewicz, 2016; Gaulke & Sharpton, 2018).

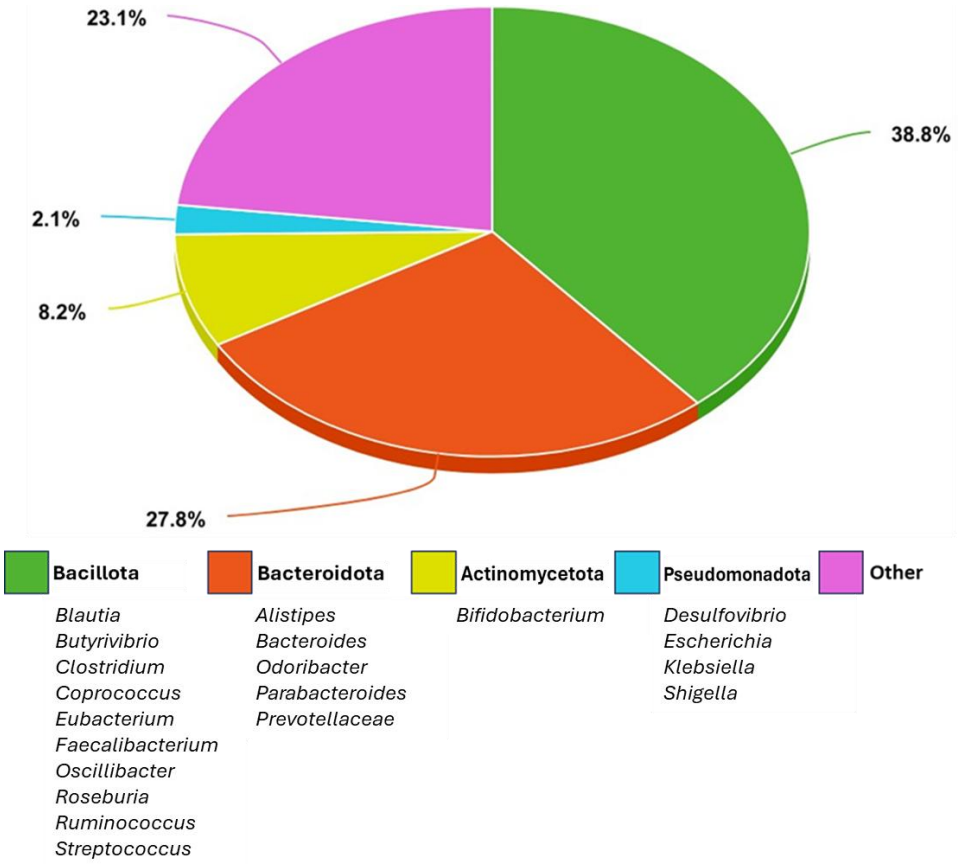


Figure 1-1: Major bacterial phyla constituting the human gut microbiota and their relative abundance. Based on average composition of the gut microbiota of healthy European individuals. Bacillota and Bacteroidota are the dominant groups present in the human colon, with Actinomycetota and Pseudomonadota less abundant. Whilst species and genus level composition varies between individuals, the phyla level remains fairly conserved. Adapted from D'Argenio & Salvatore (2015)

Unsurprisingly, the type and amount of dietary fibre influences the composition, diversity and richness of the gut microbiome (Makki *et al.*, 2018). An example of this in humans is the well documented *Bacteroides-Prevotella* ratio. Individuals on a diet higher in animal fat and protein, generally from Western populations, tend to have a higher proportion of *Bacteroides* in the gut, whereas those on a plant based, high fibre diet, such as hunter-gatherer populations, have increased abundance of the genus *Prevotella* (De Filippo *et al.*, 2010; Wu *et al.*, 2011; De Filippo *et al.*, 2017). Hunter-gatherer populations have higher microbial richness and biodiversity compared to urban populations, with increased Bacteroidota and decreased Bacillota abundance, and almost no Actinomycetota (Schnorr *et al.*, 2014). Furthermore, the disappearance of many taxa present in Hunter-Gatherer and ancient humans has been noted in industrialised human populations. These are known as VANISH (volatile and/or associated negatively with industrialised societies of humans) taxa (Carter *et al.*, 2023; Sprockett & Coyte, 2023). These losses in diversity are likely due to many factors including dietary changes, improved sanitation and use of antibiotics (Sprockett & Coyte, 2023; Carter *et al.*, 2023).

Alterations in the gut microbiota composition can result from multiple factors including diet, exposure to toxins, drugs, or pathogens. Dysbiosis is associated with pathogenesis of intestinal and extra-intestinal diseases including Inflammatory Bowel Disease, obesity, diabetes, hypertension, ischemic stroke and many more (Carding *et al.*, 2015). Improved understanding of the mechanisms of glycan utilisation in the dietary tract, and hence the impact of diet on microbial composition, is undoubtedly critical to elucidate interactions between diet, microbiome and health.

1.2 Plant Cell Wall Material

1.2.1 Overview of Plant Cell Wall Structure

Dietary fibres exposed to members of the gut microbiome are often in the context of whole plant cell wall material, the most important terrestrial source of biomass. Plant cell walls are composite structures conferring rigidity and strength to the cell, allowing it to withstand osmotic pressures. These walls are also involved in signalling, allowing control of growth, cell shape and cell differentiation (Cosgrove, 2005).

The major constituents of plant cell walls are cellulose, hemicelluloses, and pectins. Cellulose is in the form of microfibrils, which are interlinked by hemicellulosic polymers, which themselves are crosslinked by pectins (**Figure 1-2**). In secondary cell walls, which are produced once a plant has finished growing, these structural polysaccharides are crosslinked by lignins, phenolic compounds which increase mechanical strength of the plant cell wall (Cosgrove & Jarvis, 2012). The organisation and relative abundance of different carbohydrate types varies significantly depending on plant type, tissue type and developmental stage (Cosgrove & Jarvis, 2012), and intake of lignified secondary cell wall material is generally low in the human diet.

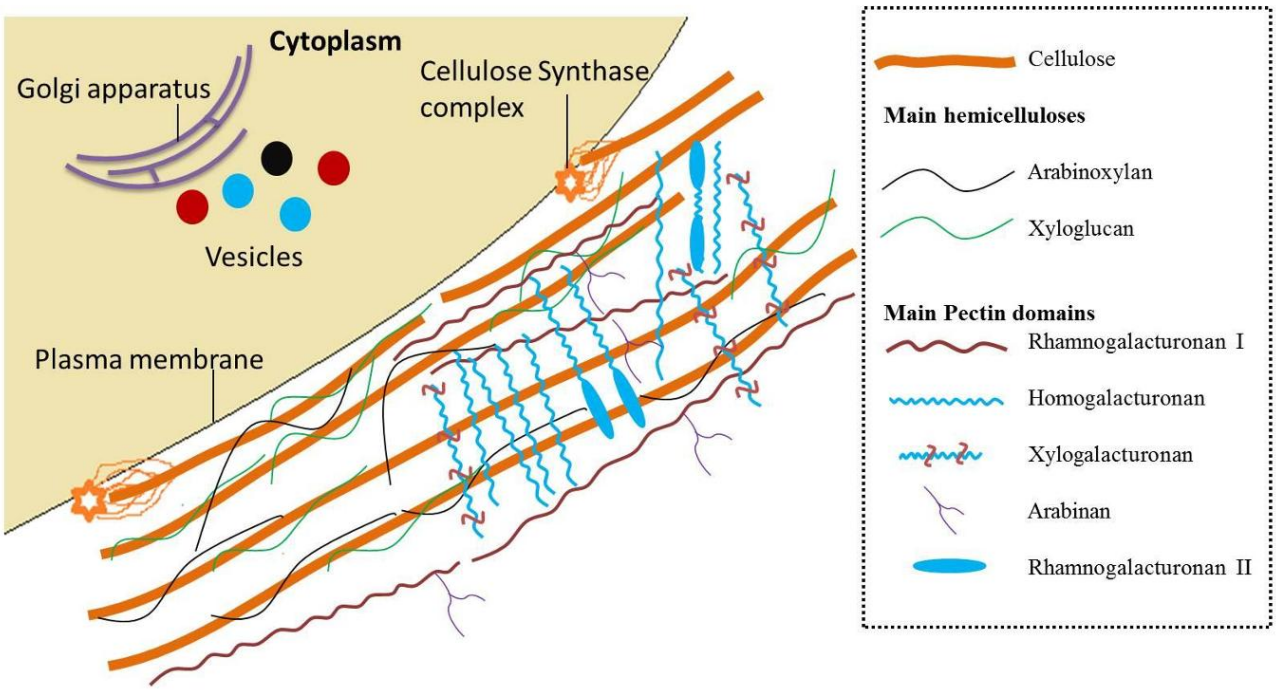


Figure 1-2: Structure of the primary plant cell wall. The primary cell wall consists of cellulose microfibrils embedded in a matrix formed from hemicelluloses and pectins. For simplicity, the left side demonstrates crosslinking between cellulose and hemicelluloses such as xyloglucan, xylans and mannans, without pectins. The right had side demonstrates pectins, such as rhamnogalacturonan I, homogalacturonan and xylogalacturonan, which themselves crosslink hemicellluloses, in particular xyloglucans. Adapted from Swart (2013), Cosgrove (2005).

1.2.2 Major Carbohydrate Components of the Plant Cell Wall

Cellulose is the most common abundant organic compound on earth and is the main structural component of plant cell walls (Klemm *et al.*, 2005)(**Figure 1-2**). Cellulose consists of linear glucose chains linked by β-1,4-glycosidic bonds (**Figure 1-3**). The majority of cellulose in plant cell walls is in a crystalline form; linear cellulose chains aggregate, pulled together by hydrogen bonding and van der Waals forces, to form microfibrils, which have a crystalline structure (Rongpipi *et al.*, 2019). In plant cell walls these microfibrils appear to be in bundles, of various sizes and in close association with hemicelluloses. The crystalline structure of such cellulose makes it fairly recalcitrant to enzymatic degradation by members of the gut microbiome as cellulose

molecules are packed so closely together as to prevent penetration by enzymes and small molecules. Cellulose microfibrils also contain regions of less ordered, amorphous cellulose, which is hydrolysed at a much faster rate, such that degree of crystallinity of a cellulose sample is important in determining its enzymatic digestibility (Park *et al.*, 2010; Estela & Luis, 2013). Whilst cellulose degradation in the human gut is minimal, multiple members of the rumen microbiome are specialised for its breakdown likely due to increased reliance of ruminants on microbial fermentation for host energy provision (Flint *et al.*, 2008).

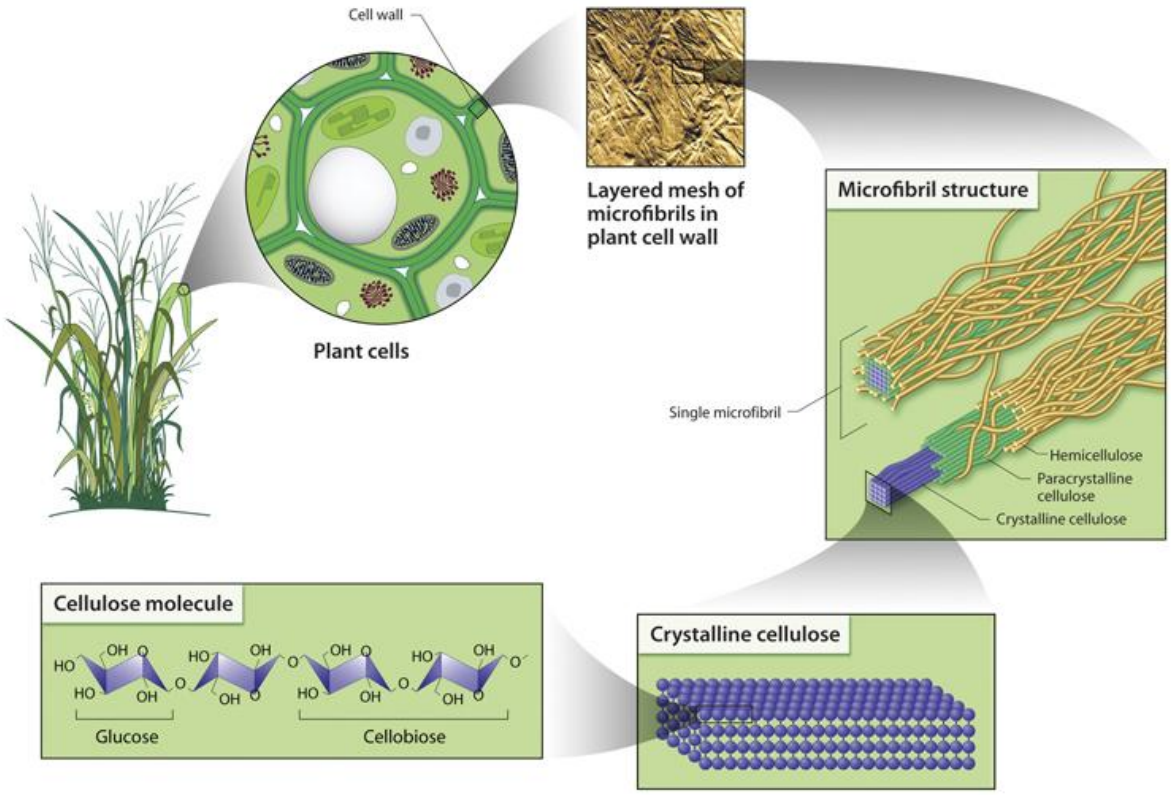


Figure 1-3: Structural organisation of cellulose within the plant cell wall. Cellulose consists of linear β-1,4- D-glucose chains joined by hydrogen bonds and van de Waals forces to form microfibrils. These microfibrils are interlinked by hemicelluloses and pectins within the plant cell wall structure. From US Department of Energy Genomic Science program http://genomicscience.energy.gov.

Hemicelluloses make up a group of heterogenous polysaccharides, with varying structures, typically constituting 15 to 35% dry mass of plants (Gírio *et al.*, 2010). Hemicelluloses include xyloglucans, xylans, mannans, glucomannans and β-glucans. These may contain pentoses (β-D-xylose, α-L-arabinose), hexoses (β-D-mannose, β -D-glucose, α-D-galactose) and/or uronic acids (α-D-glucuronic, α-D-4-O-methylgalacturonic and α-D-galacturonic acids), and sometimes have small amounts of α-L-rhamnose and α-L-fucose (Gírio *et al.*, 2010). Generally, hemicelluloses possess a β1-3 or β1-4 linked backbone with varying levels and types of branching (Scheller & Ulvskov, 2010). An overview of hemicellulose structures is provided in **Figure 1-4**. Hemicelluloses are often easily degraded by particular taxa of the gut microbiome, provide an important carbon source to microbes and are important for microbial homeostasis (Zhang *et al.*, 2022a).

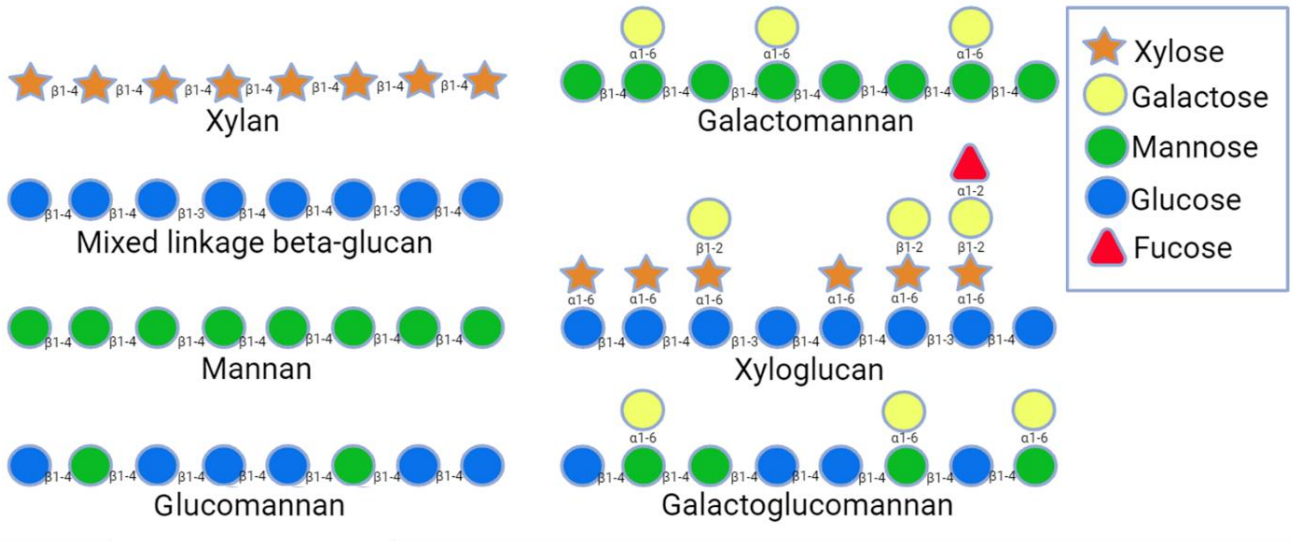


Figure 1-4: Schematic structures of hemicelluloses found within the plant cell wall. Monosaccharides are represented according to the Symbol Nomenclature for Glycans (SNFG) (Neelamegham *et al.*, 2019). Xylans can vary in type and amount of decoration of the β1-4 xylose backbone. Adapted from glycopedia.eu.

Pectins are structural heteropolysaccharides including homogalacturonan, rhamnogalacturonan I, rhamnogalacturonan II and xylogalacturonan, all of which are rich in galacturonic acid (Mohnen, 2008). Pectins possess a linear α-1,4 linked D-galacturonic acid chain, known as the homogalacturonic backbone, with different types, lengths, and complexities of branches (Tan & Nie, 2020) (**Figure 1-5**). Many members of the gut microbiota possess pectin-degrading capacity, and pectin-enriched diets lead to alteration of microbiome diversity and composition, indicating the importance of pectins in digestive function and health (Koutsos *et al.*, 2017).

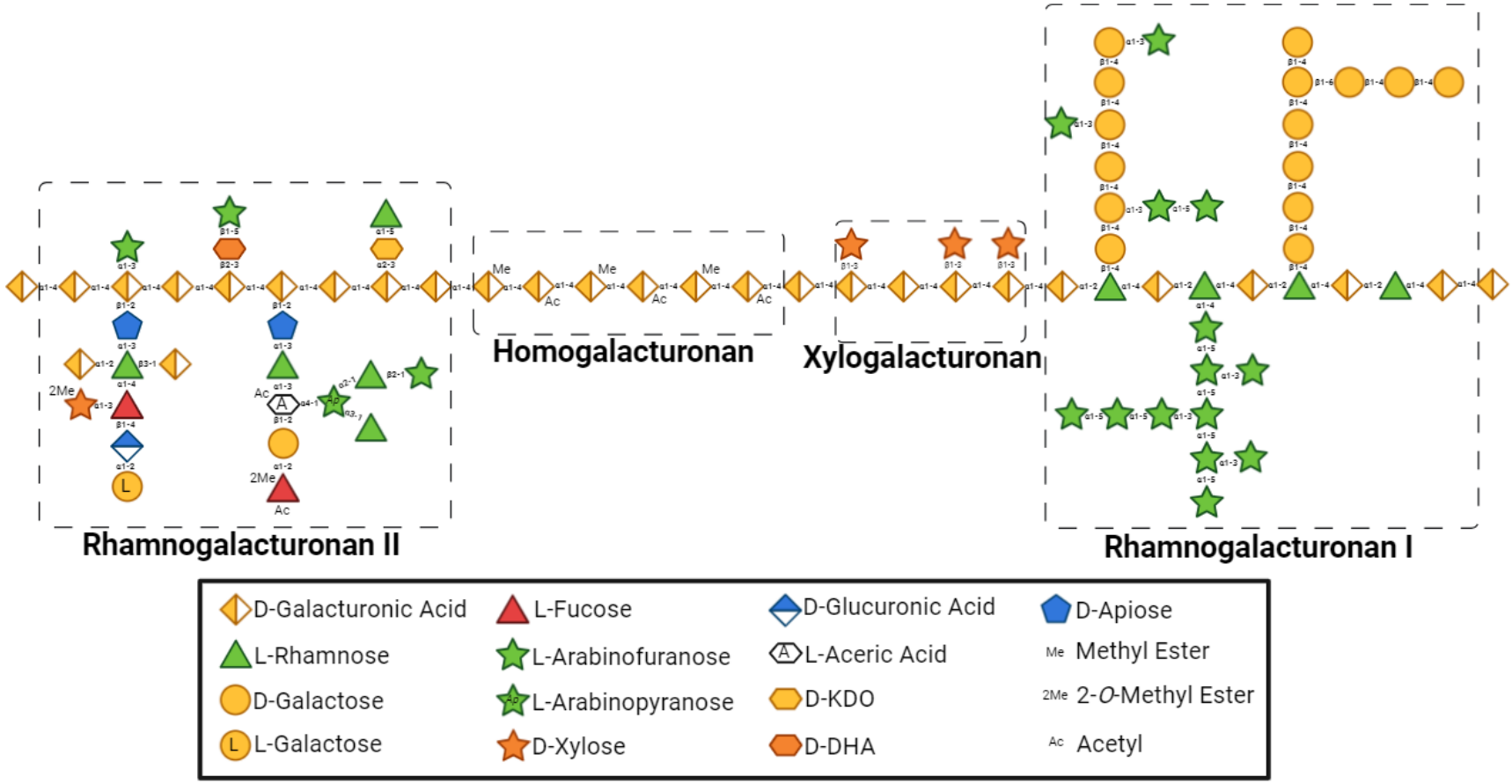


Figure 1-5: Schematic representing structure of pectins found within plant cell walls. Monosaccharides are represented according to the Symbol Nomenclature for Glycans (SNFG) (Neelamegham *et al.*, 2019). Adapted from Tan and Nie (2020) and Ndeh *et al.* (2017)

1.2.3 Structure of Xylans

Xylans are the most abundant hemicelluloses found in nature, occurring in grasses, cereals, herbs and hardwoods, constituting more than 30% dry weight of the terrestrial plant cell wall of many species (Scheller & Ulvskov, 2010). Xylans are characterised by a linear β-1,4 linked xylose backbone, commonly substituted with glycosyl side chains. Based on the nature of these glycosyl substitutions, xylans are generally characterised into three groups (**Figure 1-6**). Glucuronoxylans (GX) are substituted at O-2 with α-glucuronic acid (GlcA) or methyl-α-glucuronic acid (MeGlcA). These GXs are major components of the secondary cell walls of woody and herbaceous dicots, for example beech and birch plants (Ebringerová *et al.*, 2000.; Kulkarni *et al.*, 2012). Arabinoxylan (AX) is a significant component of cereal hemicellulose, found mainly in the outer bran layer of wheat, rye, barley and oats. The xylose backbone of AX is substituted at O-2 and/or O-3 with α-L-arabinofuranose units (Chen *et al.*, 2019). Glucuronoarabinoxylans (GAX), found in the cell wall of grasses such as rice and sorghum or cereals such as corn, are complex structures with more heterogeneity in the type and position of substitutions, and as yet their exact structures and branching points have not been elucidated. As well as α-(1 → 3) and/or α-(1 → 2)-L-arabinofuranose, β-(1 → 3) xylose, α-(1 → 2) GlcA and α-(1 → 2) MeGlcA branch points, GAXs may also possess di- and tri-saccharide branches, containing additional glycosyl units such as D-xylose and L-galactose (Xu *et al.*, 2022; Agger *et al.*, 2010). Complex GAXs are exemplified by corn xylan, CX, with a high proportion of backbone xylopyranose moieties substituted, 40% monosubstituted at O-2 or O-3 and 20% doubly substituted at both (Chanliaud *et al.*, 1995). There have been recent studies into the structure of CX, focussing on its structure, glycosyl constituents and linkage analysis (Xu *et al.*, 2022). CX consists of repeated subunits with complex branching

patterns, flanked by regions of lower complexity. 12 different glycosidic linkage types were found within CX, including 4 arabinosyl, 5 xylosyl and 3 galactosyl linkages, indicating the huge complexity of this substrate. Furthermore, we still know little about the relative positioning of side chain decorations, posing difficulties in complete elucidation of its microbial degradation.

Such complex GAX are recalcitrant to degradation by many known xylanolytic systems (Agger *et al.*, 2010), but their high prevalence in staple food crops worldwide, especially within cereal grains, makes their breakdown an important area of study.

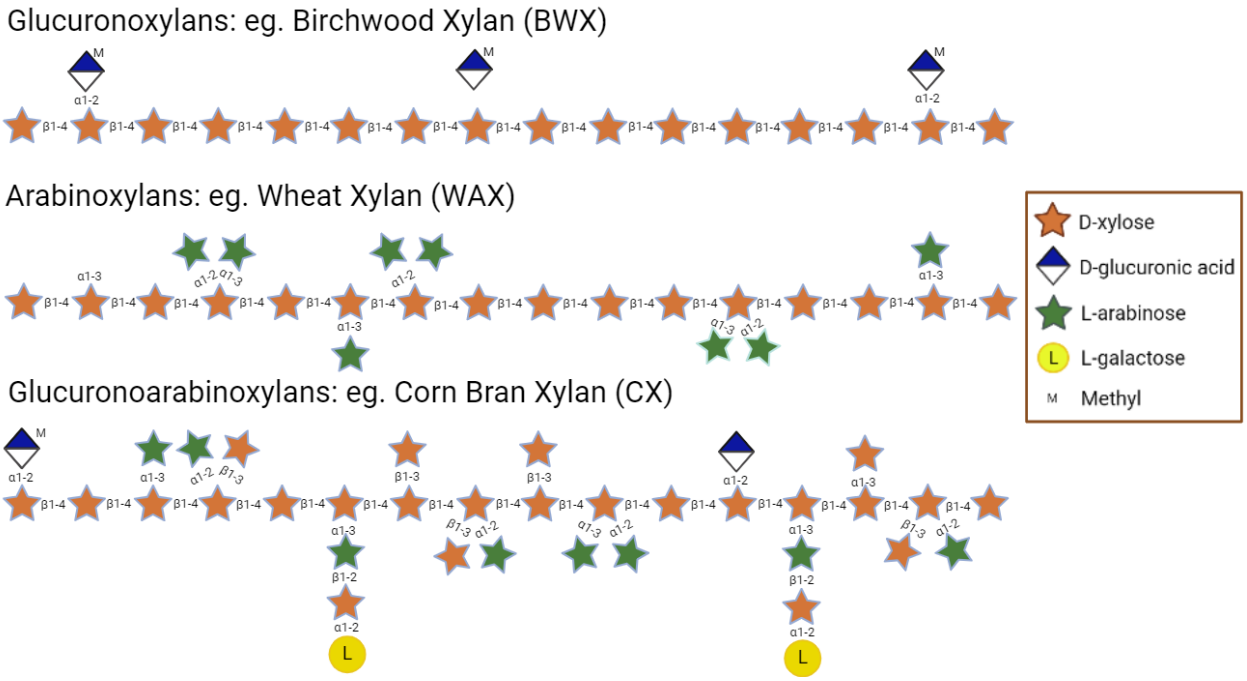


Figure 1-6: Schematic representing structure of three major classes of xylans found within terrestrial plant cell walls. Xylans utilised in this study include birchwood xylan (BWX), wheat arabinoxylan (WAX), and corn xylan (CX). Structures demonstrated are models and there is likely lots of natural variation in branching, particularly in GAX. CX is representative of highly branched and decorated xylans which are more recalcitrant to enzymatic degradation. OSX is an AX with fewer arabinose sidechains relative to WAX shown here. Glucuronoxylans (GX) are common in hard woods such as beech and birch and rarely present in the human diet. Arabinoxylans (AXs) are particularly prevalent in cereals such as wheat, rye and oat, whereas complex glucuronoarabinoxylans (GAX) are common in corn as well as grasses, and some fruit and vegetables. Monosaccharides are represented according to the Symbol Nomenclature for Glycans (SNFG) (Neelamegham *et al.*, 2019). Adapted from Rogowski *et al.* (2015)

1.3 Enzymes Involved in Degradation of Plant Cell Wall Material

1.3.1 Carbohydrate Active Enzymes (CAZymes)

Enzymes, and their associated domains, involved in the synthesis, degradation and modification of carbohydrates and glycoconjugates are referred to collectively as Carbohydrate Active Enzymes (CAZymes). This includes glycoside hydrolases (GHs), glycosyltransferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CEs), auxillary activities (AAs). These enzymes can be associated with non-catalytic carbohydrate binding modules (CBMs) which facilitate enzymatic catalysis. Information on CAZymes is stored in a carefully curated database, known as CAZy, dedicated to the display and analysis of genomic, structural and biochemical information on CAZymes (Drula *et al.*, 2022). Within the CAZy database, CAZymes are assigned into sequence-based families, built around at least one biochemically characterised member. Enzymes grouped into the same family based on sequence similarity may not share the same activity or substrate specificity, and as such many families contain members with different specificities. However, specificity for either alpha or beta linkages is almost always (with few exceptions) conserved due to conservation in the catalytic apparatus.

1.3.2 Glycoside Hydrolases (GHs)

GHs are a diverse group of enzymes involved in the hydrolysis of the glycosidic bond between two or more carbohydrate moieties, or between and carbohydrate and non-carbohydrate moiety (Henrissat & Davies, 1997) (**Figure 1-7**). There are currently 185 sequence-based GH families within the CAZy database, several of which contain

multiple sub-families. Members of a particular GH family display the same fold, and in most cases the catalytic apparatus and mechanism is conserved.

As GHs are involved in the catalysis of carbohydrates, they are of great interest when studying the mechanisms of glycan breakdown by the gut microbiome. In fact, bioinformatic analysis of GH enzymes encoded within the genomes of gut microbiome bacteria is an important method for prediction of glycan utilisation capabilities for individual species, or the microbiome as a whole (Kaoutari *et al.*, 2013). Likewise, human gut or rumen metagenomics can be important in the discovery of novel GH families or specificities (Viborg *et al.*, 2017; Neves *et al.*, 2021).

GHs can be characterised as endo-acting or exo-acting enzymes. Exo- enzymes cleave the substrate at the end, most commonly the non-reducing end, whereas endo- enzymes release oligosaccharides by cleaving at defined sites within a glycosidic chain (Davies & Henrissat, 1995) (**Figure 1-7**).

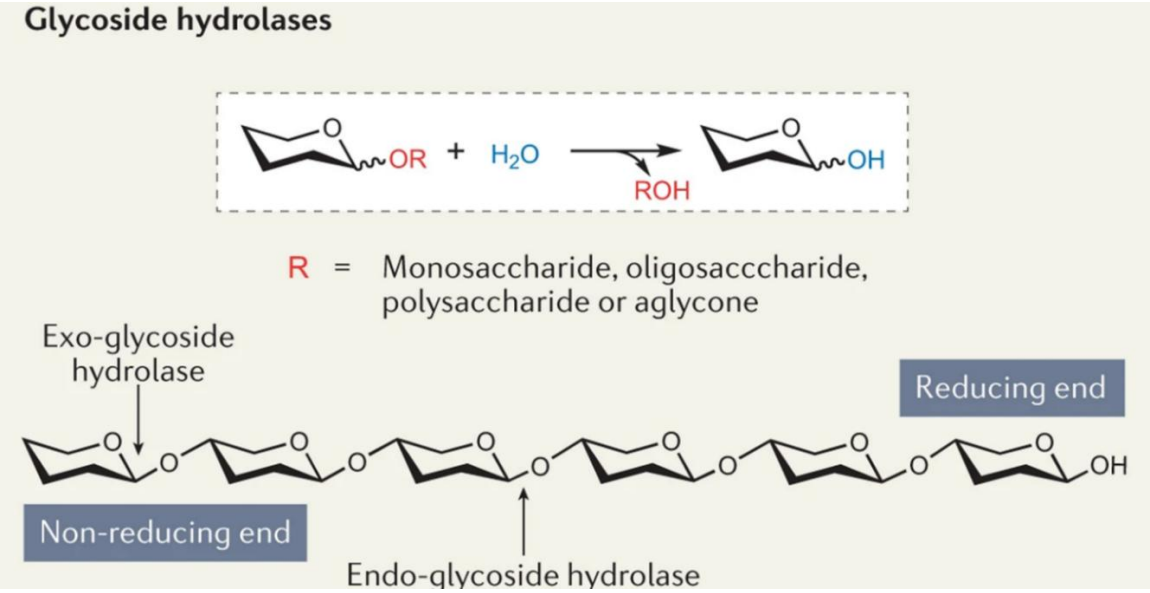


Figure 1-7: **Activity of Glycoside Hydrolase (GH) enzymes.** GHs are enzymes which catalyse the cleavage of glycosidic bonds between two or more carbohydrate moieties. Endo- acting GHs cleave internally within a glycan chain releasing an oligosacchairde product, whereas exo- acting enzymes cleave the terminal sugar, most frequently but not always from the non-reducing end. From Wardman *et al.* (2022).

The active site morphology of GHs is related to their endo- or exo- mode of action. Endo-acting GHs have an open cleft allowing binding and cleavage in the middle of a chain (**Figure 1-8A**) whereas in exo-acting enzymes the active site often forming a pocket or groove, which will only accommodate the terminal sugar (**Figure 1-8B**). Some GHs, such as cellobiohydrolases, present a tunnel topography, allowing processive exo- activity against a glycosidic chain, allowing the enzyme to release the product whilst remaining bound to the polysaccharide chain (**Figure 1-8C**) (Davies & Henrissat, 1995).

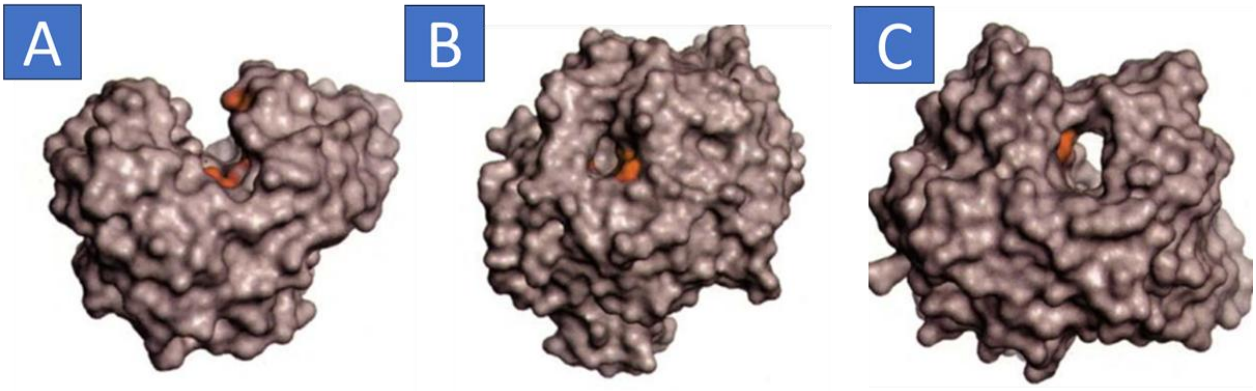


Figure 1-8: Topology of GH active sites. Panel A shows the active site morphology of endoglucanase E2 from *Thermobifida fusca* with an open cleft topography of endo-active GHs. Panel B shows a glucoamylase from *Aspergillus awamori* demonstrating a pocket topography, as commonly seen in exo-enzymes and panel C shows the tunnel topography of *Trichoderma reesei* cellobiohydrolase II with processive exo-activity. Proposed catalytic residues are shaded in red. Figure adapted from Davies and Henrissat (1995).

Generally, the catalytic activity of a GH enzyme in hydrolysis of a glycosidic bond involves two amino acid residues, a general acid-base and a nucleophile. Based on the spatial positioning of these catalytic residues hydrolysis occurs via either an inverting or a retaining mechanism (**Figure 1-9**). Inverting GH reactions occur when

the general acid and nucleophile, normally glutamic acid or aspartic acid, are approximately 10Å apart, sufficient space to allow activation of water by the base catalyst concurrently with attack of the glycosidic bond by the catalytic acid residue (**Figure 1-9A**). The reaction operates via a one-step single displacement mechanism, involving an oxocarbenium ion-like transition state, ultimately resulting in net inversion of the anomeric configuration (**Figure 1-9A**). Retaining GHs follow a double-displacement mechanism with formation and subsequent hydrolysis of a covalent glycosyl-enzyme intermediate. Each step involves an oxocarbenium ion-like transition state (**Figure 1-9B**). In retaining enzymes, the catalytic residues, typically glutamate and aspartate, are closer together, approximately 5 to 6 Å apart, than in inverting enzymes due to the separate attack of water in the two-step mechanism. In the first step, known as the glycosylation step, the nucleophile attacks the anomeric centre, forming a glycosyl enzyme intermediate. At the same time the acid catalyst pronates the glycosidic oxygen, cleaving the glycosidic bond. The second step is known as the deglycosylation step, where the covalent glycosyl enzyme intermediate is hydrolysed by water, with the other residue, now acting as a base, deprotonating the water molecule as it attacks (Ardèvol & Rovira, 2015; McCarter & Withers, 1994). This mechanism was first proposed by Koshland, (1953).

For the vast majority of GH families, catalytic machinery, mechanism and geometry around the glycosidic bond is conserved within the family, allowing predictions to be made when characterising new members of a family.

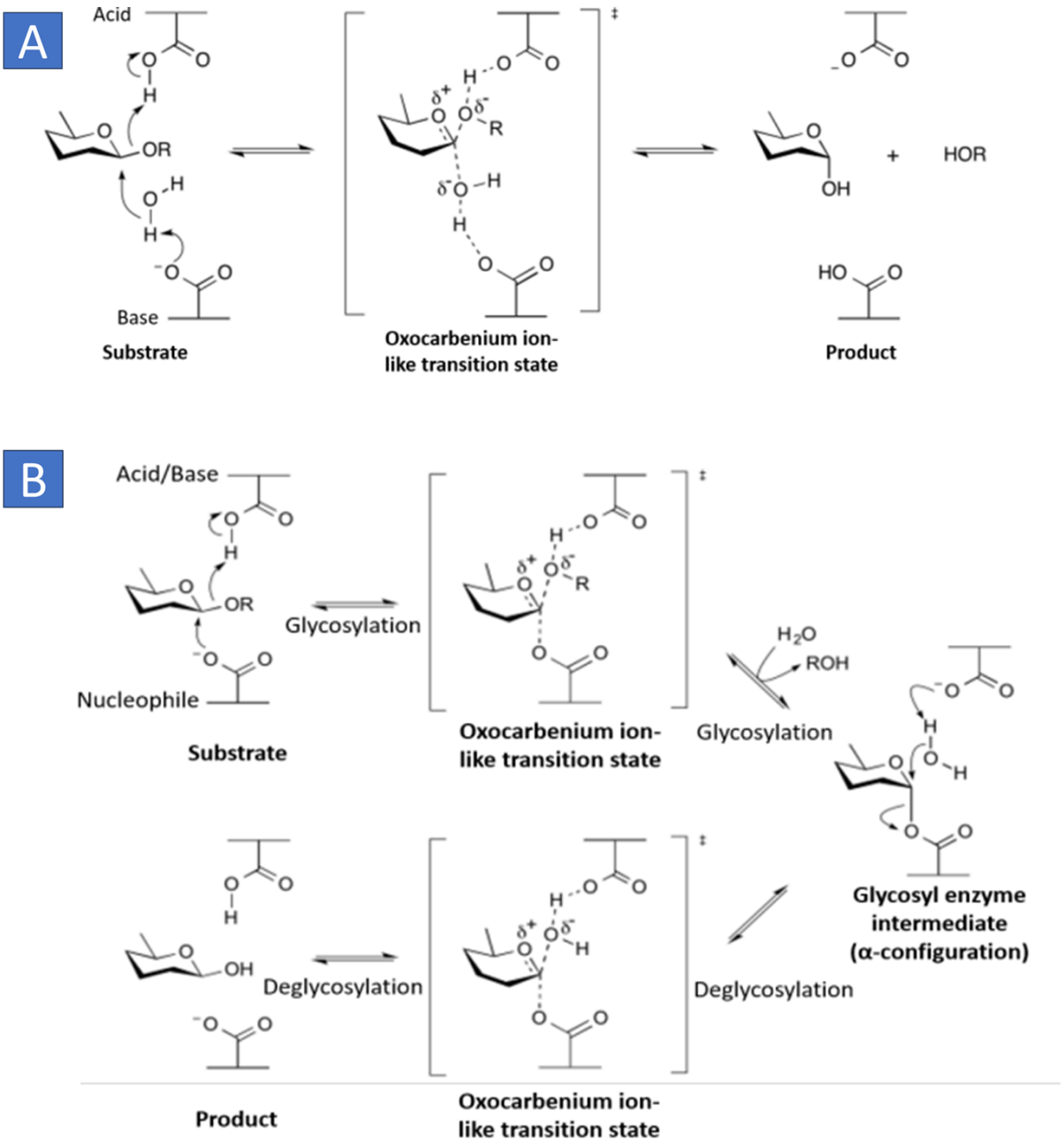


Figure 1-9: **Schematic diagram of inverting (A) and retaining (B) mechanisms of Glycoside Hydrolase (GH) enzymes.** Transitions states, Glycosyl-enzyme intermediate, Acid/base and nucleophile residues are indicated. Figure adapted from cazypedia.org/index.php/Glycoside_hydrolases.

1.3.3 GH Families Involved in Xylan Degradation

As mentioned previously, xylans are a major hemicellulose component of plant cell wall, and their abundance in the diet of humans, and to an even greater extent, ruminants, makes understanding xylan degradation by the microbiota a critical area of research. The debranching and degradation of xylans by many families of GHs has been well characterised (Labourel *et al.*, 2016; Rogowski *et al.*, 2015; Despres *et al.*, 2016; Biely *et al.*, 2023). Due to the structural heterogeneity and variety of glycosidic linkages within xylans these xylan-active GHs possess wide ranging activities included endoxylanases, β-xylosidases, α-L-arabinofuranosidases, α-glucuronidases and α-galactosidases, which work synergistically for the complete degradation of xylans in biological systems (**Figure 1-11**) (Dodd *et al.*, 2010). Endoxylanase enzymes target β1-4, or sometimes β1-3, glycosidic bonds within a xylan chain, degrading xylans into shorter oligosaccharide chains. The majority of enzymes possessing endoxylanase activity belong to GH families 5, 8, 10, 11, 30 (**Figure 1-10**), however endoxylanases are also found within families 7, 43, 51, 63 and 98. Enzymes belonging to many of these families do not display activity against complex GAX such as corn xylan. Only GH98 or GH5 subfamily 21 (GH5_21) enzymes have been proven capable of GAX degradation without prior debranching of the xylan chain (Rogowski *et al.*, 2015).

Endoxylanase GH Family	Specificity
GH5_21	AX and GAX –requirement for Ara decorations
GH5_34	AX – requirement for Ara decorations
GH10	Linear xylans, AX and GX - can accommodate some Ara or GlcA/mGlcA decoration
GH11	Linear xylans, AX and GX - can accommodate some Ara or GlcA/mGlcA decoration
GH30	GX – strong preference for GlcA/mGlcA on GX
GH98	GAX – requirement for Ara and GlcA/mGlcA decorations

Figure 1-10: Overview of specificity of glycoside hydrolase (GH) families with endoxylanase activity. GH families 5, 10, 11, 30 and 98 are the key families possessing endoxylanase activity with specificities for arabinoxylans (AX), glucuronoxylans (GX) and/or glucuronoarabinoxylans (GAX).

β-xylosidases display exo-activity, releasing single xylosyl units from the end of xylans or xylo-oligosaccharide chains (**Figure 1-11**). Such enzymes belong to GH families 3, 39, 43, 52 and 54. Other enzymes target side chains of the xylose backbones. Xylan targeting α-L-arabinofuranosidase enzymes target α-L-arabinose, or β-D-xylose substitutions O-2 or O-3 linked on the xylan backbone. Whilst some α-L-arabinofuranosidases specifically target single substitutions, others target double O-2 and O-3 decorated residues, removing the O3 linked sugar (Rogowski *et al.*, 2015). α-L-arabinofuranosidases are found in GH families 3, 43, 51, 54 and 62. Xylan specific α-glucuronidases belonging to families 67 and 115 target GlcA or MeGlcA substitutions. GH67 enzymes target decorations of the xylose residue at the non-reducing end, whereas GH115 α-glucuronidases can remove GlcA or mGlcA from xylose residues in the middle of the chain. A GH95 member encoded by hemicellulytic human gut symbiont *Bacteroides ovatus* possesses α-L-galactosidase activity, removing galactose decorations from xylans (Rogowski *et al.*, 2015).

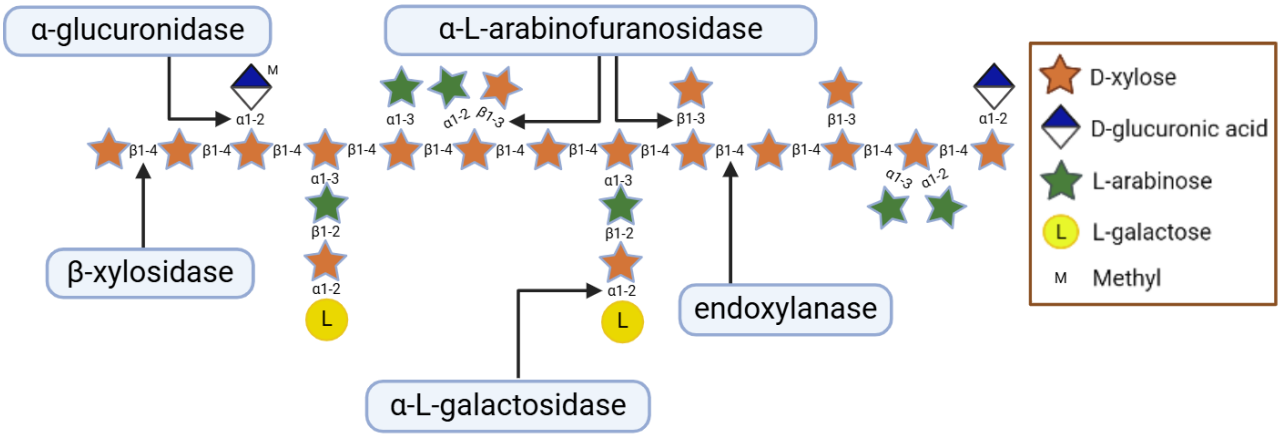


Figure 1-11: Xylan debranching and xylosidase enzymes and their target glycosidic bonds. Due to the structural heterogeneity and complexity of naturally occurring xylans, their completely degradation requires concerted activity of multiple GH enzymes. The xylan shown is a model GAX structure. The relative level of decoration and types of sugars will vary and in many cases are unknown in different xylans.

1.3.4 Carbohydrate Esterases and their Role in Xylan Degradation

Carbohydrate esterases (CE) are responsible for the de-O-acylation or de-N-acylation of substituted saccharides, in which sugars can play the role of the alcohol/amine or the acid. CEs show great diversity in substrate specificity and in structure, reflected in the 20 families within the CAZy database (Biely, 2012).

Polysaccharides, in particular xylans, are often highly acetylated, with mono- or di-acylation at positions 2 and 3 of backbone xylopyranose residues (**Figure 1-12**). Acetylation of xylose backbone residues makes the polysaccharide more resistant to xylanase- or xylosidase- mediated hydrolysis, and also reduces solubility in water, meaning that deacetylation of xylans is an important process in their degradation. Acetylxylan esterases, which liberate acetic acid esterifying the xylose backbone, are mainly found in CE families 1,4,5 and 6 (Biely *et al.*, 2016).

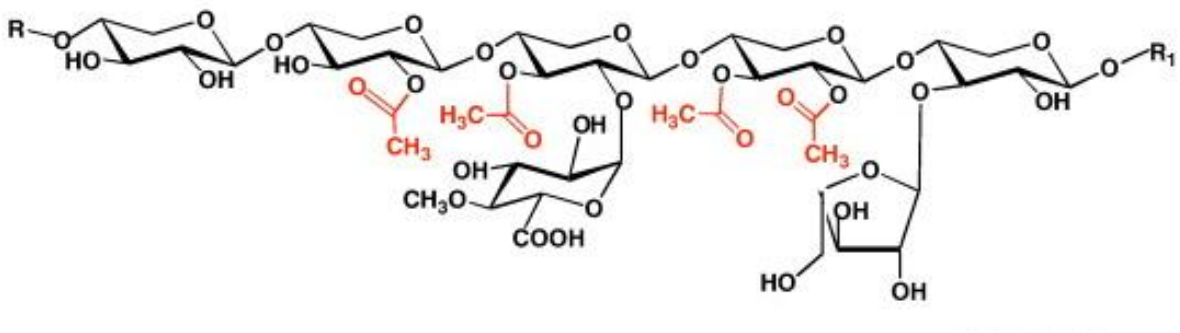


Figure 1-12: Acetylation of a Glucuronoarabinoxylan (GAX) from annual plant. Adapted from Biely (2012)

Although the action of CEs is likely critical for the complete degradation of unprocessed xylans, in the context of whole plant cell wall material in the human gut, xylan substrates utilised in this study have undergone extraction processes, involving alkali-treatment and hence will be almost completely deacetylated (Biely *et al.*, 2016). As such, we do not look in depth into the mechanisms of CEs in this study.

1.3.5 Polysaccharide Lyases and Glycosyltransferases

PLs are a group of non-hydrolytic enzymes which cleave uronic acid containing polysaccharide chains via a β-elimination mechanism. PLs are divided into 42 sequence-based families (Lombard *et al.*, 2010). GTs are involved in the formation of glycosidic bonds. As such, GTs are important in the biosynthesis of all known disaccharides, oligosaccharides and polysaccharides via the transfer of sugar moieties from activated donor to specific receptor molecules, forming glycosidic bonds. As with GH enzymes, GTs act via an inverting or a retaining mechanism (Lairson *et al.*, 2008).

Whilst it is important to acknowledge the activities of PLs and GTs, and the presence of encoding genes in members of the gut microbiome, the mechanisms of these enzymes are not important for the work presented here.

1.3.6 Carbohydrate Binding Modules (CBMs)

CBMs do not possess catalytic activity, but rather are sugar binding modules that fold into a structurally discrete module forming part of a larger multi-modular protein (Boraston *et al.*, 2004). CBMs are most commonly associated with GHs but are also found in PLs and GTs. At present (June 2024), CBMs are grouped into 103 sequence-based families. Members of these families generally display similar structural folds, and carbohydrate binding function, however several families do exhibit diversity in targeted carbohydrate ligands. CBMs can be classified into three main types based on their target ligand, determined by binding site architecture (Gilbert *et al.*, 2013). Type A CBMs bind to crystalline cellulose and chitin, with planar binding sites rich in aromatic residues creating a flat, hydrophobic platform for binding to the planar faces of their crystalline ligands. Type B CBMs are endo-type and bind internally to single glycan chains. Binding sites appear as grooves or clefts with binding subsites to accommodate longer sugar chains. Type C CBMs bind to termini of glycans, normally the non-reducing end, and binding sites appear as pockets recognising mono-, di- or tri-saccharides (Gilbert *et al.*, 2013). Type A, B and C CBM modules are depicted in **Figure 1-13**.

Many CBM domains are linked to their cognate catalytic domains via a flexible linker region, whereas some are more directly abutted onto the catalytic domain. It is as yet unclear whether these differences equate to any functional difference. Generally, the role of a CBM is to enhance catalytic efficiency by binding to a carbohydrate ligand. CBMs carry our three main roles. The first is targeting the enzyme to a distinct region of the substrate, for example the reducing end or internal polysaccharide chains, dictated by the morphology of the binding site. CBMs also increase the enzyme concentration in close proximity to its substrate improving efficiency of degradation

(Hervé *et al.*, 2010). Finally, CBMs can adhere to particular cell wall components, allowing the attached catalytic domain to exhibit catalytic activity onto neighbouring carbohydrate substrate (Gilbert *et al.*, 2013).

CBMs do not undergo conformational changes when binding to their ligand, but the topography of the binding site is complementary to the shape of the target ligand (Boraston *et al.*, 2004). Interactions between CBMs and their carbohydrate ligand are driven by extensive hydrogen bonding between polar amino acid residues and hydroxyl groups of the carbohydrate ligand, but mostly by the position and orientation of aromatic amino acid residues (mainly tyrosine and tryptophan but occasionally also phenylalanine) forming a hydrophobic platform for stacking interactions with the planar face of sugar rings (Boraston *et al.*, 2004). Furthermore, coordinated metal ions in the binding site can directly interact with the target ligand. Interactions are generally weak, making them easily reversible. Some CBMs display promiscuity and have evolved the ability to recognise more than one bond linkage or monosaccharide type, such as plant cell wall recognising CBMs, which are often able to recognise cellulose and hemicelluloses. Occasionally, individual CBMs may also possess multiple binding sites, allowing a multivalent interaction with the target ligand, although this is rare (Boraston *et al.*, 2004).

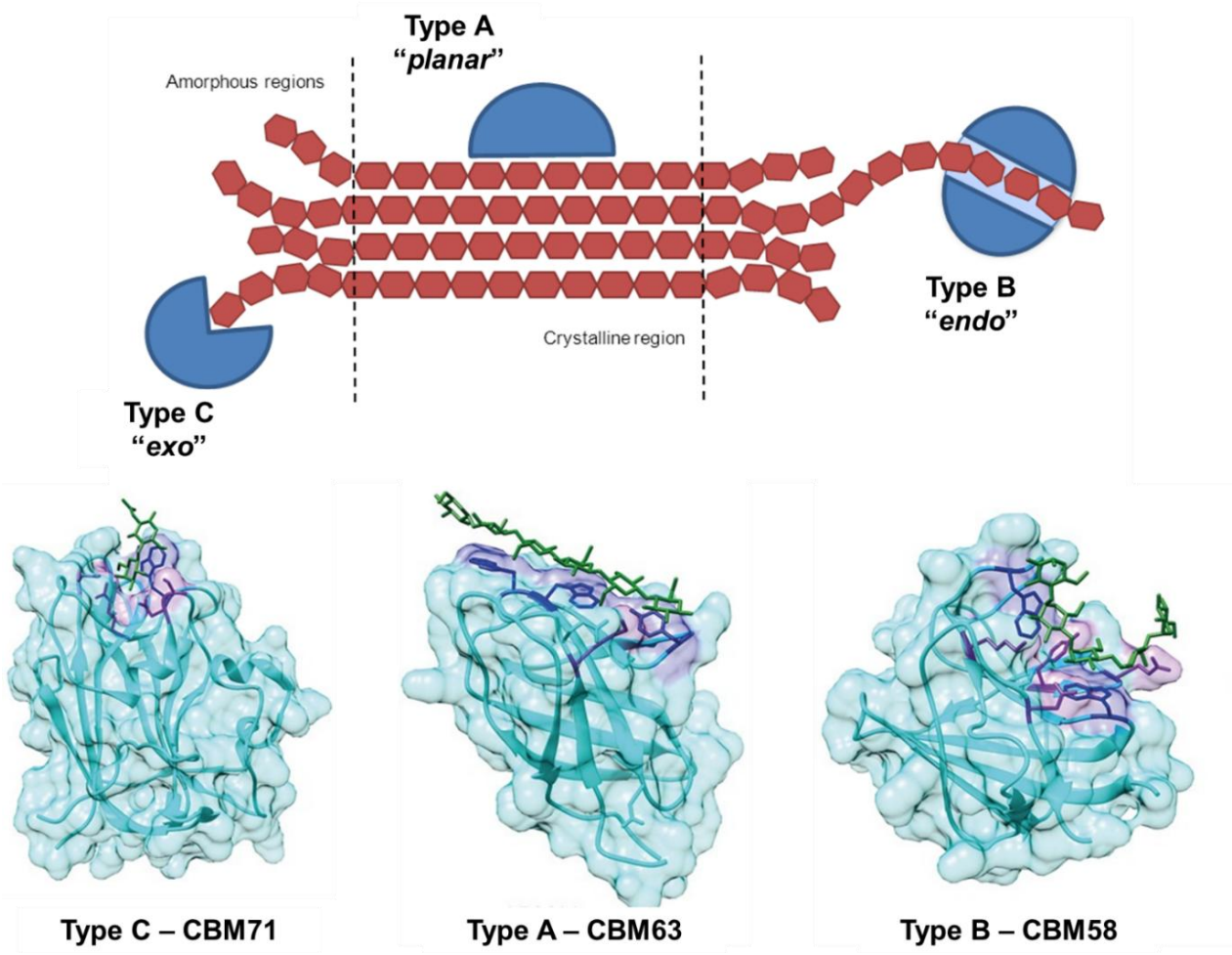


Figure 1-13: Three main binding site types of Carbohydrate Binding Modules. A) Schematic of the main CBM types binding to different regions of a polysaccharide substrate. B) Example topology of three binding site types, Type A: CBM63 from the *Bacillus subtilis* expansin (EXLX1) in complex with cellohexaose (PDB ID 4FER), Type B: CBM58 from the *Bacteroides thetaiotaomicron* α-amylase with maltopentaose (PDB ID 3K8L), Type C: CBM71 from the *Streptococcus pneumoniae* β-galactosidase in complex with β-D-galactopyranosyl-1,4-N-acetyl-D-glucosamine (PDB ID 4CUB). Adapted from https://www.cazypedia.org/index. php/Carbohydrate-binding_modules and Armenta *et al.* (2017).

1.4 Glycan Degradation by Members of the Bacteroidota Phylum – Polysaccharide Utilisation Loci

1.4.1 Importance of *Bacteroides* in Polysaccharide Degradation

Bacteroidota and Bacillota are the most abundant phyla present in the human and monogastric gut, as well as in the rumen (D'Argenio & Salvatore, 2015). The major

genera constituting gut Bacteroidota are *Alistipes, Bacteroides, Parabacteroides, Odoribacter* and *Prevotella* (Ndeh & Gilbert, 2018; D'Argenio & Salvatore, 2015). Members of the Bacteroidota phylum, in particular *Bacteroides* and *Prevotella* spp., have high capacity for degradation of complex polysaccharides due to a vast repertoire of CAZymes in their genomes and as such are considered general glycan utilisers (Kaoutari *et al.*, 2013; Lapébie *et al.*, 2019).

1.4.2 Polysaccharide Utilisation Loci and the Bacteroidota Sus-like Paradigm

Bacteroidota from wide-ranging environmental niches employ similar strategies in glycan degradation, allowing this broad capacity for complex carbohydrate breakdown. This involves co-localised, co-upregulated gene clusters known as Polysaccharide Utilisation Loci, or PULs. Genes within PULs encode proteins which collectively confer ability to breakdown a particular glycan, or closely related group of glycans (Martens, Koropatkin, *et al.*, 2009; Martens *et al.*, 2008). Together, the gene products of PULs orchestrate the detection, enzymatic degradation and transport of complex polysaccharides (Martens, Koropatkin, *et al.*, 2009). The first characterised PUL was the starch utilisation system (Sus) of *Bacteroides thetaiotaomicron* (**Figure 1-14**), discovered and named by the Salyers Lab from studies of starch breakdown by this species (D'Elia & Salyers, 1996; Tancula *et al.*, 1992; Anderson & Salyers, 1989; Foley *et al.*, 2016). This gene cluster consists of 8 genes: *susRABCDEFG*, where SusR is an inner membrane spanning regulator which senses maltose and triggers upregulation of the sus genes. SusDEF lipoproteins function in binding of starch to the outer membrane, such that starch can be hydrolysed by SusG α-amylase. Resulting maltooligosaccharides are subsequently transported into the periplasm by SusC transporters. SusD-like proteins are specialised glycan binding proteins which occur

are associated with with SusC-like TonB Dependent Transporters. Once in the periplasm, maltooligosaccharides are further degraded into mono- and disaccharides by SusA, neopullulanase and SusB α-glucosidase (Foley *et al.*, 2016; Martens, Koropatkin, *et al.*, 2009). This paradigm for glycan utilisation occurs across all members of the Bacteroidota phylum and has allowed for designation of PULs (Martens, Koropatkin, *et al.*, 2009; Sonnenburg *et al.*, 2005). All PULs contain adjacent gene pairs encoding SusC- and SusD-like proteins, which function in concert with specific CAZymes, regulators and glycan binding proteins for the degradation of a particular glycan (Martens, Koropatkin, *et al.*, 2009). Bioinformatic tools now exist to predict PULs within the genomes of Bacteroidota based on the presence of tandem SusC/SusD-like pairs (Terrapon *et al.*, 2015). Predicted and experimentally characterised PULs can be found within the database, PULDB, containing 68500 PUL predictions from 2065 Bacteroidota species (Terrapon *et al.*, 2018).

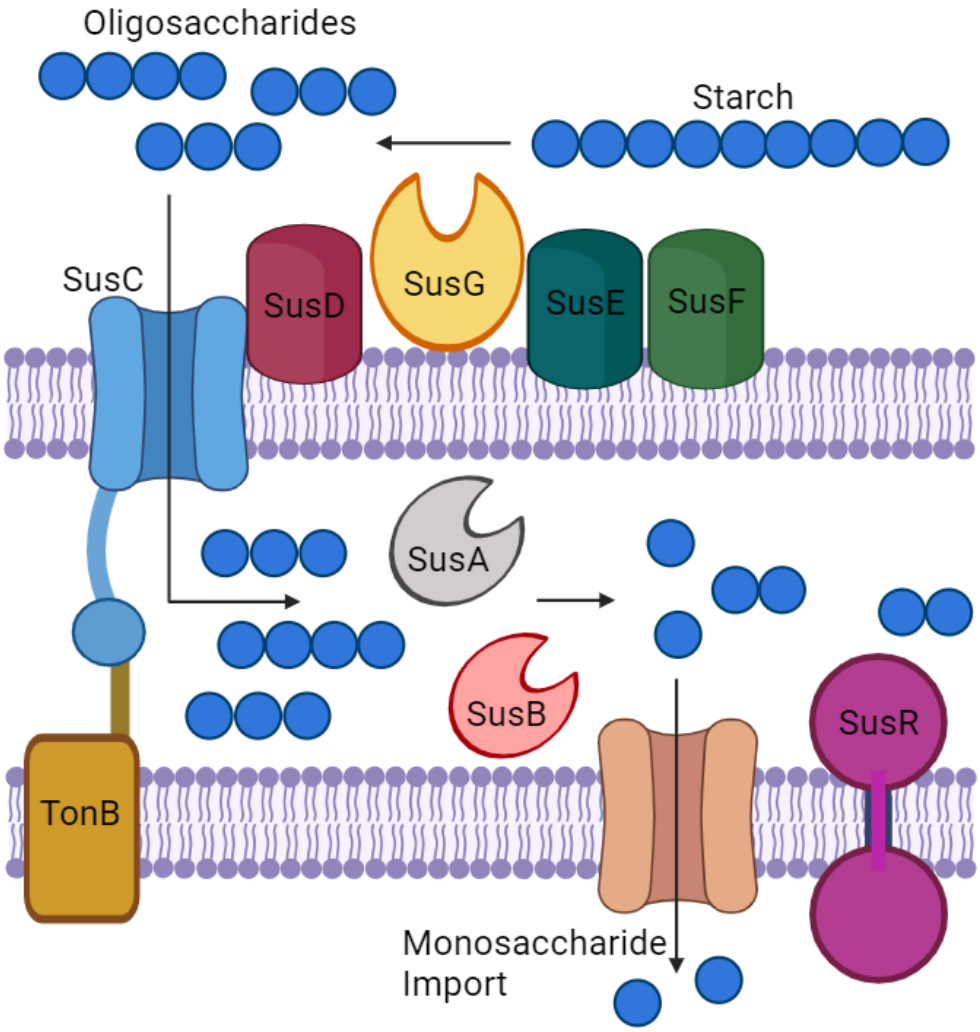


Figure 1-14: Schematic model of the *B. thetataiotaomicron* Starch Utilisation System (Sus). The *B. theta* Sus consists of 8 proteins, SusRABCDEFG, including 3 outer membrane binding proteins, Sus C, E and F, which bind starch at the cell surface. Hydrolysis of starch to oligosaccharides by the membrane anchored SusG α-amylase generates oligos small enough for transport into the periplasm via SusC, a TonB dependent membrane transporter. In the periplasm, SusA neopullulanase and SusB α-glucosidase degrade oligosaccharides into glucose, which is subsequently transported into the cytoplasm. SusR is an inner membrane spanning regulator which recognises maltose in the periplasm and upregulates SusABCDEFG in response. Adapted from Foley *et al.* (2016).

1.4.3 Role of SusC and SusD like Proteins in Nutrient Acquisition by Bacteroidota

SusC/D-like pairs are the hallmark of Bacteroidota PULs and as such play a critical role in the functioning of glycan degradation systems. Studies have shown that SusC

and SusD protein molecules are tightly associated and form a complex at the cell surface, both components of which are essential for binding to substrate as well as for transport into the cell periplasm (Reeves *et al.*, 1997; Cho & Salyers, 2001; Glenwright *et al.*, 2017).

SusD-like proteins are involved in oligo- or polysaccharide binding at the cell surface, often the inaugural step in glycan degradation. SusD-like outer membrane lipoproteins are secreted proteins consisting of an α-helical fold with an N-terminal lipid tail allowing them to associate with the outer membrane. These SusD-like proteins are also involved in passing oligosaccharides to SusC-like TonB dependent receptors for transport into the periplasm (Bolam & Koropatkin, 2012; Koropatkin *et al.*, 2008). SusD-like proteins are significantly larger and structurally divergent from other known carbohydrate binding proteins. Their conservation in PULs demonstrates their functional importance, playing a distinct role in nutrient acquisition compared to other glycan binding proteins (Foley *et al.*, 2016).

SusC-like proteins are a novel type of TonB-dependent transporter (TBDT) that work in concert with a SusD-like lipoprotein (**Figure 1-15**). A vast variety of TBDTs are encoded within Bacteroidota genomes. Whilst PUL-encoded SusC-like TBDTs are involved in the uptake of oligosaccharides, others may be important for the uptake of various other nutrients (Bolam & van den Berg, 2018; Pollet *et al.*, 2021). TBDTs have many roles in gram negative bacterial species, although have mostly been studied in *Escherichia coli,* including sensing and adapting to environmental stimuli, uptake of iron siderophores and vitamin B12, and can be important in pathogenicity (Koebnik, 2005; Torres *et al.*, 2001). TonB dependent transporters consist of an outer membrane spanning 22-stranded β-barrel which is completely occluded by a plug domain which inserts into the barrel from the periplasmic space (**Figure 1-15**) (Bolam & van den

Berg, 2018). TonB-dependent transport relies on TonB, an inner membrane associated protein, which spans the periplasm. An inner membrane complex is formed between TonB, ExbB and ExbD, and ExbBD is responsible for generation of proton motive force which energises the active transport process through the TBDT. The N-terminal outer membrane spanning helix of TonB interacts with the ExbBD complex allowing for the transduction of this energy to the outer membrane where the C-terminal domain of TonB is associated with part of the TonB dependent transporter known as the TonB box, tranducing energy to the TBDT for the active transport process (Noinaj *et al.*, 2010). Although TonB-dependent transporters are ubiquitous across gram negative bacteria, association of SusC-like TBDTs with SusD-like proteins seems to be unique to Bacteroidota PULs (Pollet *et al.*, 2021).

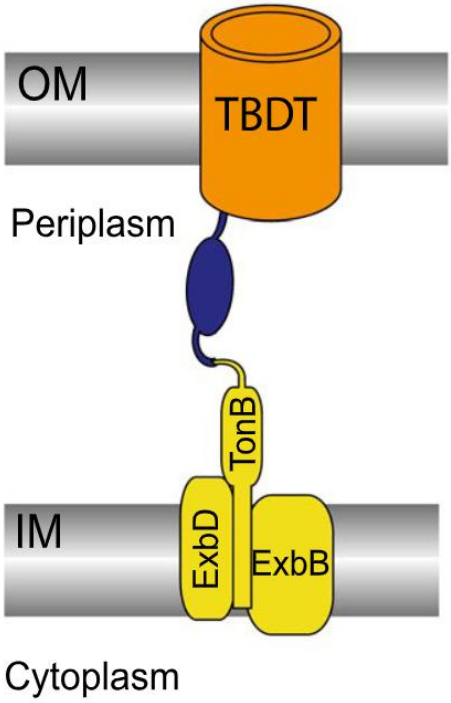


Figure 1-15: Classic architecture of TonB dependent transporters found in gram negative bacteria. Outer membrane spanning barrel domain is displayed in orange, and the plug domain in dark blue. This TonB dependent receptor is anchored to the inner membrane TonBExbBD complex, which is responsible for providing energy for the active transport process. From Pollet *et al.* (2021).

Structural studies have demonstrated that SusCD-like may complexes exist as $SusC_2D_2$ tetramers, creating a twin barrel structure which mediates substrate uptake via a “pedal bin” mechanism (**Figure 1-16**). SusC transporters form the barrel, capped by their cognate SusD. It is proposed that in the absence of ligand, SusD-like proteins move away from the SusC transporter in a hinge-like fashion, opening the lid and exposing the binding site to the external environment to facilitate ligand binding (Glenwright *et al.*, 2017; Gray *et al.*, 2021; Bolam & van den Berg, 2018). Closing of the SusC lid in the presence of ligand, with subsequent Ton-B mediated removal of the plug domain, allows transport of oligosaccharide into the periplasm (**Figure 1-16**). Recent studies show that in levan and dextran PUL-encoded utilisation systems of *B. thetaiotaomicron* outer membrane components assemble on the core SusCD transporters, forming stable glycan-utilising machinery assembled on the outer membrane (White *et al.*, 2023).

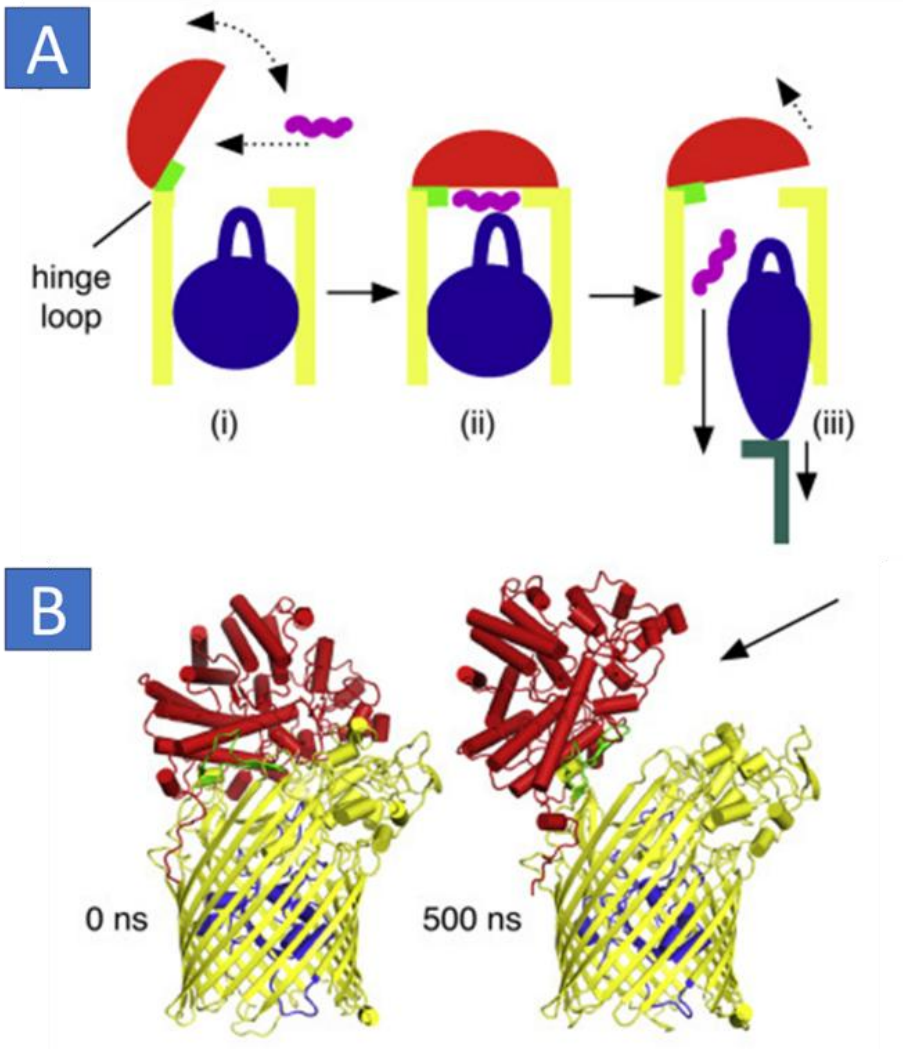


Figure 1-16: Architecture of SusCD-like complexes from *B. thetataiotaomicron.* A. Schematic representation of the pedal bin mechanism for nutrient acquisition by SusCD-like complexes (i) shows the SusD lid in its dynamic open state. (ii) upon ligand binding the SusD lid closes on the SusC transporter. (iii) Ton B complex moves the plug domain, leading to formation of a channel in the SusC transporter, allowing diffusion of the substrate into the periplasm. In the absence of substrate, the SusD-lid is reset to the open state (i). B. Molecular dynamics simulation of susCD complex showing SusD lid opening in the absence of ligand. For clarity, only one SusCD dimer is shown. Adapted from Bolam & van den Berg (2018).

1.4.4 Regulation of Bacteroidota Polysaccharide Utilisation Loci

The ability to sense and respond to carbohydrate stimuli via the upregulation of PUL components is critical for competitive advantage within the highly populated gut ecosystem. PUL regulation occurs via a positive feedback loop in which early

degradation products produced by low-level constitutive transcription of PUL components (Sonnenburg *et al.*, 2005) activates the sensor system, leading to upregulation of the PUL, and hence enhanced degradation of a specific polysaccharide (McKee *et al.*, 2021). This continues until the polysaccharide target is depleted, or a polysaccharide of higher preference to the bacterium is detected (McKee *et al.*, 2021). PUL regulation is commonly mediated by one of three mechanisms: SusR-like sensor/regulators, extracytoplasmic function (ECF) sigma/anti-sigma factors or, most commonly, hybrid two component systems (HTCS) (Martens, Koropatkin, *et al.*, 2009). Although SusR was essential for regulation of the first discovered sus system (D'Elia & Salyers, 1996), the role of such proteins in regulation of glycan degradation appears to be an exception, as homologues are not commonly found in documented PULs (Grondin *et al.*, 2017). SusR is an inner membrane spanning receptor, with a sensor domain exposed to the periplasm and a cytoplasmic DNA binding domain. SusR detects presence of maltose and starch-derived oligosaccharides and binds to promoter regions of Sus genes, activating their expression (D'Elia & Salyers, 1996).

HTCS are cytoplasmic membrane-spanning proteins consisting of the components of a classical two component system: a transmembrane sensor histidine kinase and a DNA-binding response regulator, fused to a large periplasmic carbohydrate-sensing domain (Lowe *et al.*, 2012; Sonnenburg *et al.*, 2006). *B. thetaiotaomicron* possesses 32 HTCS encoding genes, 17 of which are located within PULs, indicating their importance in regulation of polysaccharide degradation (Sonnenburg *et al.*, 2006). Structural studies into a PUL-encoded *B. thetaiotaomicron* HTCS involved in regulation of heparin and heparan sulphate degradation demonstrate a method of transmembrane signalling distinct to that of classical two component systems (Lowe

et al., 2012). The third type of regulatory systems utilised by gut Bacteroidota for control of polysaccharide degradation are ECF sigma-anti-sigma pairs, commonly found in PULs targeting host-derived polysaccharides, such as O-glycans (Martens *et al.*, 2008). It is proposed that SusC-like proteins are coupled to the periplasmic region of an anti-sigma factor. SusC-mediated oligosaccharide transport sends a transmembrane signal via the anti-sigma factor, releasing its cognate ECF sigma factor into the cytoplasm, allowing its interaction with RNA polymerase and subsequent transcription of promoters and upregulation of PUL components (Martens, Roth, *et al.*, 2009).

1.4.5 Surface Glycan Binding Proteins (SGBPs) with Polysaccharide Utilisation Loci

As well as SusD-like binding proteins, many Bacteroidota PULs contain genes encoding surface glycan binding proteins (SGBPs). These proteins are involved in recruitment of glycans to the outer membrane, where specific GH or PL enzymes can initiate degradation. SGBPs often form a complex with SusC/SusD-like proteins, such that they are cell surface localised and exposed to the extracellular environment. SGBPs seem to be a lot more divergent in sequence and structure in comparison with SusD-like glycan binding proteins and may recognise substrates identical or complementary to that of the SusD-like protein (Tamura & Brumer, 2021; Grondin *et al.*, 2017).

Despite diversity in SGBP sequences, they are N-terminal lipoproteins with extended multidomain structures composed of tandem Ig like domains, binding substrate in one or more distal C-terminal domains (**Figure 1-17**) (Grondin *et al.*, 2017; Tauzin *et al.*, 2016).

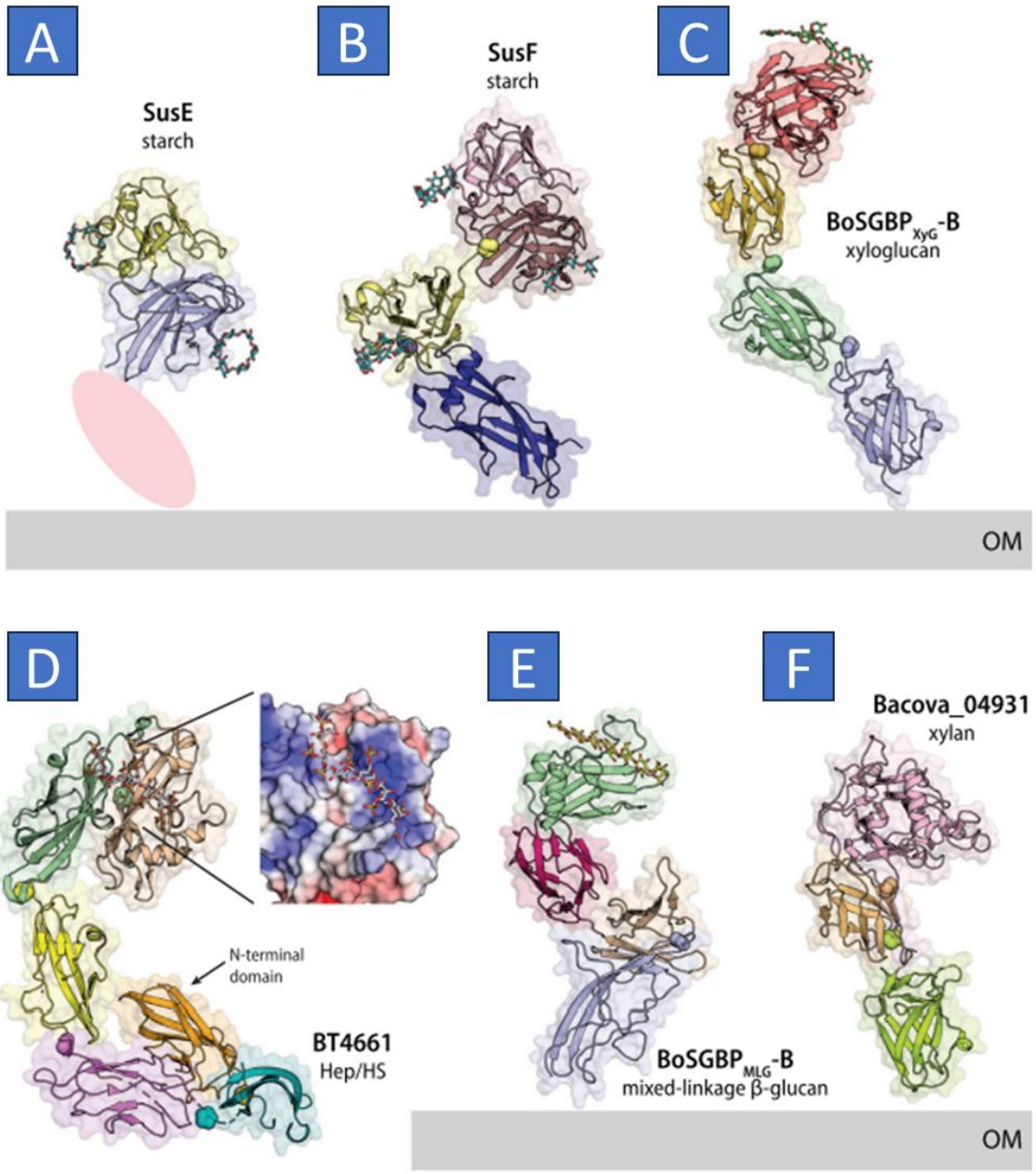


Figure 1-17: Structures of Surface Glycan Binding Proteins. Several SGBP structures have been solved to date by X-ray crystallography. Individual domains are shown in different colours. A) SusE starch binding protein from *B. thetaiotaomicron*. B) SusF starch binding protein from *B. thetaiotaomicron*. C) BoSGBPXyG-B xyloglucan binding protein from *B. ovatus*. D. BT4661 heparin/ heparin sulphate binding protein from *B. thetaiotaomicron*. E) BoSGBPMLG-B mixed linkage beta glucan binding protein from *B. ovatus*. F) Bacova_04931 xylan binding protein from *B. ovatus*. From Tamura & Brumer (2021).

1.4.6 Utilisome Structures

Outer membrane components of *Bacteroides* PULs have been shown to exist in stable complexes termed utilisomes, specialised for the efficient capture, processing and transport of a particular glycan type (White *et al.*, 2023). These utilisomes complexes consist of at least one SGBP, implicated in the initial capture of glycans, surface exposed GH enzymes for initial processing via cleavage of the glycan into shorter oligosaccharides, before passage to the SusC/D complex and transport into the periplasm via the pedal bin mechanism described in Section 1.4.3 (**Figure 1-18**). White *et al.* (2023) solved the structure of PUL-encoded outer membrane utilisomes for levan and dextran utilisation systems from *B. thetaiotaomicron* via single particle cryo-EM, demonstrating the movement of SusD- lids from the open configuration in the absence of ligand, to closed state upon ligand binding (**Figure 1-18**).

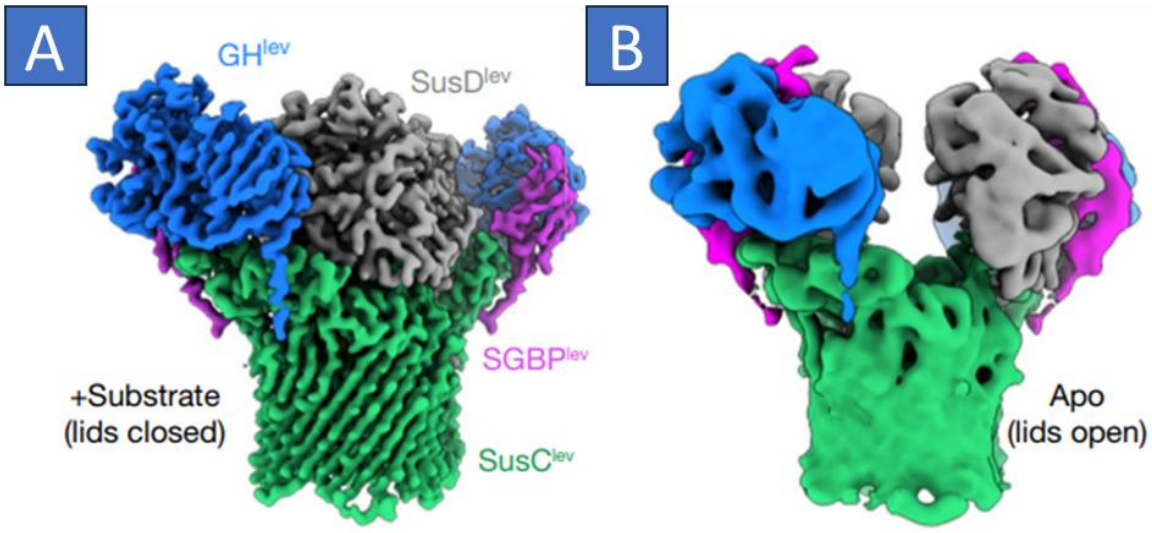


Figure 1-18: Structure of Levan utilisome from *B. thetaiotaomicron* demonstrating conformational shift upon ligand binding. A) Cryo-EM structure of the levan utilisome with short chain fructo-oligosaccharides (FOS) (DP 8–12) at 2.9 Å resolution shows the SusD lid in the closed configuration. B) Equivalent view of the apo utilisome at about 6 Å resolution, showing that the SusD- like lids are open in the absence of substrate. Adapted from White *et al.* (2023).

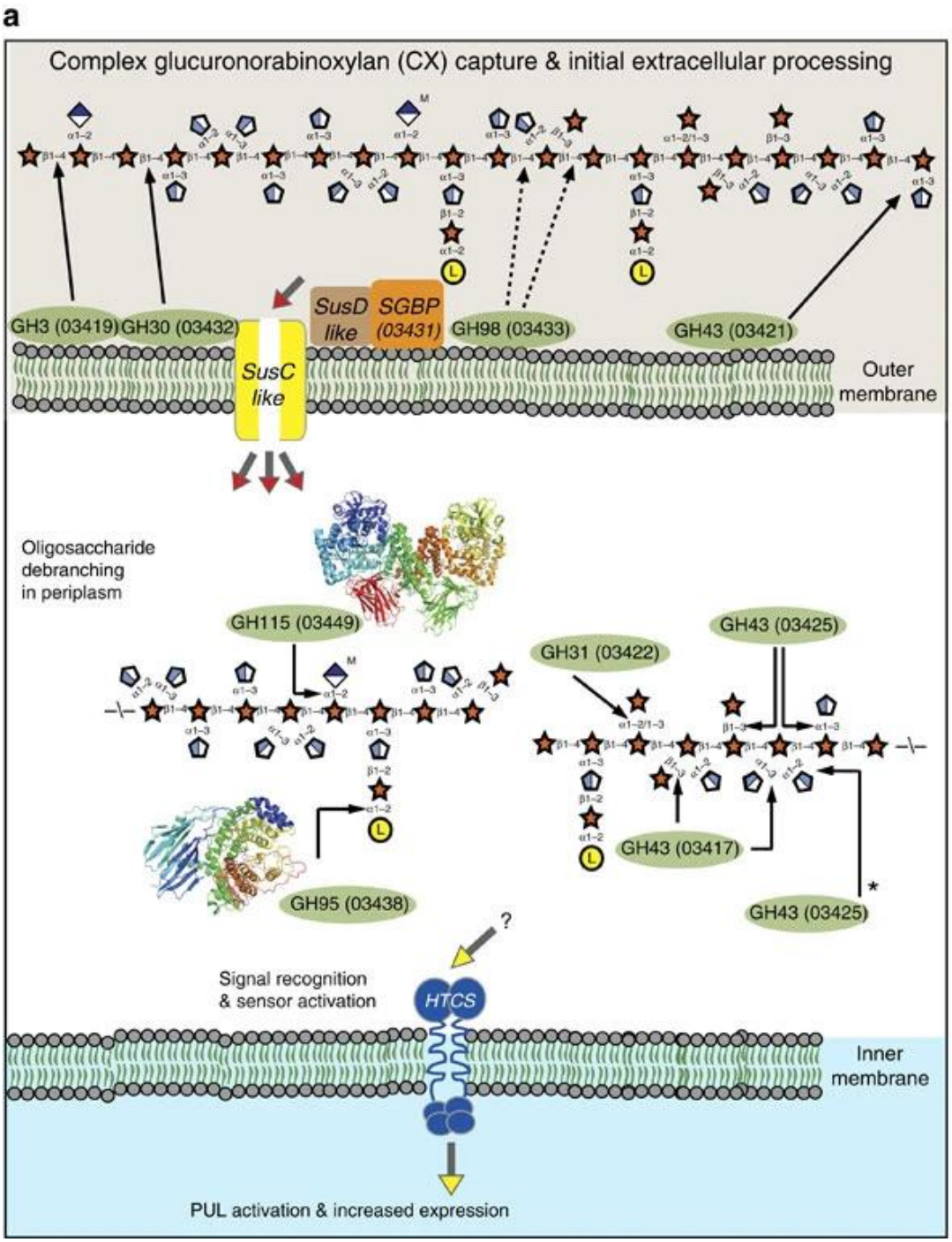
1.4.7 Xylan Targeting Polysaccharide Utilisation Loci of Gut *Bacteroides* Species

Ability to degrade xylan substrates seems commonplace, although not ubiquitous, amongst *Bacteroides* members of the human gut microbiota. Multiple *Bacteroides* species have demonstrated xylan degrading ability, including *B. eggerthii, B. cellulosilyticus, B. intestinalis, B. ovatus*, and *B. xylanisolvens* (Dodd *et al.*, 2011; Rogowski *et al.*, 2015; Despres *et al.*, 2016; Wang *et al.*, 2016; McNulty *et al.*, 2013). These species possess PUL-encoded xylan utilisation systems of varying complexities providing capabilities to degrade different types of xylans, as shown in **Figure 1-19**.

B. ovatus has evolved the capacity to degrade plant-derived hemicelluloses inaccessible to some other gut Bacteroidota such as *B. thetaiotaomicron* (Martens *et al.*, 2011). Xylans are a major component of plant hemicelluloses, and some of the most abundant naturally occuring polysaccharides. High abundance of xylans in dietary components, particularly cereal grains such as corn, oats, wheat and rye, meaning that xylans are an important target for degradation by members of the human gut microbiota as xylan utilisation likely confers a major competitive advantage.

Analysis of the *B. ovatus* ATCC 8483 genome demonstrated 112 putative PULs, 2 of which were activated by xylans (Martens *et al.*, 2011) (**Figure 1-20**). Xylan utilisation by gut generalist *B. ovatus* has since been extensively characterised (Rogowski *et al.*, 2015). This xylan degradation system is functionally dynamic, allowing divergent responses to different forms of xylan (**Figure 1-19**). Both xylan targeting PULs, a larger PUL consisting of 26 genes, and a smaller PUL of 10 genes (**Figure 1-20**), referred to as PUL-XylL and PUL-XylS respectively, are upregulated during growth on AXs such as WAX. Growth on complex CX lead to activation of PUL-XylL alone, whereas on linear xylooligosaccharides and GX only PUL-XylS was upregulated. This implies that

the smaller PUL encodes apparatus necessary for degradation of simpler xylans, the gene products of PUL-XylL are required for sequestration, degradation and metabolism of more complex xylans (Rogowski *et al.*, 2015). Growth profiles of *B. ovatus* mutant strains show that deletion of PUL-XylS doesn't influence growth on GAX but prevented growth on simpler xylans. On the contrary, knockout of PUL-XylL prevented growth on complex xylans but not on simpler AX or GX.



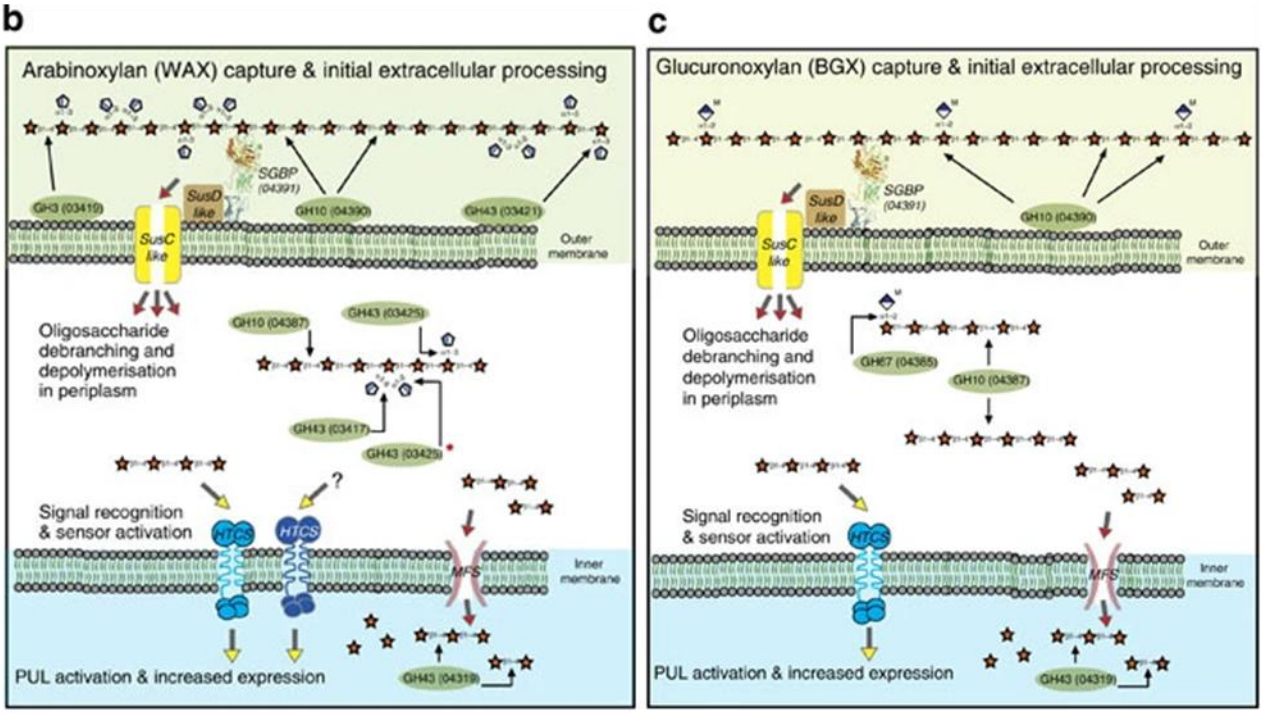


Figure 1-19: Overview of *B. ovatus* polysaccharide utilisation loci (PUL) encoded xylan utilisation systems. (A) Model for the degradation of complex glucuronoarabinoxylans (GAX), (B) arabinoxylans (AX) and (C) glucuroxylans (GX). Xylans are bound at the cell surface by outer membrane SusD-like and surface glycan binding protein (SGBP) lipoproteins and subsequently hydrolysed by membrane-bound surface endo-xylanases such as those from glycoside hydrolase family 10 or 98. Surface generated xylooligosaccharides are transported across the outer membrane into the periplasm via SusC-like ton B dependent receptor for further debranching and degradation. Specific oligosaccharides then bind to a hybrid two component system (HTCS) sensor-regulator leading to PUL activation and increased expression. Figure from Rogowski *et al.* (2015).

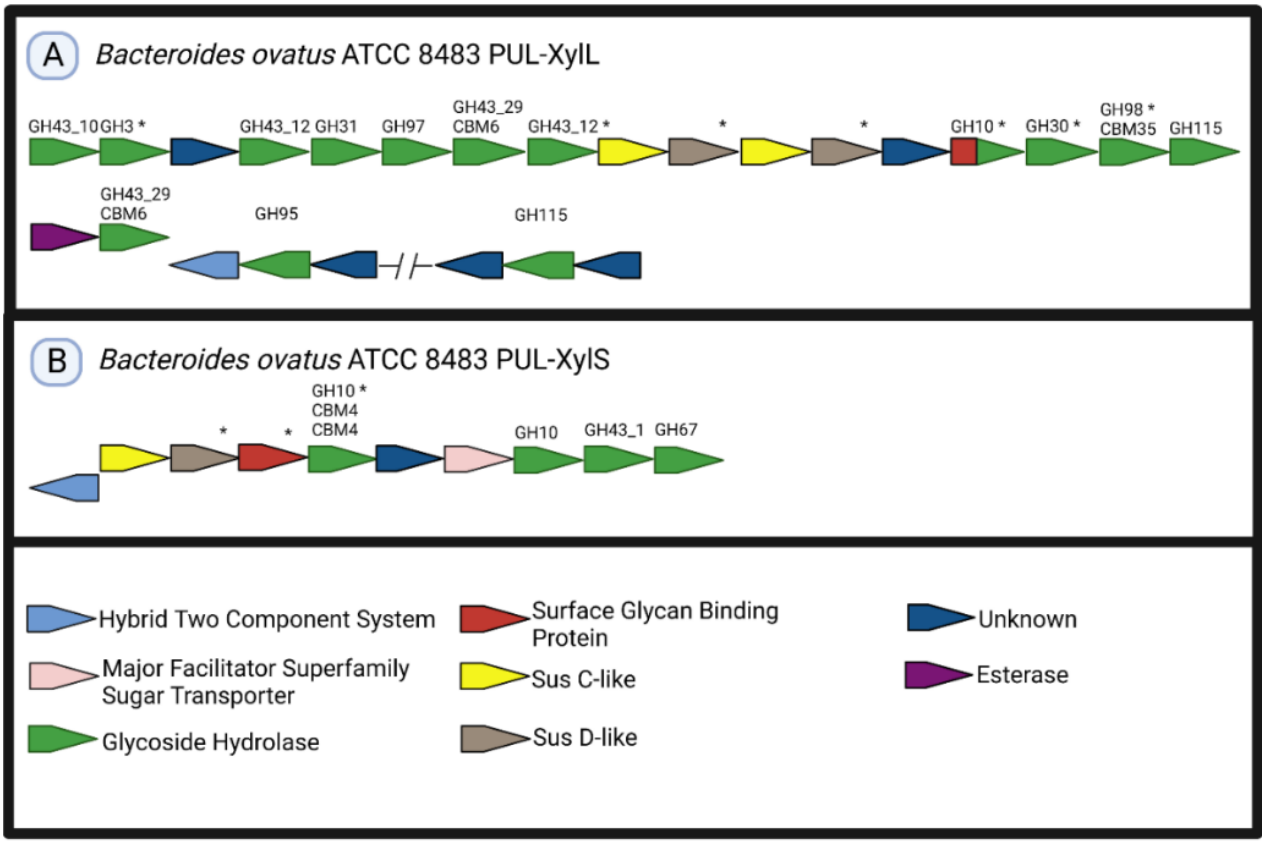


Figure 1-20: Schematic of the two PULs encoded by *B. ovatus* which are upregulated during growth on xylan and demonstrate differential expression on different types of xylan. (A) *B. ovatus* large xylan PUL (PUL-XylL), highly upregulated during growth on complex, highly decorated xylans. (B) Small xylan PUL, PUL-XylS, upregulated during growth on simple xylans. Proteins located at the cell surface are denoted by an asterisk. Adapted from Rogowski *et al.* (2015).

Analysis of xylan PULs within *B. ovatus* and *B. xylanisolvens* strains showed high levels of conservation in PUL-XylL in 22 out of 23 examined strains. These PUL-XylL homologues can be subdivided into two groups: those possessing the three genes GH30_8, GH98 and GH115 as in *B. ovatus* ATCC 8483 and those in which these 3 genes are replaced by a GH5_21 encoding gene (Rogowski *et al.*, 2015). As discussed previously, GH5_21 enzymes from *B. xylanisolvens* possess endoxylanase activity against AX and GAX (Rogowski *et al.*, 2015; Aspeborg *et al.*, 2012). Interestingly, conservation in PUL-XylL in other Bacteroidota genomes is very low, in

particular the distribution of GH98 encoding genes(Rogowski *et al.*, 2015). However, a xylan PUL homologous to PUL-XylS was identified in ruminal *Prevotella bryantii* using a transcriptomic approach (Martens *et al.*, 2011; Dodd *et al.*, 2010). Overall, although xylan utilisation is not uncommon amongst gut Bacteroidota, the dynamic xylan-utilisation machinery of *B. ovatus* is particularly adapted for degradation of divergent xylan structures, likely conferring a competitive advantage in the gut environment.

1.5 Glycan Degradation by Human and Animal Gut Bacillota

1.5.1 Importance of Bacillota in Dietary Glycan Degradation

Gram positive Bacillota are abundant in the human and animal gut microbiota, and many Bacillota play an important role in dietary glycan degradation, in particular those belonging to the genera *Butyrivibrio, Ruminococcus, Roseburia, Eubacteria* and *Clostridia* (Flint *et al.,* 2008; La Reau & Suen, 2018; Kelly *et al.,* 2010; La Rosa, Leth, *et al.,* 2019; Leth *et al.,* 2023). Whereas Bacteroidota are considered generalists with respect to polysaccharide utilisation, often encoding vast repertoires of CAZymes targeting diverse glycans, Bacillota tend to be more specialised for the degradation of specific glycan types, as first noted by Salyers *et al.* (1977). This is reflected in the number of encoded CAZymes in members of the HGM averaging 137.1 per genome for Bacteroidota compared to 39.6 for Bacillota (Kaoutari *et al.*, 2013). This difference in glycan targeting preference demonstrates nutritional specialisation of Bacillota species, potentially enhancing survival within particular gut niches or under nutritional stress. On the whole, dietary fibre degradation by Bacillota is less well elucidated compared to that of *Bacteroides*, perhaps due to the fastidious nature of Bacillota, making laboratory studies challenging. However, their unique glycan targeting strategies and high nutritional specialisation make gut Bacillota an important area of study to improve our overall understanding of plant cell wall breakdown in the gut. Different families seem to possess one of two strategies for glycan degradation: multi-enzyme complexes known as cellulomes, or polysaccharide utilisation loci somewhat analogous to those of Bacteroidota.

1.5.2 Gram Positive Polysaccharide Utilisation Loci

As research began into glycan degradation by gut Bacillota it was noted that certain members possess clusters of glycan degradation related genes similar to those of Bacteroidota PULs, except for the lack of the outer membrane SusC/SusD homologs essential for classification of loci as PULs (O. Sheridan *et al.*, 2016; La Rosa *et al.*, 2022) These gene clusters were first identified in *Roseburia* spp. and *Eubacterium rectale* and denoted as Gram positive polysaccharide utilisation loci (GpPUL) (O. Sheridan *et al.*, 2016) (**Figure 1-22**). GpPULs were defined as encoding at least a CAZyme, a transport system and a transcriptional regulator (**Figure 1-22**).

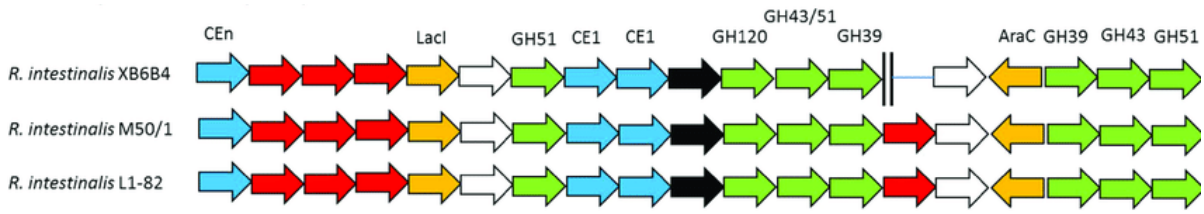


Figure 1-21: Schematic representation of Gram positive PULs (GpPULS) involved in arabinoxylan degradation by Roseburia intestinalis spp. Glycoside hydrolases are coloured in green, carbohydrate esterases in blue, ABC transporter systems in red, uncharacterised transporter encoding genes are black and regulatory genes in yellow. Adapted from Sheridan *et al.* (2016).

Unlike Bacteroidota PULs, these GpPULs often contain CAZymes which can target different glycan types, meaning that functional predictions are more challenging (O. Sheridan *et al.*, 2016). GpPULs possess one of three classes of transporters: ATP binding cassettes (ABC) transporters, phosphotransferase system (PTS) transporters or major facilitator superfamily (MFS) transporters, although ABC transporters seem to be most commonly occurring in Bacillota (Cockburn & Koropatkin, 2016). Interestingly, no homologues of *Bacteroides* PUL binding protein encoding genes (SusD-like or SGBP homologues) are found in *E. rectale/ Roseburia* gpPULs, therefore polysaccharide binding seems to involve CBM domains present in the GH enzymes,

in concurrence with the high frequency of presence of CBMs in Bacillota GH enzymes (O.Sheridan *et al.*, 2016). A model of a gpPUL encoded glycan utilisation system is shown in **Figure 1-23.** GpPULs are ubiquitous in the human gut, but also those of other monogastric animals, ruminants and insects, indicating their importance in dietary and host glycan degradation (La Rosa *et al.*, 2022).

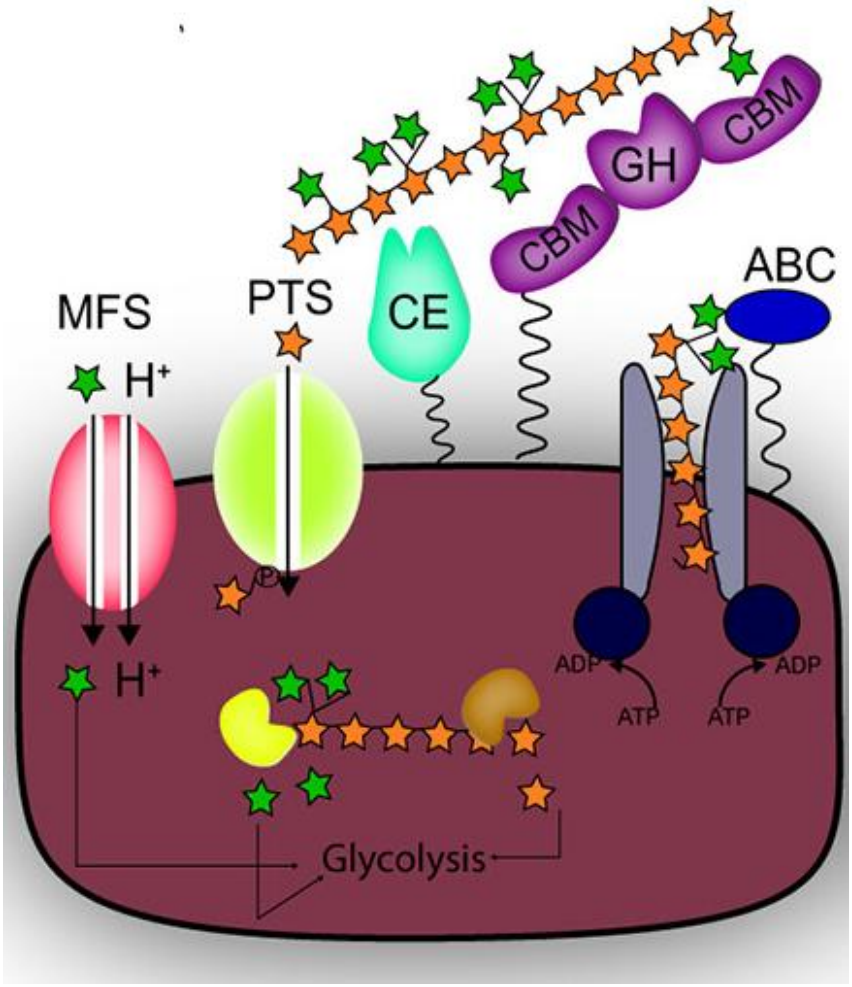


Figure 1-22: Model of a Gram-positive polysaccharide utilisation loci (GpPUL) encoded glycan utilisation system. In some *Roseburia* and *Eubacteria* species, glycan transport systems, commonly ABC transporters, are co-upregulated with CAZymes, including carbohydrate esterases and glycoside hydrolases. Catalytic domains within GpPULs are often associated with multiple carbohydrate binding module (CBM) domains. Adapted from Glowacki & Martens (2021).

1.5.3 Cellulosomal Systems

Within some intestinal niches, such as the rumen or the equine caecum and hindgut, the ability to degrade insoluble fibre construes a competitive advantage and is critical for energy supply for the host organism. In these environments, cellulose, the major constituent of plant cell wall material, is the predominant carbon source (Cockburn & Koropatkin, 2016). Certain Bacillota, such members of the *Ruminococcus* and *Clostridia* genera, from a range of ecological niches, produce multi-enzyme complexes known as cellulosomes, which bring enzymes, CBMs and substrate into close proximity, allowing the synergism for the highly efficient degradation of cellulose and cellulose-associated hemicelluloses (**Figure 1-24**) (Artzi *et al.*, 2017). The first cellulosome was described in the highly cellulolytic anaerobe *Clostridium thermocellum* (Lamed *et al.*, 1983) and was shown to allow selective, specific binding of the bacterium to its cellulose substrate (Bayer *et al.*, 1983). Cellulosomes consist of two major building block components: enzymes of ancillary proteins containing dockerin modules, and scaffoldins which are cohesin-containing structural proteins (**Figure 1-24**). Cohesin-dockerin interactions are some of the strongest non-covalent interactions in nature, and selective binding of different cohesins and dockerins dictates how the cellulosome can be assembled, the overall cellulosome archecture and integration of different enzymes into the structure (Bayer *et al.*, 2008). Scaffoldins may also contain dockerins to bind to other scaffoldins, and/or CBMs which target the cellulosome complex to specific sites on the plant cell wall substrate (Bayer *et al.*, 1998; Artzi *et al.*, 2017). Cellulosomes can be cell-free structures or attached to the cell envelope via intricate domain-anchoring proteins (Bayer *et al.*, 1998; Artzi *et al.*, 2017). Interestingly, once the structure of the inaugural *C. thermocellum* cellulosome had been elucidated, subsequent work on the structures of cellulosomes produced by other bacterial

species demonstrated differences in the size, complexity and architecture, leading to multi-enzyme complexes which vary greatly between species, and perhaps within species in different environmental conditions. The cellulosome of ruminal *Ruminococcus flavefaciens* 17 is particularly complex and intricate, requiring many complex biochemical and omic studies to elucidate (Bayer *et al.*, 2008).

Within cellulosome systems there is at least one primary scaffoldin, which incorporates enzymes directly into the cellulosome (Fontes & Gilbert, 2010). In bacteria which express cell-surface cellulosomes, multiple anchoring scaffoldins are responsible for the attachment of primary scaffoldins and their enzyme complement to the cell surface. These anchoring scaffoldins possess three repeated domains with homology to S layer (SLH) proteins, responsible for attachment of the scaffoldin to the cell surface via interactions with the peptidoglycan layer or secondary cell wall polysaccharides (Kosugi *et al.*, 2002; Fontes & Gilbert, 2010). Additionally, some very complex cellulosome structures contain adaptor scaffoldins which provide an extra link between anchoring and primary scaffoldins (Bayer *et al.*, 2008). An overview of cellulosome structure is provided in **Figure 1-24**.

CBMs within cellulosomal systems mediate the tight attachment of cellulosome systems to plant cell wall material. This primarily occurs via type A CBM3 domains within primary scaffoldins (**Figure 1-24**) which bind tightly to the surface of crystalline cellulose (Rincon *et al.*, 2007). Upon binding of the cellulosome complex to microcrystalline cellulose within the cell wall, further type B and C CBM domains within cellulosomal enzymes play a role in fine-tuning substrate recognition by bringing catalytic regions in close proximity to their specific substrate within the cell wall, enhancing catalytic efficiency (Fontes & Gilbert, 2010).

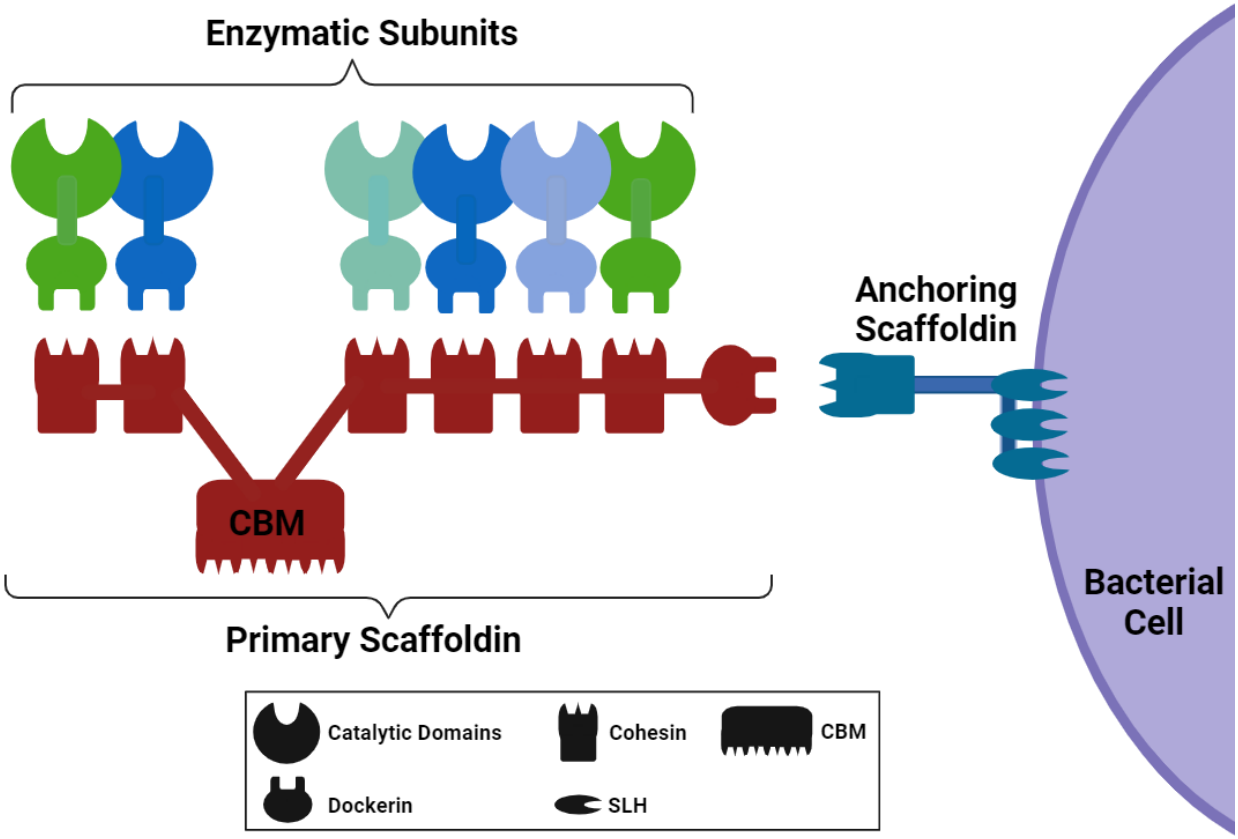


Figure 1-23: **Schematic representation of the structure of a simple cellulosome.** Cellulosomes are large multi-protein complexes specialised for the degradation of insoluble polysaccharides such as cellulose. Cellulosomes can be cell-bound or free structures. Dockerins appended to CAZyme subunits interact with cohesin molecules on the primary scaffoldin. Dockerins on this primary scaffoldin bind to cohesins on anchoring scaffoldin tethering the cellulosome to the cell surface.

Unlike in the Bacteroidota PUL paradigm, cellulosome producing Bacillota do not encode all proteins involved in glycan regulation, transportation and degradation in single co-upregulated loci. However, genes involved in cellulosome formation are commonly found in cellulosome related gene clusters. Simple cellulosome systems may be encoded by enzyme-linked gene clusters on the chromosome, which consist of a single primary scaffoldin encoding gene, followed downstream by genes encoding for dockerin-bearing enzymes (**Figure 1-25B**) (Sakka *et al.*, 2010; Bayer *et al.*, 2008). In complex cellulosome systems, where there can be multiple anchoring, adapter and

primary scaffoldins, scaffoldin encoding genes are organised into multiple chromosomal scaffoldin gene clusters (**Figure 1-25B**) (Bayer *et al.*, 2008). In many cellulosome-encoding species, such as in *Clostridium cellulolyticum,* CAZymes and scaffoldin encoding genes are found in multiple operons distributed across the genome. As such, complex global regulatory mechanisms are required to control transcription of these genes (Xu *et al.*, 2015).

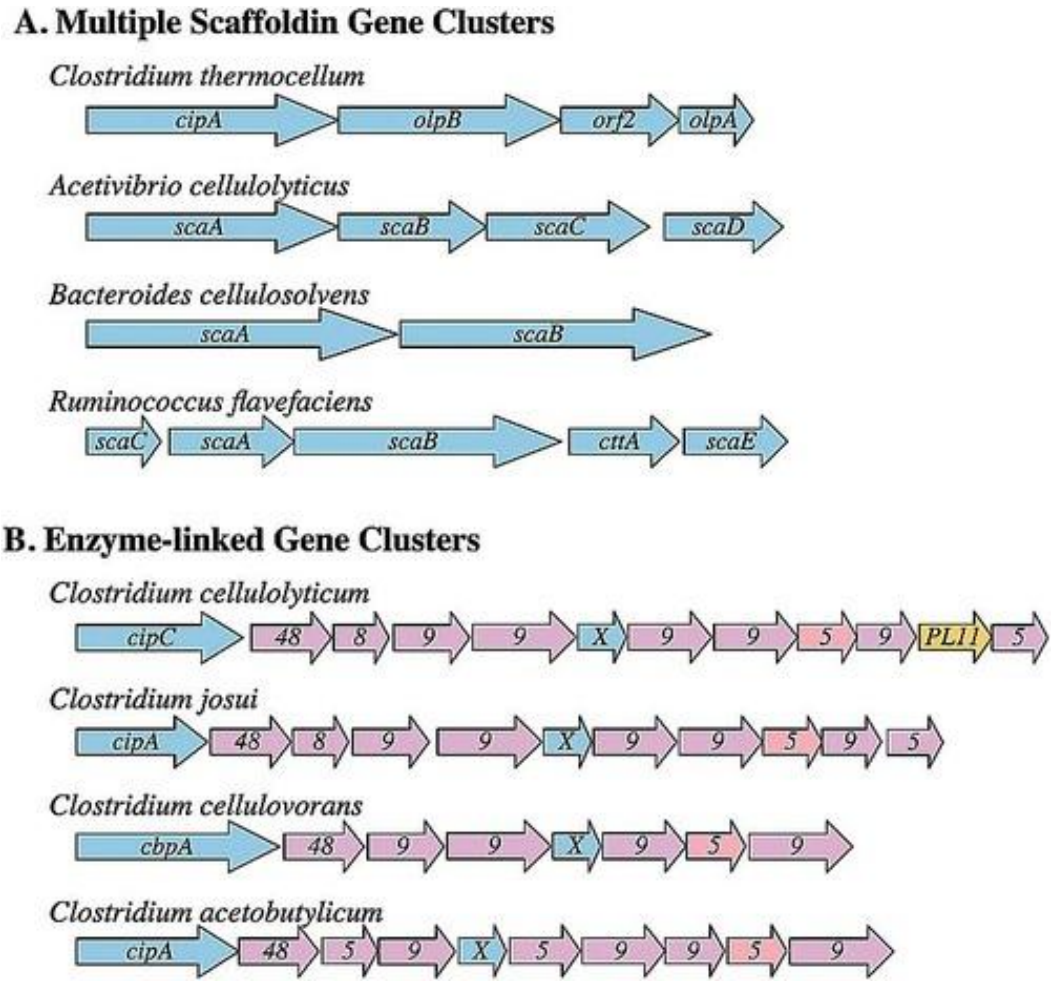


Figure 1-24: Cellulosome-related gene clusters from the genomes of different cellulosome-producing Bacillota. A) Scaffoldin gene clusters encode multiple scaffoldin encoding genes in tandem, allowing for the assembly of large, complex cellulosome systems. B) Enzyme linked gene clusters consisting of one primary scaffoldin encoding gene, followed by multiple genes encoding dockerin-containing CAZymes. From Bayer *et al.* (2008).

Although little is understood about the regulation of cellulosome transcription and translation, multiple studies show that upregulation of CAZymes is growth-substrate dependent, including for ruminal strain *R. flavefaciens* (Maša Vodovnik *et al.*, 2013; Berg Miller *et al.*, 2009), *Clostridium cellulovorans* (Han *et al.*, 2003) and for the model cellulose utiliser *Clostridium cellulolyticum* (Xu *et al.*, 2013). Transcriptomic studies of *C. cellulolyticum* suggested that transcriptional regulation of cellulosome-associated CAZymes occurred via carbon catabolite repression, where intracellular accumulation of intermediates from polysaccharide degradation prevents further transcription of CAZyme encoding genes (Xu *et al.*, 2013). Non-cellulosomal genes expression, including hemicellulases and associated transporters, are regulated via the detection of soluble extracellular sugars by two component systems (Xu *et al.*, 2013). Interestingly, in this during growth on glucose alone, cellulase enzymes were highly upregulated indicating that glucose inhibits carbon catabolite repression, enhancing cellulolysis. On the contrary, transcriptomic studies of *C. cellulovorans* showed significant reduction of cellulase transcripts in the presence of di- or monomeric sugars (Han *et al.*, 2003) This suggests differential mechanisms of CAZyme regulation between different cellulosome producing species.

1.5.4 Glycan Degradation by Cellulolytic Human Gut Symbiont *Ruminococcus champanellensis*

Although Bacillota are abundant in the human gut, due to the short passage times and the lack of reliance on degradation of dietary fibre for energy provision, cellulolytic species are rare. The first characterised cellulose utiliser from the human gut was *R. champanellensis* 18P13, closely related to the ruminal cellulolytic strain *R. flavefaciens* ATCC 19208 (Chassard *et al.*, 2012). This *R. champanellensis* strain was isolated by

Chassard *et al.* via its ability to utilise filter paper as a sole carbon source. Furthermore, genome analysis of *R. champanellensis* demonstrated that this was the first strain from the human gut encoding a wide range of cellulosomal elements including cohesion and dockerin containing proteins, of which there were 24 and 60 respectively (Ben David *et al.*, 2015). Due to its unique cellulolytic activity, *R. champanellensis* has been previously referred to as a keystone species in the human gut (Moraïs *et al.*, 2016), as its ability to degrade cellulose within dietary plant-based material may improve the accessibility of other plant cell wall hemicelluloses to members of the HGM, hence enhancing their survival within the intestinal environment. However, due to the relatively low dietary fibre content, lack of reliance on cellulose degradation for energy provision, short passage time through the human gastrointestinal tract, *R. champanellensis* is unlikely to function as a true keystone species, particularly in industrialised and western populations.

To date, studies of this *R. champanellensis* strain have been limited to prediction of cellulosome structure based on cohesin-dockerin binding studies (Ben David *et al.*, 2015), and basic analysis of the activity of recombinantly expressed dockerin containing CAZymes (Moraïs *et al.*, 2016). Cohesin-dockerin binding profiles revealed that *R. champanellensis* cellulosomes can be complex cell bound structures, or simple cell-free structures composed of a single scaffoldin (Ben David *et al.*, 2015). The 24 cohesins were found on 11 different scaffoldin proteins, termed ScaA to ScaK (**Figure 1-26**). Studies of cohesin-dockerin binding interactions revealed that up to 11 enzymes can be brought together into a large cellulosome complex. Interactions and potential cellulosome architectures are demonstrated in **Figure 1-26**.

According to CAZy, *R. champanellensis* possesses 129 CAZyme modules, more than half of which are within dockerin-containing proteins (Ben David *et al.*, 2015). 57 of

these CAZyme modules are glycoside hydrolases, belonging to 27 different families, many of which are cellulases from GH families 5 and 9 (Ben David *et al.*, 2015). *R. champanellensis* encodes 25 dockerin-containing GH enzymes. These include 10 cellulases, 4 xylanases, 3 mannanases, 2 xyloglucanases, 2 arabinofuranosidases, 2 arabinases and 1 β-glucanase. Activities of these enzymes were confirmed via analysis of reducing end production from a range of cellulose and hemicellulose substrates (Moraïs *et al.*, 2016). This suggests that cellulosome complexes may possess activity against a range of different plant cell wall components, but it is not yet known whether, or how the presence of different glycans dictate expression of cellulosomal components, or the size and architecture of assembled cellulosomes.

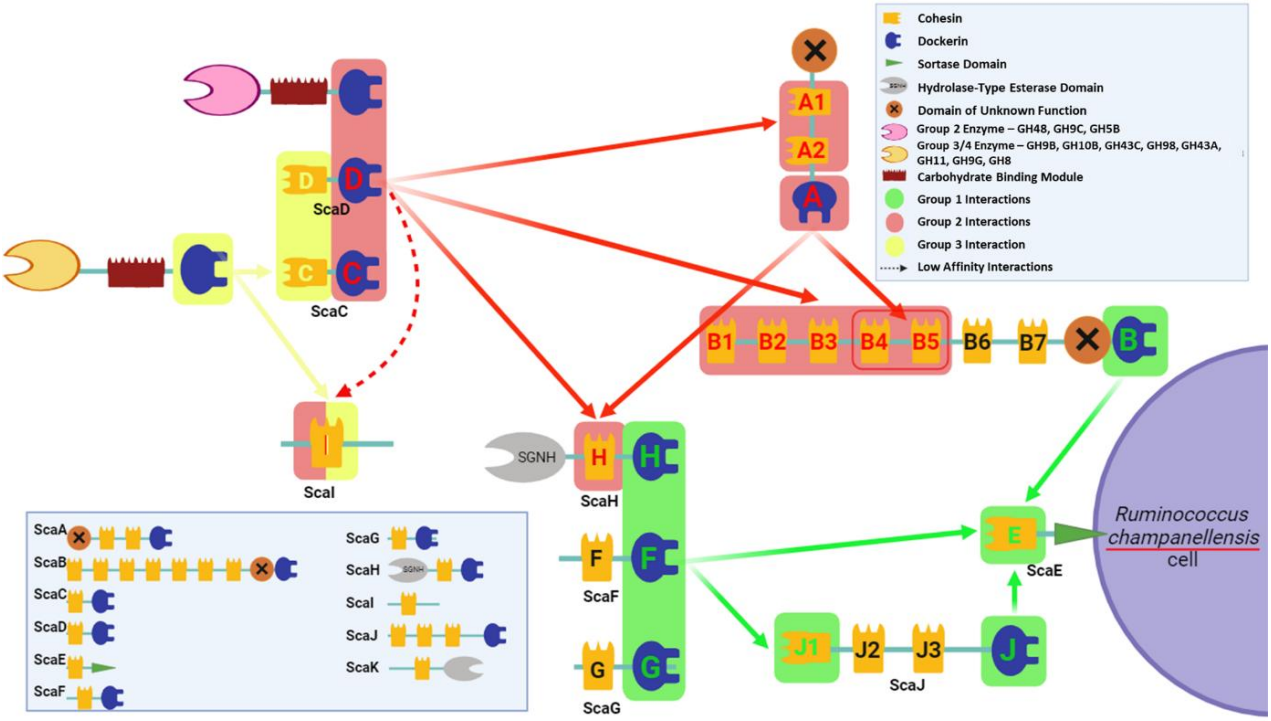


Figure 1-25: Schematic representation of the complex cellulosome system of *Ruminococcus champanellensis*. *R. champanellensis* is proposed to produce cell free and cell bound cellulosomes, involving up to eleven different scaffoldins, termed ScaA to ScaK. Cell free cellulosomes involve interactions with ScaI, whilst cell bound cellulosomes are attached to the cell surface via anchoring scaffoldin ScaE. Interactions between cohesins and dockerins are grouped into 3 different types, colour coded red, green, and yellow, dictating binding specificity and hence the overall cellulosome structure. Specificities of Sca F and Sca G are as yet undetermined. CAZymes are characterised into group 2 or group 3/4, dependent on the specificity of the integrated dockerin domain. Adapted from Ben David *et al.* (2015).

1.6 Aims and Objectives

Plant cell wall material is incredibly complex and consists of many tightly associated polysaccharides. The human and animal gut microbiota play a critical role in degradation of these macromolecular structures via the production of CAZymes, in particular GHs, which target specific polysaccharide components, ultimately contributing to nutrient supply to the host organism. Understanding the mechanisms of glycan degradation by members of the human and animal gut microbiota is critical to improve nutritional understanding and promote health. This study investigates how key members of the HGM degrade complex plant cell wall materials with a particular focus on the abundant and nutritionally important xylans.

In order to achieve this aim the project was split into four main objectives:

1. Establish xylan-utilising capacities of members of the human and animal gut microbiota via bioinformatic analysis and assessing growth characteristics.

   Although the xylan-utilisation strategies of some strains have been well characterised (Rogowski *et al.*, 2015; Dodd *et al.*, 2010), for many others they are unknown or poorly understood. Here, we elucidate the link between genome encoded CAZymes and the ability to utilise different xylans for growth.

2. Understand the structural and functional mechanisms employed by *B. ovatus* GH98 enzyme in the degradation of complex glucuronoarabinoxylans.

   Complex GAXs such as CX are common in the human diet and is recalcitrant to degradation. A *B. ovatus* encoded GH98 enzyme possesses endoxylanase activity solely against complex GAX. Here we

demonstrate functional and structural studies into the specificity of this unique activity.

3. Explore the functional binding properties of GH98-associated CBM35 domains.

   Bioinformatic analysis suggests that GH98 enzymes possessing endo-xylanase activity has a conserved C-terminal CBM35 domain. Here, we assess binding activity of the CBM35 from *Bacteroides ovatus* GH98 and characterise the binding site and specificities.

4. Understand how whole plant cell wall material is utilised by important members of the human gut microbiota.

   Studies into glycan utilisation strategies and preferences when individual bacterial strains are exposed to complex plant cell wall materials are very limited. Here, transcriptomic studies of *B. ovatus* and *R. champanellensis* grown on a complex corn-derived substrate help elucidate these complex bacterial-substrate interactions.

2. Materials and Methods

2.1 Bioinformatic analysis

2.1.1 Prediction of Polysaccharide Utilisation Loci (PUL) in Gut Bacteroidota

Predicted and characterised PULs were identified using PULDB (Terrapon *et al.*, 2018), a database containing literature-derived and predicted PULs from species across the Bacteroidota phylum (www.cazy.org/PULDB). PULs were then manually verified using the Neighbourhood Regions tool in JGI IMG/M (mg.jgi.doe.gov) (Chen *et al.*, 2023).

2.1.2 Phylogenetic Analysis of GH98 Protein Sequences

Complete GH98 sequences listed within the CAZy database (www.cazy.org) (Cantarel *et al.*, 2009) were downloaded from NCBI and used in phylogenetic analysis of the GH98 family. Phylogenetic trees were created using MEGA 11 (Tamura *et al.*, 2021), using the maximum likelihood method with 100 bootstraps. Trees were then edited using International Tree of Life (IToL) software (Letunic & Bork, 2021).

2.1.3 Sequence Similarity Networks (SSN)

For SSN analysis, the complete sequences encoding GH98 enzymes were extracted from the CAZy database. The amino acid sequences were then used to generate a SSN using the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) (Gerlt *et al.*, 2015). 540 GH98 sequences were analysed by SSN with an alignment score of 120. The SSN data were visualised using Cytoscape 3.6 (Su *et al.*, 2014).

2.1.4 GH98 Protein Domain Analysis

2.1.4.1 Protein Domain Annotation

Protein domain composition was analysed using the SMART protein annotation resource (Letunic & Bork, 2018) and Interpro (Paysan-Lafosse *et al.*, 2023). dbCAN

was subsequently used for confirmation of CAZyme module annotation (Zheng *et al.*, 2023).

2.1.4.2 Prokaryotic Signal Peptides

Signal peptides from protein sequences were predicted using SignalP 6.0 (Teufel *et al.*, 2022).

2.1.5 Alphafold2 for Protein Structure Prediction

AlphaFold2 predictions were either retrieved from the AlphaFold2 database or run using a version of AlphaFold2 hosted on google colab (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb#scrollTo=pc5-mbsX9PZC) (Jumper *et al.*, 2021).

2.1.6 Creation of phylogenetic heat maps

Phylogenetic trees were created using maximum likelihood approach with the phangorn package in R. Initially, a distance matrix was computed, and a starting tree was generated using the Neighbor-Joining (NJ) algorithm. This tree was then used as a starting point for ML optimisation. Maximum likelihood tree estimation was performed using the Whelan and Goldman (WAG) substitution model. To assess pairwise distances between taxa based on the resulting ML tree, cophenetic distances were calculated. These distances were then visualized using a clustered heatmap via the **pheatmap** package, with average linkage clustering and a diverging color scale ranging from blue to red. Row and column clustering distances were set to match the tree-based cophenetic distances.

2.2 Bacterial Strains and Plasmids

2.2.1 *Escherichia coli* strains

All *Escherichia coli* strains used in this study are listed in **Table 2-1.**

Table 2-1: *Escherichia coli* strains used in the present study

***E. coli* Strain**	**Description**	**Application**	**Reference**
BL21 (DE3)	F - dcm ompT hsdS(rB-mB-) gal l(DE3)	Protein Expression	Studier and Moffat,1986
Tuner (DE3)	F – ompT hsdSB (rB – mB –) gal dcm lacY1 (DE3)	Protein Expression	Novagen
One Shot™ TOP10	F´mcrA (mrrCB-hsdRMS-mrr) 80lacZ M15 lacX74 deoR recA1 araD139 (ara-eu)7697 galU galK rspL endA1 nupG	DNA Ligation and Replication	Invitrogen
B834 (DE3)	RB- , mB- , gal- , met-	SeMet Protein Expression	Wood, 1996

2.2.2 Bacteroidota and Bacillota strains

All *Bacteroides*. *Prevotellaceae* and *Ruminococcus* strains used in this study are listed in **Table 2-2.**

Table 2-2: Bacterial strains used in the present study

Bacteria Strain	**Genotype**	**Description**
Bacteroides caccae	ATCC 43185	Type strain, human isolate
Bacteroides cellulosilyticus	DSM 14838	Type strain, human isolate
Bacteroides cellulosilyticus	WH2	Human isolate
Bacteroides clarus	DSM 22519	Type strain, human isolate
Bacteroides eggerthii	DSM 20697	Type strain, human isolate
Bacteroides finegoldii	DSM 17565	Type strain, human isolate
Bacteroides fluxus	DSM 22534	Type strain, human isolate
Bacteroides intestinalis	DSM 17393	Type strain, human isolate
Bacteroides ovatus	ATCC 8483	Type strain, human isolate
Phocaeicola plebeius	DSM 17135	Type strain, human isolate

Bacteroides salyersiae	DSM 18765	Type strain, human isolate
Bacteroides xylanisolvens	DSM 18836	Type strain, human isolate
Ruminococcus champanellensis	DSM 18848	Type strain, human isolate
Segatella copri	DSM 18205	Type strain, human isolate
Segatella bryantii	B14	Type strain, bovine isolate
Segatella albensis	DSM 11379	Type strain, ovine isolate
Xylanibacter muris	DSM 103722	Murine isolate
Palleniella intestinalis	DSM 103738	Murine isolate
Paraprevotella clara	DSM 19731	Type strain, human isolate
Paraprevotella xylaniphila	DSM 19681	Type strain, human isolate

2.2.3 Plasmids

E. coli expression plasmids used in the present study are shown in **Table 2-3.** Plasmids encoding Bacova_03417, Bacova_03425, Bacova_03438, Bacova_03449 and Cj98 were previously cloned in the same lab by Artur Rogowski.

Table 2-3: Vectors used in the present study

Plasmid	Size (kb)	Description	Application	Reference
pET21a	5.44	amp^r, T7, lac $lacI^q$	Expression Vector	Studier, 1991 (Studier, 1991)
pET28a	5.37	kan^r, T7, lac $lacI^q$	Expression Vector	Novagen
pGEX-6P-1	4.98	amp^r, lac $lacI^q$	Expression Vector	Amersham

2.3 Microbiology

2.3.1 Chemicals, Enzymes and Media

Unless stated otherwise, all media, antibiotics, chemicals and enzymes were sourced from Sigma Chemical Company (Sigma Aldrich, UK). 18.2 MΩ, ultrapure

water was used in all experiments. Water was used as the solvent for all solutions unless otherwise stated.

2.3.2 Centrifugation

Bacterial cultures 100-1000 mL were centrifuged at 5,000 rpm in 500 mL centrifuge tubes (Nalgene) for 10 minutes using a Beckman J2-21 centrifuge with a JA-10 rotor at 4 °C. Bacterial cultures of 50 mL were centrifuged at 15,000 rpm for 30 minutes in 50 ml centrifuge tubes (Nalgene) using a Beckman J2-21 centrifuge with a JA-25.5 rotor at 4 °C. Cultures of 2-50 mL were harvested by centrifugation at 5,000 rpm in 25 ml universal tubes (Sterilin) using Hettich Mikro 220R benchtop centrifuge with fixed angle rotor at 4 °C. Reactions < 2 mL were centrifuged in Eppendorf tubes (2 or 1.5 mL) at 13,000 rpm using a Heraeus Instruments Biofuge pico benchtop centrifuge.

2.3.3 Bacterial Growth Substrates

2.3.3.1 Xylans and Cellulose

Oat spelt xylan (nano) and birchwood xylan (BWX) were commercially sourced from Sigma Aldrich, UK, Low viscosity wheat arabinoxylan (WAX) and rye arabinoxylan (RAX) were purchased from Megazyme International. Corn xylan (CX) was kindly provided by Madhav Yadav (United States Department of Agriculture) following preparation as described by Rogowski *et al.* (2015). Cellulose sources used in this study were Avicel microcrystalline cellulose powder and Whatman no.1 cellulose filter paper.

2.3.3.2 Preparation of Corn Derived Growth Substrate

Corn derived growth substrate, representative of a whole cell wall material, was prepared from ground corn kernels. Ground corn underwent a pepsin-pancreatin digest as described by the Boisen Method (Bindelle *et al.*, 2007; Boisen & Fernández,

1997) to simulate removal of starches, peptides, proteins and lipids which would occur in the stomach and small intestine. 25 mL 0.1M pH 6 phosphate buffer solution and 10 mL 0.2 M HCl solution were added to 0.5 g of ground corn and the pH adjusted to 2.0. 1 mL of 25 mg mL^{-1} porcine peptin (2000 FIP-U/g) (Merck) was added, sealed and incubated at 37 ºC for 2 hours with gentle agitation. Following pepsin hydrolysis, 10 mL 0.2M pH6.8 phosphate buffer solution was added with 5 mL 0.6 M NaOH solution. pH was adjusted to 6.8, prior to addition of 1 mL fresh pancreatin solution at 100 mg mL^{-1}. The mixture was then incubated at 37 ºC for 4 hours with gentle agitation. Following incubation, the mixture was dialysed into dH_2O for 48 hours, or until complete removal of monosaccharides, as judged by running on TLC, Corn derived substrate was then frozen to -80 °C and lyophilised in a Christ Alpha 1-2 Freeze Drier at -60 °C, prior to autoclaving and storage at -20 ºC.

2.3.4 Media and Growth Conditions

2.3.4.1 Growth Media and Antibiotics

Media used for the culture of bacterial strains are listed in **Table 2-4.** A range of antibiotics were used for the selection of desired transformants or appropriate bacterial species, which are listed in **Table 2-5**.

Histidine-Hematin was formed from 1.9 mM Hematin in 0.2 M Histidine solution. pH was adjusted to 8.0 with 10 M NaOH and dissolved overnight with shaking at 37 °C.

Table 2-4: Growth media used in the present study.

Media	Application	Composition	Details
Luria Broth (LB)	*E. coli* growth	25g/L LB powder (Sigma Aldrich, UK)	Prepared in ultra-pure water, pH adjusted to 7.2, autoclaved
Brain Heart Infusion (BHI)	*R. champanellensis* + *Prevotellaceae* overnight growth	37g/L BHI powder (Sigma Aldrich, UK)	Prepared in ultra-pure water, autoclaved, supplemented with 0.1% (v/v) Hematin-Histidine
Tryptone Yeast Extract Glucose (TYG)	*Bacteroides* spp. overnight growths	See ***Table 2-6***	Prepared in ultra-pure water, autoclaved, supplemented with 0.1% (v/v) Hematin-Histidine
Yeast Casitone Fatty Acid (YCFA)	*R. champanellensis* + *Prevotellaceae* selective growth	See ***Table 2-8***	Prepared in ultra-pure water, autoclaved, supplemented with 0.1% (v/v) Hematin-Histidine and 5% rumen fluid
Bacteroides Minimal Media (MM)	*Bacteroides* spp. selective growths	See ***Table 2-7***	Prepared in ultra-pure water, filter sterilised
SelenoMethionine Complete Medium	*E. coli* growth for SeMet incorporated protein expression	21g/L SelenoMet Medium Base (Molecular Dimensions) supplemented with 40mg/L L-selenomethionine (Anatrace)	Prepared in ultra-pure water, autoclaved prior to addition of 50mL SelenoMet nutrient mix solution

Table 2-5: Selective antibiotics used in growth media

Antibiotic	Working Concentration	Application	Storage
Ampicillin	100 mg mL^{-1}	pET21a, pGEX-6P-1	-20 °C
Kanamycin	10 mg mL^{-1}	pET28a	-20 °C

2.3.4.2 Sterilisation

All solutions, media, and glassware were sterilised by autoclaving using an Astell Hearson 2000 Series Autoclave or a Prestige© Medical Series 2100 Clinical Autoclave at 121 ºC, 32 or 15 lb / inch-2 for 20 min. Solutions were also filter sterilised using sterile Millipore filter disc (0.22 µm pore size) (Supor® Acrodisc® 3.2) and a suitable sterile syringe (Plastipak®, Becton Dickinson).

2.3.4.3 *E. coli* Media and Growth Conditions

E. coli strains listed in **Table 2.1** were grown in liquid LB media (**Table 2.4**) at 37 ºC with aeration in a rotary shaker at 180 rpm with appropriate antibiotic, as indicated in **Table 2-5**. Agar plates were made by adding 2% w/v bacteriological agar (Sigma Aldrich, UK) to LB, autoclaving, and allowing to cool to 55 ºC prior to addition of selective antibiotic. 25 mL molten LB-agar was poured into sterile plastic Petri dishes (ThermoFisher, UK) to form plates. *E. coli* colonies were grown by spreading culture across the plate with a sterile L- spreader. Plates were incubated agar-down at 37 ºC for 16 hours in a stationary incubator.

2.3.4.4 *Bacteroides* spp. Media and Growth Conditions

Bacteroides spp. listed in **Table 2-2** and *P. plebeius* were routinely grown statically in 5 mL Tryptone Yeast Extract Glucose (TYG) media (**Table 2-6**) under anaerobic conditions using an A35 anaerobic cabinet (Don Whitley Scientific, UK) at 37 ºC.

Table 2-6: Components of Tryptone Yeast Extract Glucose (TYG) Agar

Component	Amount per Litre
Tryptone Peptone	10 g
Yeast Extract	5 g
Glucose	2 g
Cysteine (Free base)	0.5 g
1 M KPO_4 pH 7.2	100 mL
Vitamin K solution, 1 mg mL^{-1}	1 mL
TYG salts (MgSO4 0.5 g L^{-1}, NaHCO3 10 g L^{-1}, NaCl 2 g L^{-1})	40 mL
0.8% $CaCl_2$	1 mL
$FeSO_4$, 0.4 mg mL^{-1}	1 mL
Resazurin, 0.25 mg mL^{-1}	4 mL

For xylan utilisation experiments *Bacteroides* spp. were grown at in *Bacteroides* Minimal Media (**Table 2-7**) supplemented with 5 mg mL^{-1} final concentration of xylan. 2X *Bacteroides* Minimal Media was filter sterilised and subsequently combined in equal parts with 10 mg mL^{-1} chosen xylan substrate dissolved in dH_2O.

Table 2-7: Components of 2X *Bacteroides* Minimal Media. For composition of 10X *Bacteroides* Salts, Balch's Vitamins, Trace Mineral Solution, Purine/ Pyrimidine Solution and Amino Acid Solution see Supplementary Table X.

Component	Amount per 50 mL
10X *Bacteroides* Salts	10 mL
Balch's Vitamins	1 mL
Trace Mineral Solution	1 mL
Purine/ Pyrimidine Solution	1 mL
Amino Acid Solution	1 mL
Vitamin K3 solution, 1 mg mL^{-1}	100 µL
FeSO4, 0.4 mg mL^{-1}	100 µL
0.8% $CaCl_2$	100 µL
0.1M $MgCl_2$	100 µL
Haematin-Histidine	100 µL
Vitamin B12 0.01 mg mL^{-1}	50 µL
L- cysteine	100 mg

2.3.4.4 *Prevotellaceae* spp. and *Ruminococcus champanellensis* Media and Growth Conditions

Prevotella and *Ruminococcus* spp. were routinely grown in Brain Heart Infusion (BHI) media supplemented with 0.1% (v/v) Hematin-Histidine. These species were grown under anaerobic conditions at 37 ºC using an A35 anaerobic cabinet (Don Whitley Scientific, UK). As these species are strictly anaerobic all inoculations were performed in the anaerobic cabinet. *R. champanellensis* was grown using the Anaerocult A Incubation System (Sigma Aldrich, UK) inside the anaerobic cabinet. In order to

assess growth of these strains on a sole carbon source (xylan, avicel and corn-derived substrate), we utilised modified Yeast Casitone Fatty Acid (YCFA) media supplemented with 10 mg mL^{-1} growth substrate and 5% (v/v) rumen fluid. Preparation of YCFA media is described in **Table 2-8**.

Table 2-8: Preparation of Modified Yeast Casitone Fatty Acid (YCFA) Media. For constituents of Balch's Vitamins see Supplementary Table X.

	Component	**Amount per Litre**
	Casitone	10 g
	Yeast Extract	2.5 g
	$MgSO_4$ x 7 H2O	45 mg
	$CaCl_2$ x 2 H2O	90 g
	K_2HPO_4	450 mg
	KH_2PO_4	450 mg
	NaCl	900 mg
	Resazurin	1 mg
	Acetic acid	1 mL
	Propionic acid	1.9 mL
	iso-Butyric acid	90 µL
	n-Valeric acid	100 µL
	iso-Valeric acid	100 µL
Boil and allow to cool		
	NaHCO3	4 g
	L-Cysteine	1 g
Autoclave and cool to 55 ºc		
	Balch's Vitamins	10 mL

2.3.4.5 Rumen Fluid Preparation

Rumen fluid utilised in growth studies was kindly provided by Karen Scott, University of Aberdeen. Following collection, rumen fluid was filtered into suitable plastic containers, autoclaved and subsequently centrifuged in aliquots at 10,000 rpm at 4 ºC for 30 minutes. Aliquots were then stored at -20 ºC for long term storage.

2.3.5 Bacterial Culture Preparation and Monitoring

Bacterial cultures were routinely prepared by inoculation of 5 mL of rich TYG or BHI medium with 20 µL of a glycerol stock of desired species and grown to stationary phase for at least 14 hours in controlled anaerobic conditions inside the A35 anaerobic chamber (Don Whitley Scientific) at 37 °C to produce active bacterial cultures. These stationary phase cultures were then used to inoculate *Bacteroides* Minimal Media or YCFA. 10 µL of culture was inoculated per 200 µL well in the plate reader, Cultures were prepared in Corning Costar 96 well plates (Sigma Aldrich, UK) and growth was monitored using an Epoch microplate spectrometer (Biotek Instruments Ltd.) within the anaerobic chamber. The OD_{600} was collected at 15-minute intervals. Each growth curve was completed in triplicate and data averaged. Control wells with no bacterial inoculum were run to ensure that no contamination had occurred. Background absorbance of the control was taken as a baseline and subtracted from all samples. For *Ruminococcus champanellensis* and *B. ovatus* growths on insoluble growth substrates, growth was enumerated via counting colony forming units. Three 1:100 serial dilutions were performed and plated out onto BHI agar plates in triplicate. Data were analysed in Prism 10 (GraphPad). Cultures larger than 200 µL were performed in 10 mL sterile glass tubes which were plugged with cotton wool, preventing contamination. Tubes were placed in the anaerobic chamber and OD600 was monitored using a manual CO 75000 spectrophotometer (Sigma Aldrich, UK).

2.3.6 Storage of DNA and Bacteria

Stocks of all bacterial strains were stored in Cryovials with 25% v/v glycerol at -80 °C. Bacterial colonies on agar plates were stored at 4 °C for a maximum of four weeks. DNA was stored at -20 °C in Elution Buffer (EB, 10 mM Tris/HCl buffer, pH 8.5).

2.4 Molecular Biology

2.4.1 Transformation of Chemically Competent *E. coli*

E. coli strains listed in **Table 2-1** were made chemically competent by Carl Morland, following the protocol outlined by Cohen *et al.* (1972). Competent cells were stored in aliquots of 100 µL with 25% v/v glycerol at -80 °C.

An aliquot was thawed on ice for five minutes, prior to mixing with 2 µL of plasmid DNA then incubated on ice for 30 minutes. Cells were then heat shocked at 42 °C for 2 minutes and immediately placed back on ice. For plasmid propagation, cells were plated on LB agar containing the appropriate antibiotic, and plates were incubated inverted at 37 °C for 16 hours.

During transformation of ligation mixtures, or for protein expression, a recovery step was required. Heat shocked cells were mixed with 250 µL of sterile LB and incubated at 37 °C with agitation (180 rpm) for 1 hour. This culture was then plated on LB agar containing the appropriate antibiotic, and plates were incubated inverted at 37 °C for 16 hours.

2.4.2 Propagation of Plasmids

Following the transformation of DNA into Top10 cells, single colonies were picked and inoculated into 5 mL of LB with appropriate antibiotic. Cultures were incubated for 16

hours at 37 °C for 16 hours with agitation at 180 rpm. Cells were harvested by centrifugation at 5000 rpm for 10 minutes, and the supernatant carefully removed.

2.4.3 DNA Isolation

Plasmid DNA was purified with QIAspin Prep Kit (QIAGEN, UK) as per manufacturer's instructions.

2.4.4 Quantification of DNA

DNA concentration was determined with NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific) by measuring the absorbance at 260 nm, using EB buffer (QIAGEN) or Elution buffer (Sigma) as blanks. 2 µL aliquots were used to determine DNA concentration.

2.4.5 Polymerase Chain Reaction (PCR)

2.4.5.1 PCR for Target Gene Amplification

The polymerase chain reaction (PCR) developed by Mullis & Faloona (1987), was used to amplify target genes throughout this study. This technique requires a thermostable DNA polymerase, a DNA template, 2 oligonucleotide primers – one complimentary to each strand at the sites which flank the region of DNA to be amplified, and dNTPs. Dimethyl Sulfoxide (DMSO) (Sigma Aldrich, UK), as added to improve primer binding to the DNA template. Reactions were performed using a PHC-3 thermocycler (BioRad, USA). Primers were designed at approximately 20 bases in length, with a G/C content of around 40%, a melting temperature of greater than 45 °C and less than 5 °C difference between the 2 primers. Primer parameters were calculated using the online "OligoCalc" tool at http://biotools.nubic.northwestern.edu/OligoCalc. If possible, primers were also designed such that they contain two or three G or C bases at both ends of the primer,

which would make the primer termini anneal well to the template strand and increase the amplification efficiency. When required, restriction site sequences were added to 5´-ends of primers, preceded by a 3 bases CTC overhang to facilitate the restriction cutting of the PCR product for ligation into appropriate vectors. Primers were manufactured by Thermo Fisher Scientific (ThermoFisher, UK) and supplied lyophilised. Primers were dissolved in sterile double distilled water to 100 pmol μL^{1}.

PCR reactions were performed using KOD HotStart DNA Polymerase kit (Merck, UK) in 0.2 mL Eppendorf tubes. Components added to PCR reactions are listed in **Table 2-9** and the standard thermocycler program in **Table 2-10**. Primers are listed in Appendix **Table 8-1.** Negative controls lacking a DNA template were included for all PCR runs.

Table 2-9: Standard PCR reagents for a 50 µL reaction. DNA Polymerase, 10X buffer. dNTPs and $MgCl_2$ were provided in the KOD Hot Start Polymerase kit (Merck, UK).

Component	Volume
KOD Hot Start DNA Polymerase	1 µL
10 × Buffer for KOD DNA Polymerase	5 µL
$MgCl_2$ (25 mM)	3 µL
dNTPs (2 mM each)	5 µL
Oligonucleotide primer forward (5 µM)	5 µL
Oligonucleotide primer reverse (5 µM)	5 µL
DMSO (100 %)	2 µL
DNA template	1 µL (approx. 100 ng)
PCR Grade Water	23 µL

Table 2-10: Standard PCR Thermocycler Program

Temperature	Time	Number of Cycles	Stage
95 °C	1 min	1	Initial Denaturation
95 °C	30 seconds	34	Denaturation
50-55 °C	30 seconds		Annealing
68 °C	1 min/kb fragment size		Elongation
68 °C	10 mins	1	Final Elongation
4 °C	∞		Storage

2.4.5.2 PCR for Site Directed Mutagenesis (SDM)

Mutagenesis of single amino acids was carried out using the site directed mutagenesis (SDM) method. The site-directed mutagenesis method utilizes an appropriate double stranded recombinant plasmid DNA and two synthetic oligonucleotide primers, synthesised by Thermo Fisher Scientific (ThermoFisher, UK) (**Appendix Table 8-1**) containing the desired mutation flanked by 10-15 nucleotides that fully complemented the DNA template. The oligonucleotide primers are extended during PCR temperature cycling by using a thermostable KOD polymerase. Standard SDM PCR reagents and thermocycler programs are shown in **Table 2-11** and **Table 2-12** respectively.

Table 2-11: **Standard PCR reagents for a site directed mutagenesis (SDM).** DNA Polymerase, 10X buffer. dNTPs and $MgCl_2$ were provided in the KOD Hot Start Polymerase kit (Merck, UK).

Component	Volume
KOD Hot Start DNA Polymerase	1 µL
10 × Buffer for KOD DNA Polymerase	5 µL
$MgCl_2$ (25 mM)	3 µL
dNTPs (2 mM each)	5 µL
Oligonucleotide primer forward (5 µM)	5 µL
Oligonucleotide primer reverse (5 µM)	5 µL
DMSO (100 %)	2 µL
DNA template	1 µL (approx. 100 ng)
PCR Grade Water	23 µL

Table 2-12: Standard PCR Thermocycler program for Site Directed Mutagenesis (SDM)

Temperature	Time	Number of Cycles	Stage
95 °C	1 min	1	Initial Denaturation
95 °C	50 seconds	20	Denaturation
50-60 °C	50 seconds		Annealing
68 °C	1 min/kb fragment size		Elongation
68 °C	5 mins	1	Final Elongation
4 °C	∞		Storage

After thermocycling, reactions were submitted to a *DpnI* digest to remove methylated *E. coli* DNA. 3 µL *DpnI,* 6 µL *DpnI* buffer and 1 µL sterile dH_2O were added to 50 µL PCR product. and reactions were incubated at 37 °C for 2 hours, before transformation into Top10 competent cells.

2.4.6 Agarose Gel Electrophoresis of DNA

Separation and determination of the sizes of linear DNA fragments were carried out by electrophoresis through submerged horizontal gels (Brody & Kern, 2004). Agarose gels were prepared at 0.8% w/v by dissolving 0.4 g of agarose in 50 mL of 1X TBE buffer (89 mM Tris-borate and 2 mM EDTA, pH 8.3). The suspension was boiled to dissolve agarose, cooled, and 0.5 µg mL^{-1} of ethidium bromide was added to allow for visualisation of DNA under UV light. Molten gel was poured into mini-gel trays (Applied Biosystems) and a comb added to form wells. Once set, the gel was submerged in 1X TBE buffer and samples were mixed 1:10 with 10X loading buffer (0.25 % bromophenol blue, 50 % v/v glycerol, 10X TBE) and added to wells. 5 µl of DNA HyperladderTM I (Bioline) was run alongside to determine DNA sizes. Gels were run at 30 to 70 volts for 1 hour using a Bromma 2197 Power Pack (Biorad, USA).

The size of linear DNA fragments was determined by comparison to the standards of a known length in the DNA ladder based on their electrophoretic motility. Migration rate through the agarose gel is inversely proportional to the Log10 of the size of the oligonucleotide fragments.

2.4.7 Visualisation of DNA

Images of gels were captured using a gel documentation system (Bio-Rad Gel Doc 1000, Molecular Analyst™/PC Windows Software), and photographs were produced by Mitsubishi Video Copy Processor (Model P68B) with Mitsubishi thermal paper.

2.4.8 Purification of DNA Fragments

2.4.8.1 PCR Purification

PCR products and restriction- digested inserts were purified using QIAquick PCR Purification Kit (Qiagen, UK) as described in manufacturer's instructions.

2.4.8.2 DNA Gel Extraction

PCR products and restriction digested vectors were separated by agarose gel electrophoresis, using high purity SeaKem Gold agarose (Lonza Bioscience) at 0.8% w/v and TBE buffer, run at 10 volts for 10 hours. Bands of appropriate size were excised from the gel using a clean scalpel blade and extracted using Qiagen QIAquick Gel Extraction Kit according to manufacturer's instructions.

2.4.9 Molecular Cloning

2.4.9.1 Digestion with Restriction Enzymes

Where appropriate, double-stranded DNA was digested with relevant restriction endonucleases (Thermo Scientific). 1 µg of DNA was incubated with 1 unit of appropriate endonuclease in 1X or 2X of appropriate buffer. 20-50 µL reactions were incubated at 37 °C for 1 hour. Digested fragments were purified with QIAquick PCR purification Kit (QIAGEN) as per manufacturer's instructions and analysed with gel electrophoresis.

2.4.9.2 Ligation of Vector and Insert DNA

Appropriately endonuclease digested insert and vector DNA, excised from an agarose gel, were used for ligations. Digested fragments were ligated with the Rapid DNA Ligase Kit (ThermoFisher, UK). Insert to vector ratio was calculated using NEBioCalculator (https://nebiocalculator.neb.com/#!/ligation) according to the formula

$$Amount\ of\ insert\ (ng) = \left(\frac{Insert\ length\ (bp)}{Vector\ length\ (bp)}\right) \times Amount\ of\ vector\ (ng)$$

20 µl reactions were prepared in with 1 unit of Rapid Ligase and 1X Ligation Buffer, where the molar ratio of insert:vector was routinely taken at 3:1 but could be adjusted depending on the insert size. Ligations were incubated at room temperature for 90

minutes, then 2 µL was transformed into Top10 or CC118 chemically competent *E. coli* (Section 2.4.1).

2.4.9.3 Automated DNA Sequencing

Sequencing of DNA fragments was performed by Eurofins Genomics (Germany) using the TubeSeq Sanger Sequencing service. Each clone was sequenced in the forward and reverse directions, using standard sequencing primers T7 (TAATACGACTCACTATAGGG) and T7term (CTAGTTATTGCTCAGCGGT) for cloning into pET21a and pET28a. For pGEX-6-P clones, sequencing primers pGEX-3 (GGAGCTGCATGTGTCAGAGG) and pGEX-5 (CTGGCAAGCCACGTTTGG) were used. 10 µL of DNA was sent for each sequencing reaction. Sequencing results were aligned and compared to the original gene in Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/).

2.4.10 Overexpression of Recombinant Proteins

Expression plasmids, carrying the DNA sequence corresponding to the protein of interest, were transformed into BL21 or Tuner competent *E. coli* cells (Section 2.4.1). Cells were plated onto LB-agar containing the appropriate antibiotic and incubated at 37 °C for 16 hours. Colonies were then inoculated into 5 mL LB-antibiotic and subsequently inoculated into 1 L LB- antibiotic in a 2 L flask, which was grown at 37 ºC in a shaking incubator until an OD600 of 0.4- 0.6 was reached. Flasks were cooled 16 ºC before 1 mL 0.2 M isopropylthio-β-D-galactoside (IPTG) was added. Flasks were incubated at 16 ºC for 16 hours, shaking at 180 rpm. Addition of IPTG induces protein expression controlled by lacO in laclq carrying plasmids.

Cells were harvested by centrifugation at 5000 rpm for 10 minutes. Pellets from 1 L of culture were resuspended in 10 mL of TALON buffer (20 mM Tris-HCl, 300 mM NaCl,

pH8.0). Resuspended pellets could be frozen at -20 ºC until use. The cell suspension was kept on ice and sonicated for 2 min using a B. Braun Labsonic U sonicator set at low intensity ~42 watts and 0.5 second cycling. Sonicated cell suspension was transferred to a 50 mL centrifuge tube (Nalgene) and cell debris pelleted at 15000 rpm for 30 min at 4 ºC. Supernatant comprising the cell-free extract (CFE), was retained for protein purification by Immobilised Metal Affinity Chromatography (IMAC). Cell debris was resuspended in TALON buffer and analysed by Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) alongside other IMAC fractions.

2.4.11 Overexpression of Selenomethionine (SeMet) Proteins

A single colony of B834 (DE3) containing the expression plasmid was inoculated into a 10 mL overnight culture in LB with relevant antibiotics and grown at 37 °C 180rpm for around 16 hours. Overnight cultures were centrifuged at 2000 *g* for 10 minutes, supernatant removed, and pellet washed in 10 mL SeMet medium (SelenoMethionine Medium Complete (Molecular Dimensions). This was centrifuged again to remove residual LB, and the pellet resuspended in 1 L SeMet medium supplemented with 40 mg/L L-selenomethionine (Anatrace). SeMet medium cultures were grown at 37 °C until OD600nm = 0.6-1.0 before induction of protein expression by addition of 0.5 mM IPTG. Cultures were induced for 16 hours at 16 °C prior to harvesting cells as described in section 2.4.10.

2.4.12 Purification of Proteins

2.4.12.1 Buffers

Proteins were purified in TALON Buffer, as described in section 2.4.10. For further analysis, proteins were kept in TALON buffer or exchanged into 50 mM HEPES pH 7.5.

2.4.12.2 Immobilised Metal Affinity Chromatography (IMAC)

pET21a and pET28a vectors are designed to introduce a hexa-Histidine tag into the N- or C- terminus of the recombinantly expressed protein. His-tagged proteins can be purified by IMAC as the side chains of the histidine interact with the electropositive transition metal immobilised in the column. This interaction can be disrupted by imidazole, allowing His-tagged protein to be eluted from the column using an imidazole gradient. This purification procedure was carried out in TALON buffer.

TALON™ (Clontech Laboratories Inc.) columns containing 2 mL TALON™ resin with cobalt ions were prepared by washing with 10 column lengths of water before equilibrating with 2 column lengths of TALON buffer. Cell free extracts (CFEs) were filtered (0.45 µm) and poured through the column and eluent was collected (flow-through). Columns were washed with 20 mL TALON buffer to remove unbound proteins. The specifically bound protein was then eluted with 3 x 2 mL fractions of 10 mM imidazole in TALON buffer, followed by 5 x 2 mL fractions of 100 mM imidazole. All 8 fractions were collected and analysed by SDS-PAGE.

2.4.12.3 Glutathione-S-Transferase Tagged Protein Purification by Glutathione Affinity Chromatography

The pGEX-6P-1 vector introduces a Glutathione-S-Transferase (GST) tag into the recombinant protein, which can be used to purify the protein by affinity chromatography using Glutathione Sepharose High Performance resin (Cytiva). Addition of reduced glutathione to the column then allows elution of the GST-tagged protein.

1 L of Tuner cells containing pGEX-6P-1 plasmids were resuspended in 10 mL PBS buffer, pH 7.3 (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4,1.8 mM KH_2PO_4) prior to sonification.

Chromatography columns containing 1 mL Glutathione Sepharose High Performance resin (Cytiva) were equilibrated with 5 column lengths of PBS buffer. Cell free extracts (CFEs) were filtered (0.45 µm) and poured through the column and eluent was collected (flow-through). The column was then washed with 3 column lengths of PBS buffer to remove non-specifically bound protein. Bound protein was then eluted by adding 5 x 3 mL fractions of 50 mM TRIS-HCl with 10 mM reduced glutathione, pH 8.0, to the column. All 5 fractions were collected and analysed by SDS-PAGE.

2.4.12.4 Size Exclusion Chromatography (SEC)

Further protein purification was performed by size-exclusion chromatography (SEC), using a HiLoadTM 16/600 SuperdexTM 200 Prep grade column (Cytiva) in conjunction with an ÄKTA pure 25 chromatography system (Cytiva). The column was equilibrated with two column volumes of TALON buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl) prior to use. Protein sample was concentrated down to approximately 1 mL and loaded into a clean, equilibrated 1 mL loop. Proteins were separated on the column using TALON buffer and an isocratic elution program consisting of 1.5 column volumes with a 1 mL/min flow rate. Proteins were collected in fractions using an F9-R fraction collector (Cytiva) and subsequently visualised by SDS-PAGE to identify size and purity. Fractions containing high yields of pure desired protein were pooled for further use.

2.4.13 Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Proteins were visualised by SDS-PAGE as described by Laemmli (1970) (LAEMMLI, 1970) to assess the size, relative purity and expression level of a protein. 12.5% polyacrylamide gels were prepared using the AE-6450 apparatus from ATTO corporation (Genetic Research Corporation), composed of 2 12 cm x 10 cm glass plates sealed by a rubber gasket. Polyacrylamide gels were formed of a stacking gel, sitting on top of a resolving gel. Composition of polyacrylamide gels and SDS-PAGE running buffer are described in **Table 2-13** and **Table 2-14** respectively. Resolving gel was poured into the plates, covered in water and allowed to polymerise. The water was then removed, and the stacking gel poured on top of the resolving gel. Resolving gel was allowed to polymerise with a comb in place for the formation of wells. The comb and rubber gasket were removed once the gel had set. Protein samples were prepared by mixing 10 µL sample with 7 µL loading buffer (**Table 2-14**) and boiling for 3 minutes to ensure denaturation. The plates were placed in a gel tank filled with SDS-PAGE running buffer, and samples were loaded into the wells of the gel. Molecular weight standard SigmaMarker^{TM} wide range ladder (Sigma) was loaded alongside samples and used to assess protein size. Gels were electrophoresed at 35 A per gel for 1 hour.

Table 2-13: Preparation of 12.5% SDS-PAGE gels

Gel Component	Reagent	Volume per Gel
Resolving gel	0.75 M Tris-HCl, 0.2% SDS, pH8.8	3.2 mL
	40 % acrylamide *	1.9 mL
	dH_2O	1.1 mL
	10% (w/v) ammonium persulphate	30 µL
	TEMED	10 µL
Stacking Gel	0.25 M Tris-HCl, 0.2% SDS, pH8.8	1.25 mL
	40 % acrylamide *	0.25 mL
	dH_2O	1 mL
	10% (w/v) ammonium persulphate	20 µL
	TEMED	7 µL

*BDH Electran acrylamide, 3% (w/v) bisacrylamide

Table 2-14: Buffers used in SDS-PAGE

Component	Reagent	Volume or Concentration
Running Buffer (for 1 L)	32 mM Tris/190 mM glycine, pH 8.3	350 mL
	SDS	0.1% (w/v)
Loading Buffer (for 10 mL)	SDS	10% (w/v)
	0.25 M Tris-HCl, pH 8.8	5 mL
	Glycerol	25% (w/v)
	β-mercaptoethanol	2.5 mL
	Bromophenol blue dye	0.1% (v/v)

Following electrophoresis, proteins were visualised by staining with InstantBlueTM stain (Expedeon) for at least 15 minutes, the gels were destained in distilled water for 16 hours. Gels were photographed using a Bio-Rad Gel Doc 1000, Molecular

AnalystTM/PC windows Software). Photographs were printed with a Video Copy Processor P68B (Mitubishi) on thermal paper (Mitubishi).

2.4.14 Protein Quantification

Protein concentration was determined using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher, USA) and the Beer-Lambert's equation:

$$A=\varepsilon lC$$

A = Absorbance 280 nm - Absorbance 320nm
ε = Extinction coefficient (M^{-1} cm^{-1})
C = Concentration (M)
l = Length of light path (cm)

Extinction coefficients were predicted from protein sequence using ProtParam (web.expasy.org/cgi-bin/protparam).

2.4.15 Concentration and Buffer Exchange of Proteins

Protein samples were concentrated through use of Vivaspin® 20 Concentrator tubes (20 mL) with 10kDa cut off filters (Sartorius Stedim Biotech). Centrifugation was performed at 3500g using a MSE Mistral 3000i bench centrifuge with a swing out rotor at 10 ºC (Section 2.3.2).

Where necessary, buffer exchange was performed with dialysis into 4 L of 50 mM HEPES, pH 7.5, using 13.5 kDa cut off dialysis tubing with stirring at 4 °C for 16 hours.

2.5 Biochemistry

2.5.1 Enzyme Assays

All enzyme assays were performed at standard conditions using 1 µM enzyme and 10 mg mL^{-1} substrate in 20 mM Tris-HCl, 300 mM NaCl, pH 8.0, unless stated otherwise. All reactions were performed at 37 ºC. All assays were repeated at least three times.

2.5.2 Thin Layer Chromatography (TLC)

Thin layer chromatography (TLC) is a solvent-based chromatography method which can be applied to the analysis of glycans. Different sugars will migrate through the solvent at different rates due to the differences in ionisation state, solubility and structure. Smaller and more hydrophobic compounds will migrate faster than larger complexes. The mobile phase, TLC running buffer, consists of a solution of 2:1:1 1-butanol/ acetic acid/ water. 50 mL of running buffer was poured into a glass chromatography tank (23 cm x 23 cm x 7.5 cm), tightly sealed and allowed to equilibrate for 1 hour prior to use.

Silica coated foil TLC plates (Silica Gel 60, 20 x 20 cm, Merck) were used as the stationary phase. Plates were cut to size, with a height of 10 cm. A line was drawn 10 mm up the plate, and 2 x 3 µL aliquots of sample were spotted onto the line, 10 mm apart, allowing thorough drying between spots using a hairdryer. Once dried, the TLC plate was placed in the tank and samples allowed to migrate until running reached approximately 10 mm from the top of the plate. The plate was dried with the hairdryer and placed back in the tank for another run. The plate was dried again, then immersed for 10 seconds in orcinol sulphuric acid reagent (sulphuric acid/ ethanol/ water 3:70:20 v/v, 1% orcinol). The plate was then dried again and carefully heated until sugars could be visualised. Standards consisting of known monosaccharides and oligosaccharides

were spotted on the TLC plate. In order to detect the presence of sugar in a sample quickly, 5 µL was spotted on a section of TLC plate and immersed for a few second in orcinol sulphuric acid reagent, dried carefully and heated until sugars were revealed, as previously described.

2.5.3 High Performance Anion Exchange Chromatography (HPAEC)

Sugars were separated by high-performance anion exchange chromatography (HPAEC) and analysed by pulsed Amperometric detection (PAD). This utilised a fully automated ICS-6000 system (ICS-6000 gradient pump, PAD system and autosampler) with a 100 µL loop. Separation used a CARBOPACTM PA-300 anion exchange column equipped with a CARBOPACTM PA-300 guard column (ThermoFisher). Detection enabled by PAD used a gold working electrode and a PdH reference electrode with standard Carbo Quad waveform. All samples were run at a constant flow rate of 0.25 $mLmin^{-1}$ for 110 minutes following injection. Separation of monosaccharides was achieved using a sodium hydroxide/ sodium acetate gradient, as demonstrated in **Figure 2-1.**

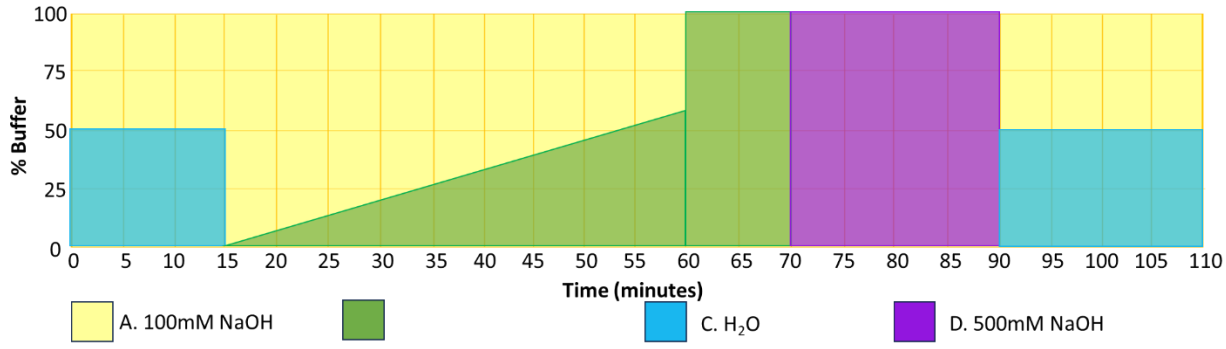


Figure 2-1: Program and buffers used for HPAEC-PAD. Monosaccharide elution program consisted of five phases. 0-15 minutes: isocratic elution for separation of uncharged monosaccharides, 15-60 minutes: linear gradient to 60% buffer B to elute charged monosaccharides, 60-70 minutes: 100% buffer B to remove any remaining sugar from the column, 70-90 minutes: 100% buffer D to remove acetate from the column, 90-110 minutes: 50% buffer A and 50% buffer C to re-equilibrate the column.

L-arabinose, D-xylose, D-galactose, D-glucuronic acid and D-methyl glucuronic acid standards were obtained commercially (Merck). Data were processed using ChromeleonTM Chromatography Management System V.6.8. Graphs created using GraphPad Prism 10.

2.5.4 Bicinchoninic Acid (BCA) Reducing Sugar Assays

The rate of polysaccharide hydrolysis can be monitored by the increase in reducing sugar formation over time. The free anomeric carbon at the end of a polysaccharide can open from its more common cyclic conformation and act as a weak reducing agent. Each time a glycosidic bond is hydrolysed a new reducing end is formed. Bicinchoninic Acid (BCA) assays can be utilised to quantify reducing end production, as reducing ends reduce Cu^{2+} ions to Cu^{+}. 2 molecules of BCA chelate with Cu^{+} forming a product which strongly absorbs light at 560 nm.

Solutions A and B were prepared according to instructions in **Table 2-15*.*** BCA reagent was freshly prepared prior to use by mixing equal parts solution A and B. 125 µL of each reaction was diluted 1:2 with dH_2O and added to 250 µL BCA reagent in a 1.5 mL Eppendorf tube. Samples were vortexed for 20 seconds then incubated at 75 °C for 30 minutes, then allowed to cool to room temperature. Samples were vortexed again, then absorbance measured at 560 nm. All spectrophotometry was performed using a Pharmacia Biotech Ultrospec 83 4000 UV/Visible spectrophotometer and all assays were performed in quartz cuvettes. Data were analysed in GraphPad Prism 10. Glucose standards at 2 to 100 µM were used for validation.

Table 2-15: Reagents used in Bicinchoninic Acid (BCA) Reducing Sugar Assays

Solution	Reagent	Amount (g)
Solution A pH 9.7 (500 mL)	Na_2CO_3	27.14
	$NaHCO_3$	12.1
	BCA sodium salt hydrate	0.971
Solution B pH 3.4 (500 mL)	$CuSO_4.5H_2O$	0.624
	L-serine	0.631

2.5.5 Isothermal Titration Calorimetry (ITC)

Binding of proteins to glycan ligands and the resulting thermodynamic parameters can be analysed by Isothermal Titration Calorimetry (ITC). Formation of protein-ligand complexes either generates or absorbs heat, resulting in exothermic or endothermic reactions, respectively. Measurement of the amount of energy required to maintain the temperature of the reaction cell, relative to the reference cell, provides data on the amount of heat released or absorbed, and hence thermodynamics of binding.

ITC experiments were performed using Malvern Panalytical MicroCal PEAQ-ITC, at 25 °C under standard conditions. Titrations were carried out in 50 mM HEPES buffer at pH 7.5. Proteins were dialysed extensively against buffer and ligands were dissolved in dialysis buffer to minimise heats of dilution. Between 80 and 200 μM protein was equilibrated in the reaction cell, whilst polysaccharide in the syringe was at 20 mg mL^{-1}. Ligand was automatically injected into the reaction cell 19 times with a stirring speed of 750 rpm. During the titration, the difference in electrical power required to maintain the temperature of the reaction cell versus the temperature of the reference cell was recorded, and from these differences the heat change on binding calculated.

As the molar concentration of polysaccharide ligands was unknown, the ligand is in cell model was used to fit data, using an estimated molar concentration of 13 µM, based on 10 available sugars on the chain at the binding site considering spatial restraints for binding. Integrated heats were fit to a single site model using MicroCal PEAQ-ITC Analysis Software v1.41.

2.5.6 Affinity Gel Electrophoresis

Binding of proteins was also analysed qualitatively by affinity gel electrophoresis, where the formation of a ligand-protein complexes distorts the migration of the protein through a native polyacrylamide gel. AE-6450 apparatus from ATTO Corporation was utilised, as for SDS-PAGE (Section 2.4.12). Native polyacrylamide gels containing soluble ligand were produced according to instructions in **Table 2-16**. Ligands were added to the gel prior to polymerisation at a concentration of 0.1% (w/v). Water was used as a negative control. 5 µg of protein was mixed with 5 µl of loading buffer (***Table 2-16***), loaded in the gel wells and run at 10 mA per gel at room temperature. Gels with and without ligands were run in the same gel box with identical samples loaded on each. BSA (15 µg) was used as a negative, non-interacting control. Proteins were stained and visualised as described for SDS-PAGE (Section 2.4.12).

Table 2-16: Composition of NATIVE-PAGE affinity gels and buffers

Component	Composition	Amount
10X Native Running Buffer	Tris base (0.25 M)	30 g L^{-1}
	Glycine (2.5 M)	187 g L^{-1}
DNA loading buffer	Bromophenol Blue	0.0025%
	Glycerol	5%
Native Gel (per gel)	dH20	6.06 mL
	10X Native Running Buffer	1 mL
	40 % Acrylamide *	1.9 mL
	10 % (w/v) Ammonium persulfate	50 µL
	TEMED	10 µL
	Substrate 1% (w/v) or water	1 mL

*BDH Electran acrylamide, 3% (w/v) bisacrylamide

2.5.7 Preparation of Substrates and Oligosaccharides

Unless otherwise specified, all glycan substrates were prepared at 10-20 mg mL^{-1} in 20 mM Tris, 300 mM NaCl buffer at pH8.0. Substrates were stored at -20 °C until required.

Oligosaccharides derived from corn xylan (CX) were created by a partial digest with Bo98 such that the resulting mixture still contained substrate for the same enzyme. Sizes of resultant oligosaccharides were analysed by TLC before purification by size-exclusion chromatography, using two Glass Econo-columnsTM (2.5 cm x 80 cm) packed with fine Bio-Gel P2 media resin with a flow adaptor (Bio-Rad) and equilibrated in vacuum-filtered/degassed 50 mM acetic acid. Samples were loaded directly onto the column which ran 50 mM acetic acid as the mobile phase at 0.2 ml/min using a peristaltic pump (LKB Bromma 2132 MicroperpexTM). 2 ml fractions were collected continuously from 12 hours after loading for 24 hours using a Bio-Rad model 2110

fraction collector. Collected fractions were then analysed by TLC for presence and purity of oligosaccharides, before being pooled, freeze-dried and stored at -20 °C.

2.6 Structural Determination by X-Ray Crystallography

2.6.1 Protein Sample Preparation

Following protein purification by IMAC and SEC proteins were concentrated to 15 mg mL^{-1} in buffer consisting of 20 mM TRIS pH 8.0, 150 mM NaCl.

2.6.2 Protein Crystallisation

Protein crystallisation was performed by a vapour diffusion sitting-drop method using an automated MosquitoR nanodrop dispensing system (TTP Labtech). Crystal trays were set up using MRC 96 well crystallisation plates (Molecular Dimensions) using 100:100 and 200:100 nanolitre ratios of protein sample to crystallisation condition. Various commercially available screening kits were used in initial screens: Index (Hampton Research) and PACT, Structure, Index, JCSG+ (Molecular Dimensions). Crystallisation trays were created at Newcastle Structural Biology Laboratory and left equilibrating at 20 °C for up to 6 months. Plates were examined weekly for crystals using a Leica MZ-6 crystallisation microscope (Leica Microsystems). Following initial screening, optimisation trays were set up in MRC 96 well crystallisation plates (Molecular Dimensions) using the dragonfly crystal system (spt Labtech) based around conditions G10 and H4 from the PACT screen. Selenomethionine labelled Bo98 E361A crystals formed from 0.02M sodium potassium phosphate, 0.1 M Bis-Tris propane, 20% w/v PEG3350 at pH 8.5. Crystals were harvested with 20% PEG400 as a cryoprotectant, before cryocooling in liquid nitrogen prior to data collection. Native Bo98, E361A was crystallised in conditions consisting of 0.2 M potassium thiocyanate, 0.1 M Bis-Tris propane and 20% w/v PEG3350 at pH 8.5. Crystals were then soaked

with 350 mg mL^{-1} CX-derived xylooligosaccharides for 16 hours, then handled as described above.

2.6.3 Structural Determination

All data collection was carried out by Arnaud Baslé (Biosciences Institute, Newcastle University) at beamline i03 at Diamond Light Source. All data processing was done under the supervision of Arnaud Baslé (Biosciences Institute, Newcastle University) and Alan Cartmell (Department of Biology, University of York). Structure refinement was performed with assistance from Alan Cartmell and Charlie Tomlinson (University of York). The CCP4 suite was used for data processing (Winn *et al.*, 2011; Agirre *et al.*, 2023). Data was scaled and merged using AIMLESS (Evans & Murshudov, 2013; Evans, 2014). The phase problem was solved for the SeMet dataset with Crank2 EP (Pannu *et al.*, 2011) and an initial model was generated with CCP4 BUILD. Iterations of refinement and manual model building were done using Refmac and Coot respectively (Murshudov *et al.*, 2011; Emsley & Cowtan, 2004; Emsley *et al.*, 2010). The model from this SeMet dataset was used to solve the phase problem for native Bo98 E361A using molecular replacement with Phaser MR (McCoy, 2007). Iterative refinement and manual model building using Refmac and Coot were again used to build the model.

All figures were generated with UCSF ChimeraX (Pettersen *et al.*, 2021).

2.7 RNASeq Transcriptomics

2.7.1 RNA isolation and purification

R. champanellensis was grown on YCFA-rumen fluid media for 1 week, supplemented with avicel or corn-derived substrate as a sole carbon source. *R. champanellensis* was also grown on YCFA-rumen fluid supplemented with glucose for 12 hours. Cells were

harvested by centrifugation and treated with RNA protect, according to manufacturer's instructions, before centrifuging at 300 rpm for 15 minutes and removing the supernatant. Cell pellets were then treated with 10 mg mL^{-1} lysozyme and incubated for 5 mins at 25 °C, prior to RNA extraction using the Monarch Total RNA Miniprep Kit (New England Biolabs) according to manufacturer's instructions. Briefly, two volumes of RNA lysis buffer were added to the cells and vortexed vigorously for 10 seconds before centrifugation at 13,000 rpm for 10 minutes to remove cellular debris and insoluble growth substrate. The supernatant was transferred to a gDNA removal column and centrifuged at 13,000 rpm for 30 seconds. The flowthrough, containing RNA, was retained. Equal volume of 95% ethanol was added to the flowthrough and mixture was transferred to an RNA purification column for centrifugation at 13,000 rpm for 30 seconds. Bound RNA was then washed with 500 µL RNA priming buffer, by centrifugation at 13,000 rpm for 30 seconds, followed by two washes with 500 µL RNA wash buffer. Column was transferred to an RNAase free tube and RNA samples were eluted in 30 µL pre-warmed nuclease free water by centrifugation at 13,000 rpm for 30 seconds. Any contaminant DNA was then removed from samples using TURBO DNase (ThermoFisher Scientific). 15 µL of sample was combined with 35 µL nuclease free water, 5 µL 10X TURBO DNase buffer and 1 µL TURBO DNase. Reactions were incubated at 37 °C for 25 minutes, mixing gently every 10 minutes, before addition of 5 µL DNase inactivation reagent and incubating at room temperature for a further 5 minutes. Samples were then centrifuged at 10000 rpm for 90 seconds, and supernatant, containing RNA, removed and stored at -80 °C prior to sequencing. The presence of intact RNA in the samples was then checked for using gel electrophoresis (100 V, 30 minutes), with a 0.8% agarose gel. Quantity and purity of RNA was assessed using NanoDrop 2000 UV-Vis spectrophotometer (Thermo Fisher Scientific)

by measuring the absorbance at 260/230 nm, using nuclease free water as a blank. 1 µL aliquots were used to determine RNA concentration.

2.7.2 RNA quality control

Prior to sequencing, quality of RNA was examined by the Newcastle Genomics Core Facility, using the TapeStation RNA ScreenTape to assess RNA concentration and integrity.

3. Xylan Utilisation by Bacteroidota Members of the Human and Animal Gut Microbiota

3.1 Introduction

Across mammalian species the microorganisms which inhabit the gastrointestinal tract play a critical role in the degradation of dietary glycans otherwise inaccessible to the host organism. Complex enzymatic systems are required in order to fully break down these complex plant cell wall materials and complete degradation often involves the synergistic activities of multiple microorganisms, demonstrating the necessity for complex composition of the gut microbiota (La Rosa *et al.*, 2022).

The composition of the gut microbiota on the phylum level plays a huge impact on overall function. Members of the phylum Bacteroidota, in particular those belonging to the genera *Bacteroides* and *Prevotella*, have broad capacities for degradation of wide-ranging plant and host-derived glycans which, in some cases, have been extensively characterised (Martens, Koropatkin, *et al.*, 2009). Bacillota such as *Ruminococci*, *Roseburia* and *Eubacteria* spp. also have glycan degrading ability, and are generally considered more specialised with respect to polysaccharide degradation (O. Sheridan *et al.*, 2016; Crost *et al.*, 2013; La Rosa, Leth, *et al.*, 2019), however in general these have not been extensively studied.

Xylans are the most abundant hemicellulose occurring in nature and are present in many components of the human diet including cereals, herbs, fruits and vegetables, and even algae (Rogowski *et al.*, 2015). Total digestibility of xylan in the human gastrointestinal tract is relatively high, at around 72% (Slavin, 1981), indicating effective mechanisms of degradation by the gut microbiota. Due to the relatively high accessibility of xylans to members of the HGM, there utilisation is an important

consideration in nutritional sciences. Xylans, and their xylo-oligosaccharide partial degradation products, have been implemented in maintaining a beneficial microbiome composition, gut homeostasis and host metabolic profile (Zhang *et al.*, 2022b; Chen *et al.*, 2020; La Rosa, Kachrimanidou, *et al.*, 2019).

Multiple *Bacteroides* species have demonstrated xylan degrading ability, including *B. eggerthii, B. cellulosilyticus, B. intestinalis, B. ovatus, and B. xylanisolvens* (Dodd *et al.*, 2011). These species encode xylan-degradation apparatus within co-upregulated polysaccharide utilisation loci (PULs), which contain, at a minimum, genes encoding a SusD-like binding protein, a SusC-like TonB dependent transporter and at least one CAZyme encoding gene (Martens, Koropatkin, *et al.*, 2009).

Rogowski *et al.,* demonstrated the highly dynamic xylan degrading capabilities of *B. ovatus*, a prominent member of the HGM (Rogowski *et al.*, 2015). *B. ovatus* strain ATCC 8483 contains two PULs upregulated during growth on Wheat Arabinoxylan (WAX), termed the large xylan PUL (PUL-XylL) and the small xylan PUL (PUL-XylS). Whilst growth on simple linear xylo-oligosaccharides led to activation of just PUL-XylS, during growth on complex corn glucuronoarabinoxylan (CX) PUL-XylL was highly upregulated and PUL-XylS relatively little. This indicates that PUL-XylS is important for the breakdown of simple xylans, whereas PUL-XylL functions in the degradation of more complex, highly decorated xylans. These differences in function can be explained by differences in the CAZymes constituting each PUL. Both PULs encode CAZymes known to function in the breakdown of simple xylans, including GH10 endo-xylanases, GH43 β-xylosidases/ α-arabinofuranosidases and GH67 α-glucuronidases. PUL-XylL contains additional xylan-degrading CAZymes: GH3 β-xylosidases, GH30 glucuronoxylanases and GH115 α-glucuronidases, as well as

CAZymes belonging to families not previously associated with xylan degradation (GH31, GH95, GH97 and GH98). This GH98 was demonstrated to possess specific endoxylanase activity against complex GAX (Rogowski *et al.*, 2015). An overview of *B. ovatus* xylan PUL encoded enzyme activity against different xylan types is provided in **Figure 1-19*.***

Depres *et al.* demonstrated that another prominent HGM member, *B. xylanisolvens* XB1A, is specialised in the breakdown of divergent xylan structures (Despres *et al.*, 2016). Upregulation of one PUL, and a remnant of another PUL was observed during growth on WAX and Oat Spell Xylan (OSX), another AX with relatively low levels of branching. This large *B. xylanisolvens* PUL possesses high homology to the *B. ovatus* XylPUL-L (Rogowski *et al.*, 2015). As discussed previously, the *B. xylanisolvens* PUL encodes a GH5 subfamily 21 in the position of the *B. ovatus* GH98. This GH5_21 enzyme also demonstrates endoxylanase activity against GAX, but also simpler AX (Rogowski *et al.*, 2015). Centanni *et al.* showed that another *B. xylanisolvens* strain, DSM 18836, and *B. ovatus* ATCC 8483 were capable of growth on a novel, complex glucuronoarabinoxylan from New Zealand Flax (NZ Flax), as well as WAX and beechwood glucuronoxylan (BEX). Interestingly, this study found that *B. cellulolyticus* DSM 14838 was incapable of utilising BEX or NZ flax xylan for growth, indicating a strong preference for simple arabinoxylans (Centanni *et al.*, 2017). *B. intestinalis* also possesses xylanolytic properties and encodes multiple xylan-active enzymes with the combined capacity for complete degradation of arabinoxylans found in the human diet (Wang *et al.*, 2016). Much like *B. ovatus, B. intestinalis* encodes a large repertoire of GHs (Kaoutari *et al.*, 2013). When *B. intestinalis* was grown on WAX, genes in two large PULs were upregulated compared to growth on xylose (Zhang *et al.*, 2014). Interestingly, when examining *Bacteroides* xylan targeting PULs,

they all seem to lack GH11 endo-xylanase enzymes, despite their role in xylan degradation by other members of the HGM including *Ruminococcus* spp. (Moon *et al.*, 2011).

Another genus of Bacteroidota prevalent in human and animal gut microbiota is *Prevotella*, however this has now been split into seven genera: *Prevotella, Segatella, Hallella, Xylanibacter, Hoylesella, Leyella, Palleniella* (Hitch *et al.*, 2022), These genera, and the closely related *Paraprevotella* and *Pseudoprevotella*, belong to the *Prevotellaceae* family. Around 50 *Prevotellacaea* species have been isolated from the oral cavity, gut or rumen, and these species are divergent in CAZyme repertoires and glycan degradation abilities (Accetto & Avguštin, 2015; Aakko *et al.*, 2020). Several *Prevotellaceae* species have been demonstrated to have strong xylan degradation abilities (Dodd *et al.*, 2011), however in comparison to *Bacteroides* these are not so well characterised.

Prevotellaceae species are not ubiquitous in the HGM, however a diet-related dominance of *Prevotellaceae* is seen in some individuals. Dominance of either *Prevotella* or *Bacteroides* in the human gut leads to two distinct enterotypes with differing fibre utilisation capacities (Chen *et al.*, 2017). A *Bacteroides*-dominated microbiota is generally associated with western diets high in protein and fats, whereas a diet rich in plant-based fibre is reported to lead to dominance of *Prevotella* species (Wu *et al.*, 2011). Next generation sequencing (NGS) studies have shown ubiquity and high prevalence of *Prevotella* species in mammalian rumens (Accetto & Avguštin, 2019), perhaps unsurprisingly due to the high plant fibre content in the diet of herbivorous mammals. Among both wild and domestic ruminants, the proportions of *Prevotellaceae* species in the rumen are relatively high at 15 to 30% (Accetto &

Avguštin, 2015). This is undoubtedly indicative of a substantial role in the fibre degrading capacities of rumen microbiota.

A lot of research on *Prevotellaceae* xylan degradation to date has been focussed on the ruminal strain *Segatella bryantii* B_14 which contains a xylan utilisation PUL, suggesting that *S. bryantii* should have the capacity for utilisation of relatively simple xylans. Indeed, Matsui *et al.* demonstrated this strain's ability to grow on glucuronoxylan or arabinoxylan (Matsui *et al.*, 2000). In the same study, *Xylanibacter ruminicola* showed preference for arabinoxylans, and unsurprisingly due to lack of xylan-targeting apparatus in its genome, *Xylanibacter brevis* did not effectively utilise any xylan substrate (Accetto & Avguštin, 2019).

Although metagenomic studies have demonstrated extensive divergence in PUL and CAZyme content of *Prevotellaceae* species (Accetto & Avguštin, 2019), differences in the xylan utilisation capacities of individual species, and the enzymatic mechanisms underlying them, have not been completely characterised. Given the high prevalence of *Prevotellaceae* in the rumen and in many human intestinal tracts, and the likelihood that many ruminal species are likely still awaiting isolation (Accetto & Avguštin, 2021; Stewart *et al.*, 2019). This is undoubtedly an important area of study if we wish to completely understand the ability of the human or rumen microbiota to utilise plant cell wall materials.

As the defining characteristic of Bacteroidota PULs, SusC/D- like pairs play an important role in the specific binding and import of glycans. Two of the SusD- like surface binding proteins from *B. ovatus* xylan PULs have been well studied, and their binding profiles characterised (Rogowski *et al.*, 2015). Unsurprisingly, Bacova_04392 from PUL-XylS binds to simple AX and GX such as WAX and BGW, but not to complex

GAX. Bo_03427, from PUL-XylL, binds to CX and BWX, but not to WAX or linear xylooligosaccharides, although these binding preferences cannot be fully explained as yet (Rogowski *et al.*, 2015). Further studies into the specificities of SusC/D-like systems from within xylan PULs are necessary to elucidate the true role of these proteins, and the possibility of utilisome formation, for the specificity of glycan degradation.

Although the xylan utilising capacities of some *Bacteroides* and *Prevotellaceae* spp. have been characterised, there are likely many other species from the HGM who contribute to xylan degradation. Improved understanding of xylan breakdown in the gut is important to elucidate the functional capacity of the HGM as a whole and has many potential important applications in the nutritional sciences.

Here we demonstrate the ability to bioinformatically predict the ability of individual *Bacteroides* and *Prevotellaceae* strains to degrade xylan based on identification of xylan specific PUL within their genomes. Only those strains possessing putative xylan PULs, identified based on the presence of key xylan-active enzymes from GH families 10, 30_8, 5_21 or 98, have the ability to utilise xylans as a sole carbon source for growth. Furthermore, we show that sequences of SusC- and SusD- like proteins from within xylan PULs cluster into distinct groups, with different tertiary structures, which may be of functional relevance within the PUL-encoded xylan utilisation systems.

3.2 Results

3.2.1 Prediction of Polysaccharide Utilisation Loci in Bacteroidota species from the human and animal gut microbiota:

Genomes of 12 *Bacteroides* and 7 *Prevotella*/ *Paraprevotella* (**Table 3-1**) clade strains from the human gut, mouse gut or rumen microbiota were selected based on previous evidence of different glycan degradation capabilities and analysed for xylan-targeting polysaccharide utilisation loci.

Table 3-1: Bacteroidota strains tested for xylan utilisation capabilities in the present study. Strains selected have been demonstrated in previous studies to be capable of degradation of at least one glycan type.

Strain	Origin	Reference
Segatella copri DSM18205	Human	(Fehlner-Peach *et al.*, 2019)
Segatella bryantii B14	Bovine rumen	(Dodd *et al.*, 2010)
Segatella albensis DSM 11379	Bovine rumen	(Matsui *et al.*, 2000)
Xylanibacter muris DSM 103722	Mouse	(Gálvez *et al.*, 2020)
Palleniella intestinalis DSM 103738	Mouse	(Gálvez *et al.*, 2020)
Paraprevotella clara DSM 19731	Human	(Morotomi *et al.*, 2009)
Paraprevotella xylaniphila DSM 19681	Human	(Morotomi *et al.*, 2009)
Bacteroides caccae ATCC 43185	Human	(Wang & LaPointe, 2020)
Bacteroides cellulosilyticus DSM 14838	Human	(Robert *et al.*, 2007)
Bacteroides cellulosilyticus WH2	Human	(McNulty *et al.*, 2013)
Bacteroides clarus DSM 22519	Human	(Ma *et al.*, 2024)
Bacteroides eggerthii DSM 20697	Human	(Dodd *et al.*, 2010)

Bacteroides finegoldii DSM 17565	Human	(Wang *et al.*, 2024)
Bacteroides fluxux DSM 22534	Human	(Déjean *et al.*, 2020)
Bacteroides intestinalis DSM 17393	Human	(Wang *et al.*, 2016a)
Bacteroides ovatus ATCC 8483	Human	(Fultz *et al.*, 2021; Rogowski *et al.*, 2015)
Phoecaeicola plebeius DSM 17135	Human	(Hehemann *et al.*, 2014)
Bacteroides salyersiae DSM 18765	Human	(Wang *et al.*, 2024)
Bacteroides xylanisolvens XBA1	Human	(Despres *et al.*, 2016)

Xylan PULs were confirmed as those containing at least one SusC/D-like pair and at least one GH encoding gene from families 10, 98, 30_8 or 5_21. All 7 of the *Prevotellaceae* strains studied possessed putative xylan PULs (**Figure 3-1**), suggesting that they all have ability to degrade at least relatively simple GX or AX. As members of the GH98 family can possess endoxylanase or galactosidase activity, where PULs were identified based on presence of the GH98 enzyme, neighbourhood regions were examined for other potential xylan active GH families, as discussed previously. Only members of two GH families, GH98 and GH5_21, are known to possess endoxylanase activity against complex glucuronoarabinoxylans (GAX), such as corn xylan (CX). Therefore, where strains possess a PUL-encoded GH98 or GH5_21, we predicted capability to also use CX for growth, although other strains may possess alternative mechanisms for complex GAX degradation, such as extensive debranching of the xylan chain. Interestingly, *Segatella albensis* DSM 11379, *Xylanibacter muris* DSM 103722 and *Palleniella intestinalis* DSM 103738 all lack the

presence of GH5_21 or GH98 encoding genes within their putative PULs. However, genome BLAST searches with characterised GH98 from *B. ovatus* ATCC 8483 demonstrated that *S. albensis* and *X. muris* genomes encode a GH98 enzyme which is not contained within a PUL. Prior to experimental analysis, it cannot be ascertained that these GH98 gene products possess endoxylanase rather than blood group sugar targeting galactosidase activity. Based on the known xylan-degrading capabilities of these strains, we have worked on the initial assumption that these will possess endoxylanase activity. This suggests that only *P. intestinalis* may lack the ability to utilise complex GAX.

Of the 12 *Bacteroides* strains, 7 possessed at least one putative xylan PUL (**Figure 3-2**). *B. cellulolyticus* WH2 possessed 3 putative xylan PULs, *B. cellulolyticus* DSM 14838, *B. eggerthii* DSM 20697, *B. intestinalis* DSM 17393 and *B. ovatus* ATCC 8483 possessed two each, and *Phoecaeicola plebius* DSM 17135 and *B. xylanisolvens* each had one putative xylan PUL. Based on this we predicted that these 7 strains would be capable of utilising xylans for growth. Due to the lack of a PUL-encoded GH98 or GH5_21 enzyme, we predicted that of the 7 strains with putative xylan PULs, only *P. plebieus* may lack ability to degrade GAX. No putative xylan PULs were identified in the following five strains: *B. caccae* ATCC 43185, *B. clarus* DSM 22519, *B. finegoldii* DSM 17565, *B. fluxus* DSM 22534 and *B. salyersiae* DSM 18765. This suggests that these strains will be incapable of utilising xylans as a sole carbon source for growth.

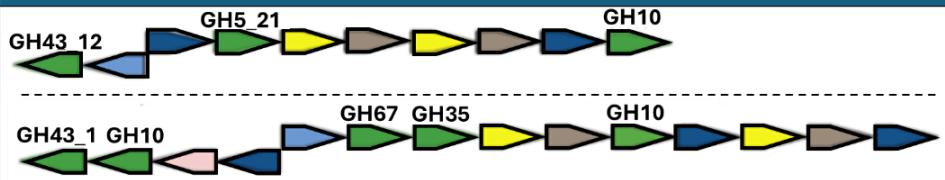
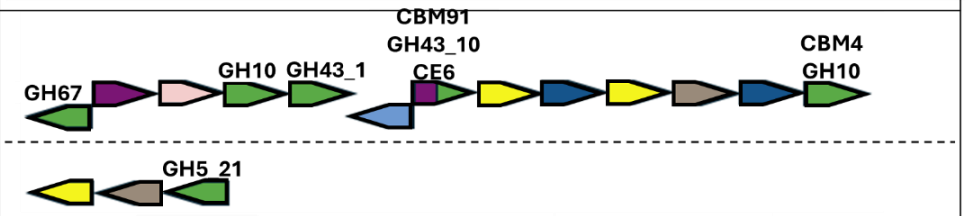
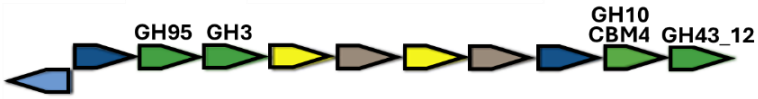
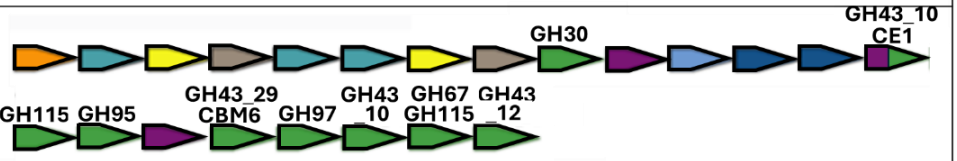
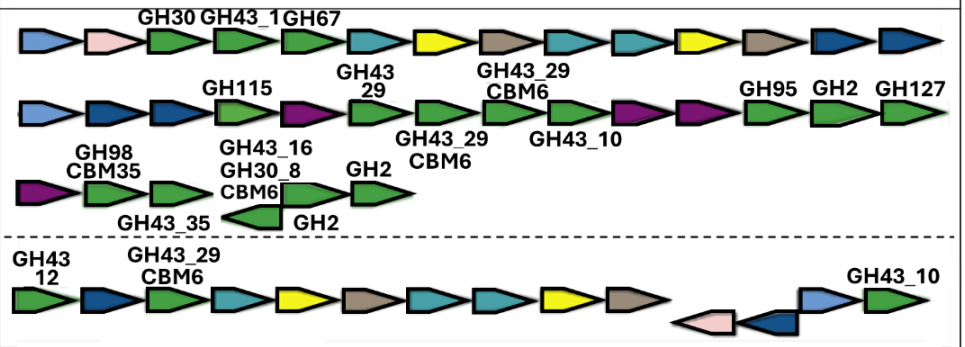
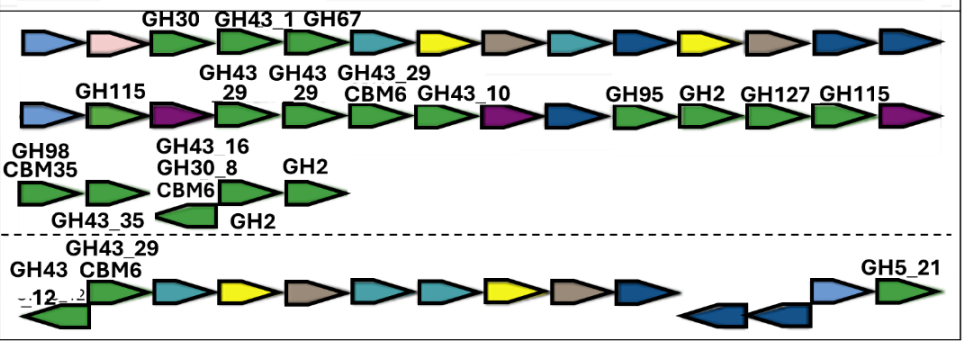
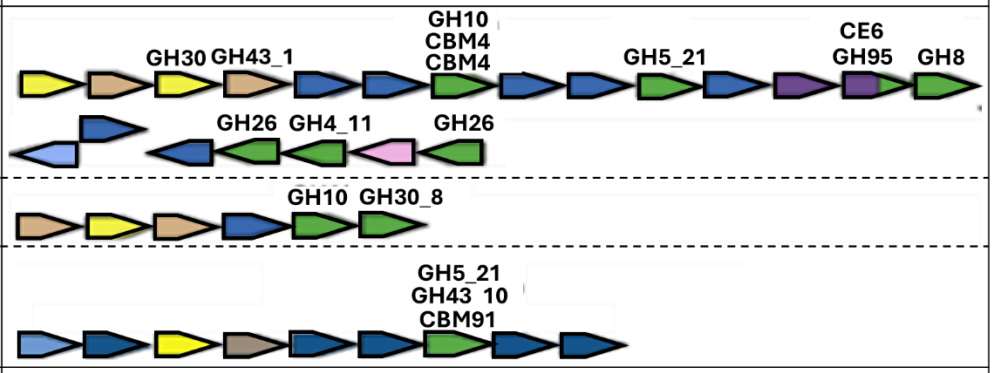
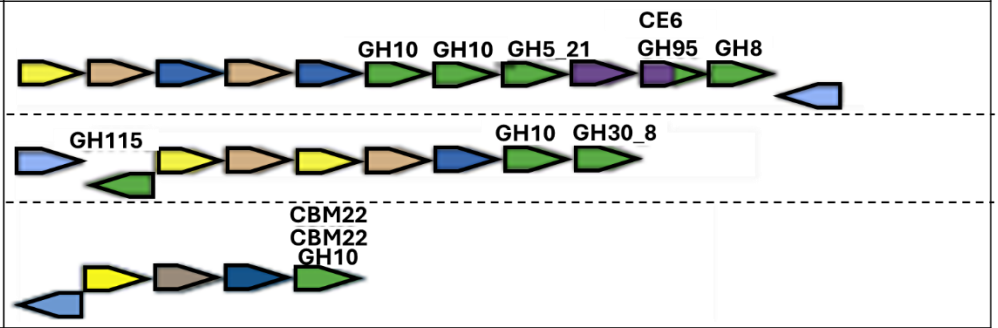
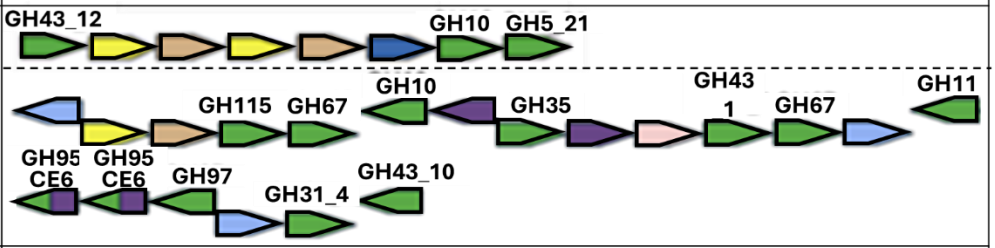
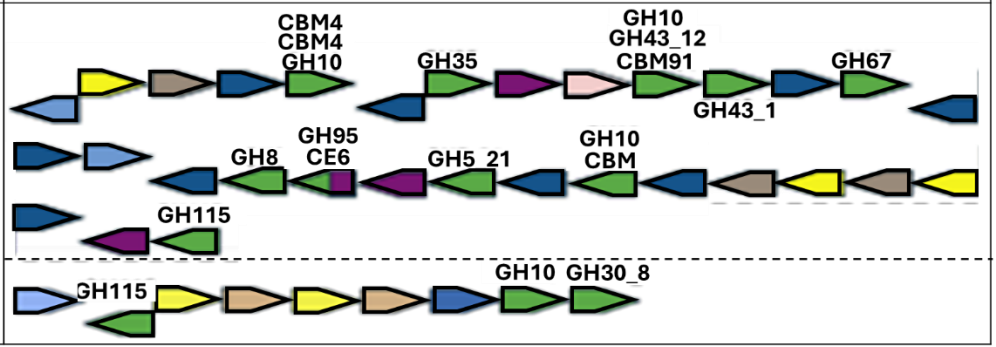
Strain	Number of predicted xylan PULs	PUL schematic
Segatella copri DSM 18205	2	
Segatella bryantii B14	2	
Segatella albensis DSM 11379	1	
Xylanibacter muris DSM 103722	1	
Palleniella intestinalis DSM 103738	1	
Paraprevotella clara DSM 19731	2	
Paraprevotella xylaniphila DSM 19681	2	

Figure 3-1: Putative xylan Polysaccharide Utilisation Loci (PULs) in Prevotellaceae species from the human or animal gut. PULs were predicted using PULDB (Terrapon *et al.*, 2018) and confirmed via exploration of neighbourhood regions of putative xylan-degrading GH enzymes in IMG/M.

Strain	Number of predicted xylan PULs	PUL schematic
Bacteroides caccae ATCC 43185	0	
Bacteroides cellulosilyticus DSM 14838	2	
Bacteroides cellulosilyticus WH2	3	
Bacteroides clarus DSM 22519	0	
Bacteroides eggerthii DSM 20697	2	
Bacteroides finegoldii DSM 17565	0	
Bacteroides fluxus DSM 22534	0	
Bacteroides intestinalis DSM 17393	2	

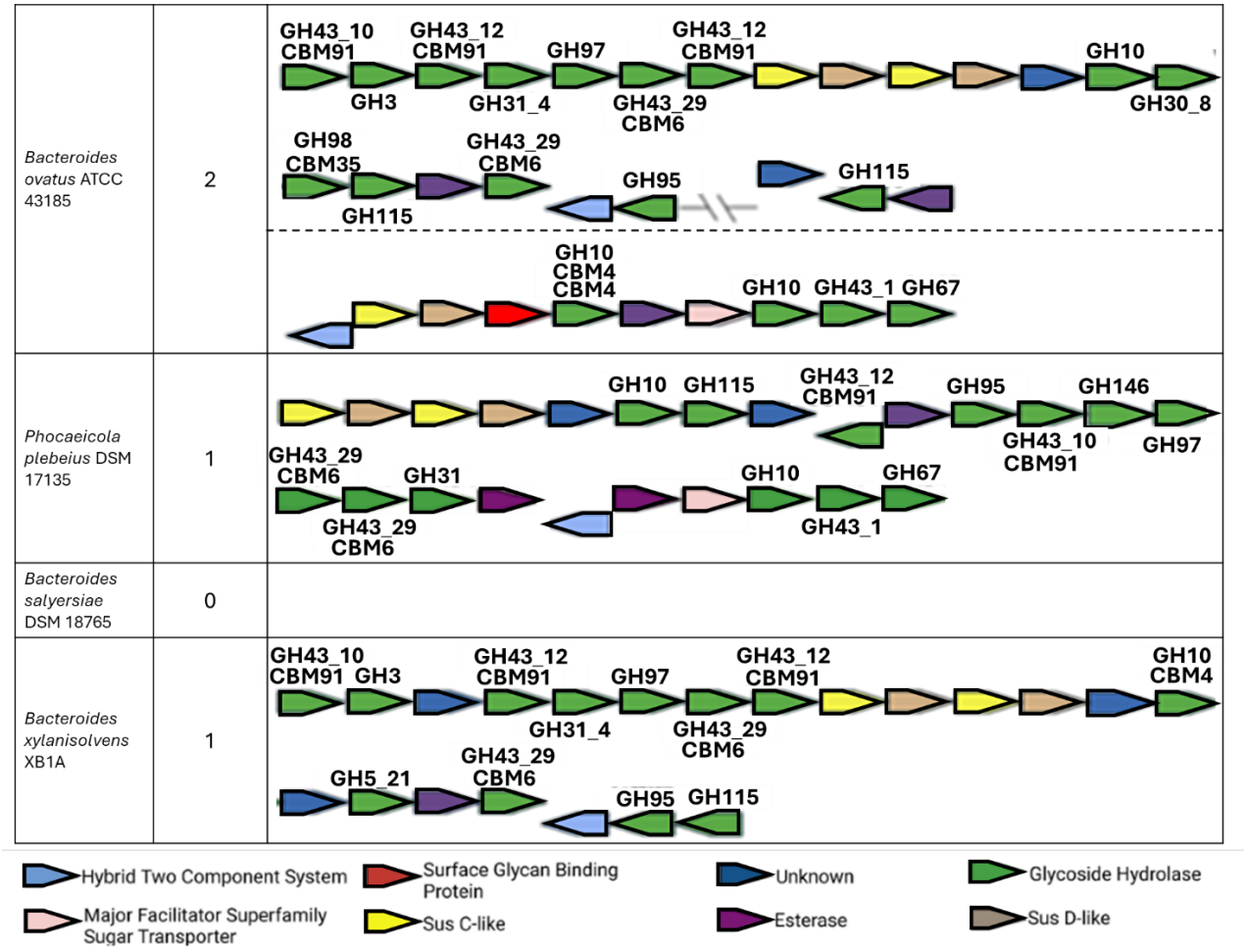


Figure 3-2: Putative xylan PULs in *Bacteroides* species from the human or animal gut. PULs were predicted using PULDB (Terrapon *et al.*, 2018) and confirmed via exploration of neighbourhood regions of putative xylan-degrading GH enzymes in IMG/M.

3.2.2 Growth of human and animal gut *Prevotellaceae* strains on different xylans

Due to the more fastidious nature of members of the *Prevotellaceae* family these strains could not be grown on *Bacteroides* minimal media, or a range of other minimal medias and evidence for xylan utilisation was only seen on YCFA media supplemented with 5% rumen fluid (**Figure 3-5**). Although YCFA media already possesses some sources of carbon which could support growth of *Prevotellaceae* spp., lack of a suitable minimal media for these species meant that we were assessing

any increase in growth in xylan-supplemented YCFA-rumen fluid, compared to media alone.

All 7 strains seemed to possess ability to utilise WAX, with all except *S. bryantii* reaching a max OD of at least 0.5 greater than growth on media alone (**Figure 3-5**). Interestingly, growth of *S. bryantii* seemed to be poor on all substrates except glucose, on which it shows biphasic growth (**Figure 3-5**). This relatively poor growth may be due to adaptation of *S. bryantii* to the ruminal niche as an obligate anaerobe, meaning that slight O_2 presence, as is likely under the growth conditions utilised in this study, may be inhibitory to growth.

In comparison to WAX, growth of *Prevotellaceae* on BWX was much poorer, with *S. bryantii, X. muris*, *P. intestinalis*, *P. clara* and *P. xylaniphila* showing only marginally higher growth on BWX compared to control media, whilst *S. copri* and *S. albensis* seemed incapable of utilising BWX as a sole carbon source for growth (**Figure 3-5**). Interestingly, these 2 strains which did not grow on BWX are the only ones lacking GH67 or GH115 α-glucuronidases within their predicted xylan PULs (**Figure 3-1**), demonstrating that removal of glucuronic acid or methyl-glucuronic acid decorations is necessary to enable utilisation of BWX. Overall, these data demonstrate a preference of *Prevotellaceae* spp. for arabinoxylan over glucuronoxylan.

Despite a preference for AX, growth on OSX, a less highly decorated AX, is poor in some species. Whilst *S. copri, S. albensis, X. muris* and *P. clara* show little to no growth on OSX, *S. bryantii, P. intestinalis* and *P. xylaniphila* show robust growth, reaching a maximum OD of at least 0.4 above media controls (**Figure 3-5**). By observation of CAZyme content of predicted xylan PULs (**Figure 3-1**) it is challenging to predict which factors determine these differences in OSX utilisation, and it may be

due to divergence in substrate specificity of different members of the same GH family between species. Recombinant expression and characterisation of more of these PUL encoded *Prevotellaceae* is necessary to truly understand substrate specificities.

Again, differences in utilisation of complex GAX, CX, were seen between *Prevotellaceae* strains (**Figure 3-5**). All except *S. copri* and *S. bryantii* showed high growth on CX. This is of interest as both these strains possess a PUL encoded GH5 subfamily 21 enzyme (**Figure 3-1**), which we predicted would have endoxylanase activity against complex GAX. Furthermore, we show that *P. intestinalis*, which lacks both GH98 and GH5_21 enzymes within PUL (**Figure 3-1**), is capable of CX utilisation. Further analysis of the *P. intestinalis* genome revealed a GH5_21 encoding gene, which likely allows this GAX utilisation.

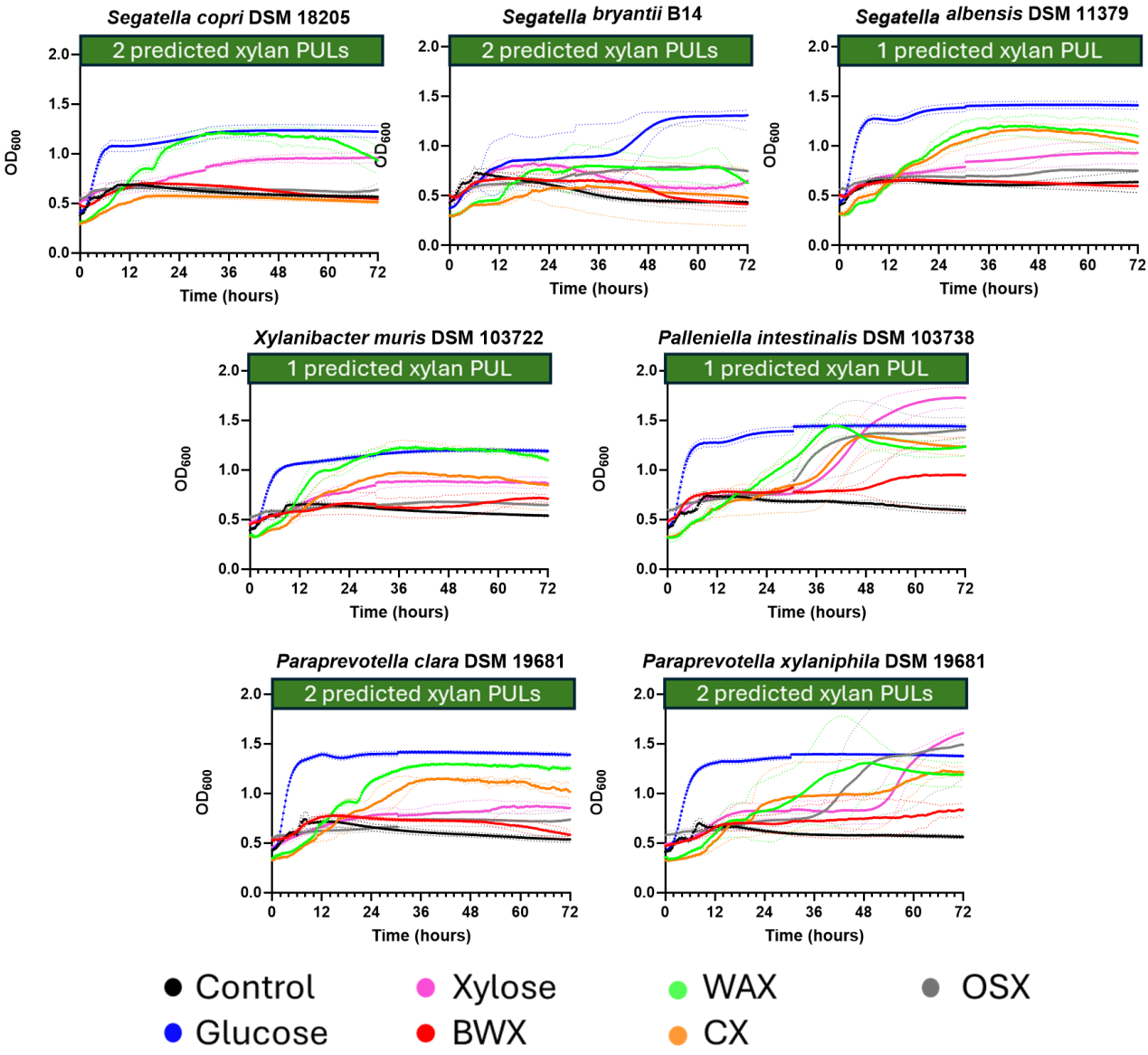


Figure 3-3: Growth of *Prevotellaceae* strains on different xylans. Seven human and animal gut *Prevotellaceae* strains were grown on YCFA media supplemented with 5% rumen fluid and 5 mg mL^{-1} carbon source for 48 hours. Cultures were run in triplicate and optical density at 600 nm was recorded every 10 minutes.

3.2.3 Breakdown products remaining after growth of *Prevotellaceae* spp. on different xylans

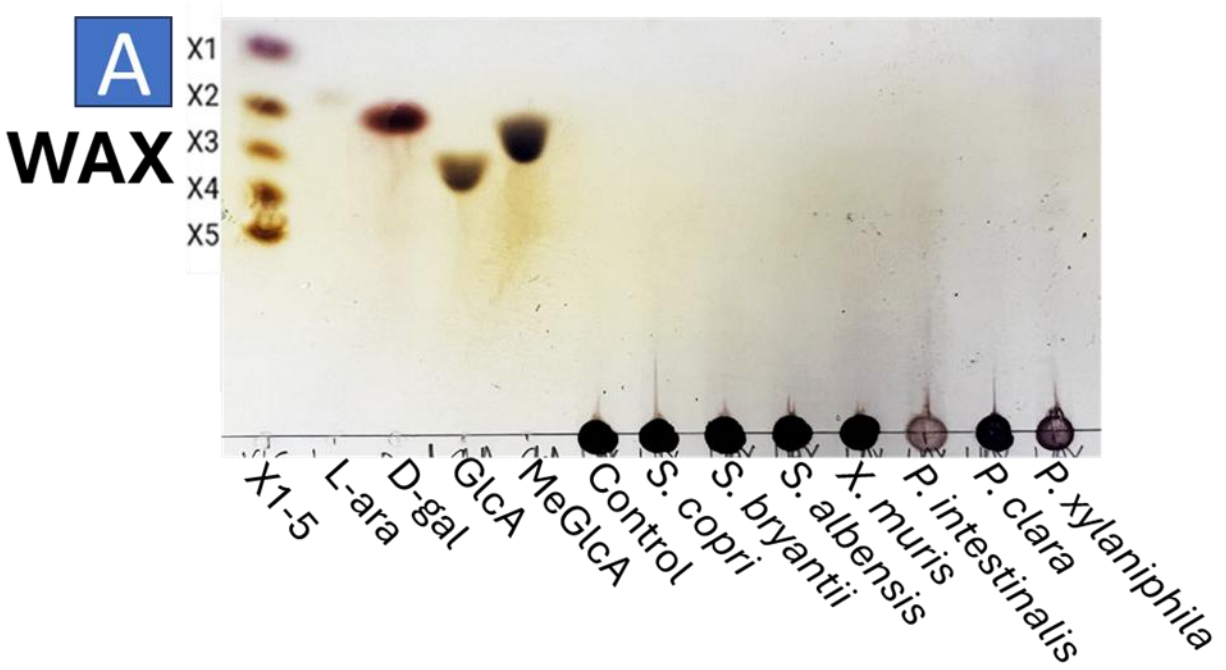
When examining the sugar content of the media after 72 hours of growth via TLC (**Figure 3-6**), during growth on WAX no small soluble oligosaccharides or monosaccharides were seen in spent media of any strains, indicating that any xylan breakdown products were taken up in the cells, as explained by the Sus-like system paradigm. Interestingly, for *P. intestinalis* and *P. xylaniphila* less substrate was seen on the origin, in comparison to the control media, as shown by a paler "spot" (**Figure 3-6A**). Material on the origin may be undigested WAX, or xylo-oligosaccharides too large to run on TLC. This suggests that a greater proportion of the large, polymeric WAX may be degraded compared to other strains, however this does not seem to correlate with a faster growth rate, or a higher maximum OD, and due to the non-quantitative nature of TLC this may not be too reliable (**Figure 3-5**).

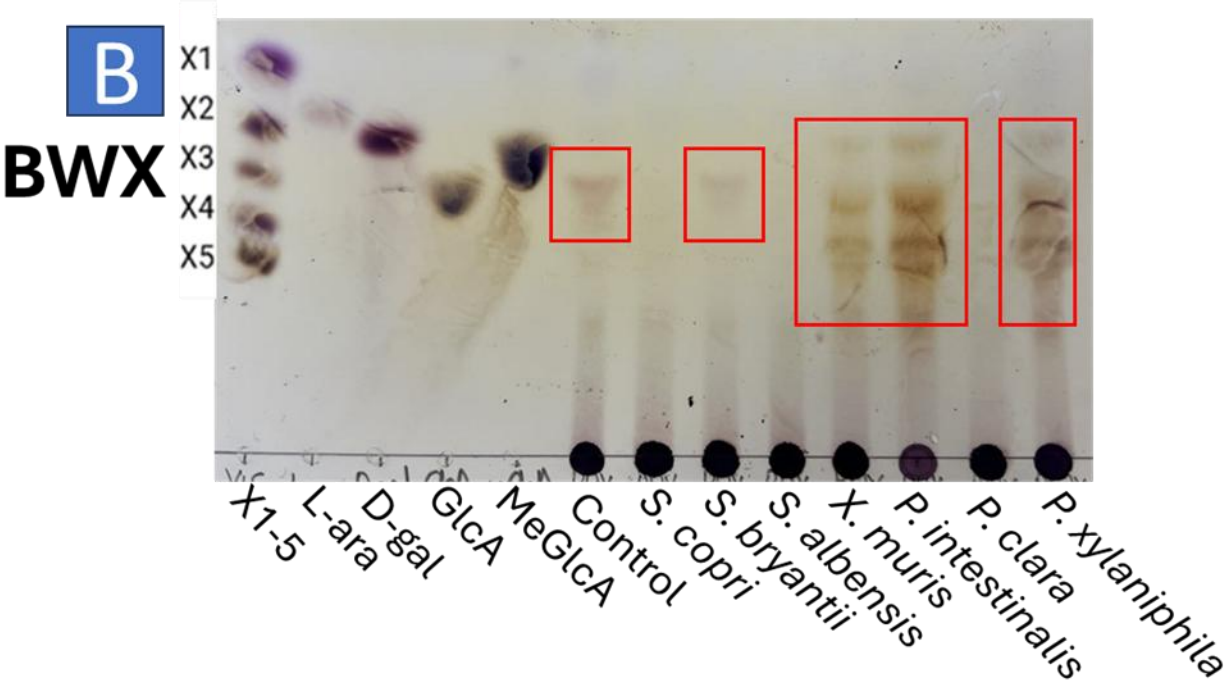
When examining spent BWX growth media on TLC (**Figure 3-6B**), a small amount of monosaccharide, potentially GlcA, can be seen in the control media, which is utilised by *S. copri* and *S. albensis,* however this alone was likely insufficient for sustained growth. For 3 strains, *X. muris. P. intestinalis* and *P. xylaniphila,* a range of small degradation products can be seen in the media after growth (***Figure 3-6B***), indicating that several small oligosaccharides resulting from the breakdown of BWX are not taken up by the bacterial cells, leading to accumulation in the media. This may explain why growth seemed poor in comparison to other xylan substrates. This poor BWX utilisation is interesting as all strains possess PUL-encoded GH10 or GH30 predicted endoxylanases (**Figure 3-1**). Studies to date show that members of these families possess activity against GX (Collins *et al.*, 2005; Biely *et al.*, 2023), so we predicted that all these *Prevotellaceae* strains should be able to utilise BWX without prior

debranching. Further studies into the activity of these predicted endoxylanase enzymes could be beneficial to further elucidate GX utilisation. Furthermore, there may be factors preventing the uptake of glucuronic acid decorated xylooligosaccharides into the cell periplasm. Overall, these data show that intestinal *Prevotellaceae* have a preference for arabinoxylans over glucuronoxylans. This may be due to a lower abundance of GX in the mammalian diet, in comparison to AX which are prominent in many cereal grains.

Patterns of OSX utilisation by *Prevotellaceae* spp. are more challenging to interpret by TLC due to presence of oligosaccharides in the control media (***Figure 3-6C***). However, following 72 hours of *P. intestinalis* and *P. xylaniphila* growth, more oligosaccharides could be seen in the media in comparison to those species who could not utilise OSX for growth (**Figure 3-5**). Interestingly, for some species, namely *S. copri, S. albensis, X. muris* and *P. clara,* some of the small oligosaccharides seen in the control media had been utilised (**Figure 3-6C**), however these were likely insufficient to allow measurable cell proliferation.

TLC analysis of media after 72 hours of growth on CX shows that for *P. intestinalis* there is accumulation of oligosaccharides and potentially monosaccharides in the media (**Figure 3-6D**), supporting the hypothesis of extracellular CX debranching as discussed previously. For all other strains, no oligosaccharide products were visible in the media (**Figure 3-6D**), indicating that for those strains capable of growth on CX (**Figure 3-5**), degradation products could be imported via SusC/D-like systems.


A
WAX
X1
X2
X3
X4
X5
X1-5
L-ara
D-gal
GlcA
MeGlcA
Control
S. copri
S. bryantii
S. albensis
X. muris
P. intestinalis
P. clara
P. xylaniphila


B
BWX
X1
X2
X3
X4
X5
X1-5
L-ara
D-gal
GlcA
MeGlcA
Control
S. copri
S. bryantii
S. albensis
X. muris
P. intestinalis
P. clara
P. xylaniphila

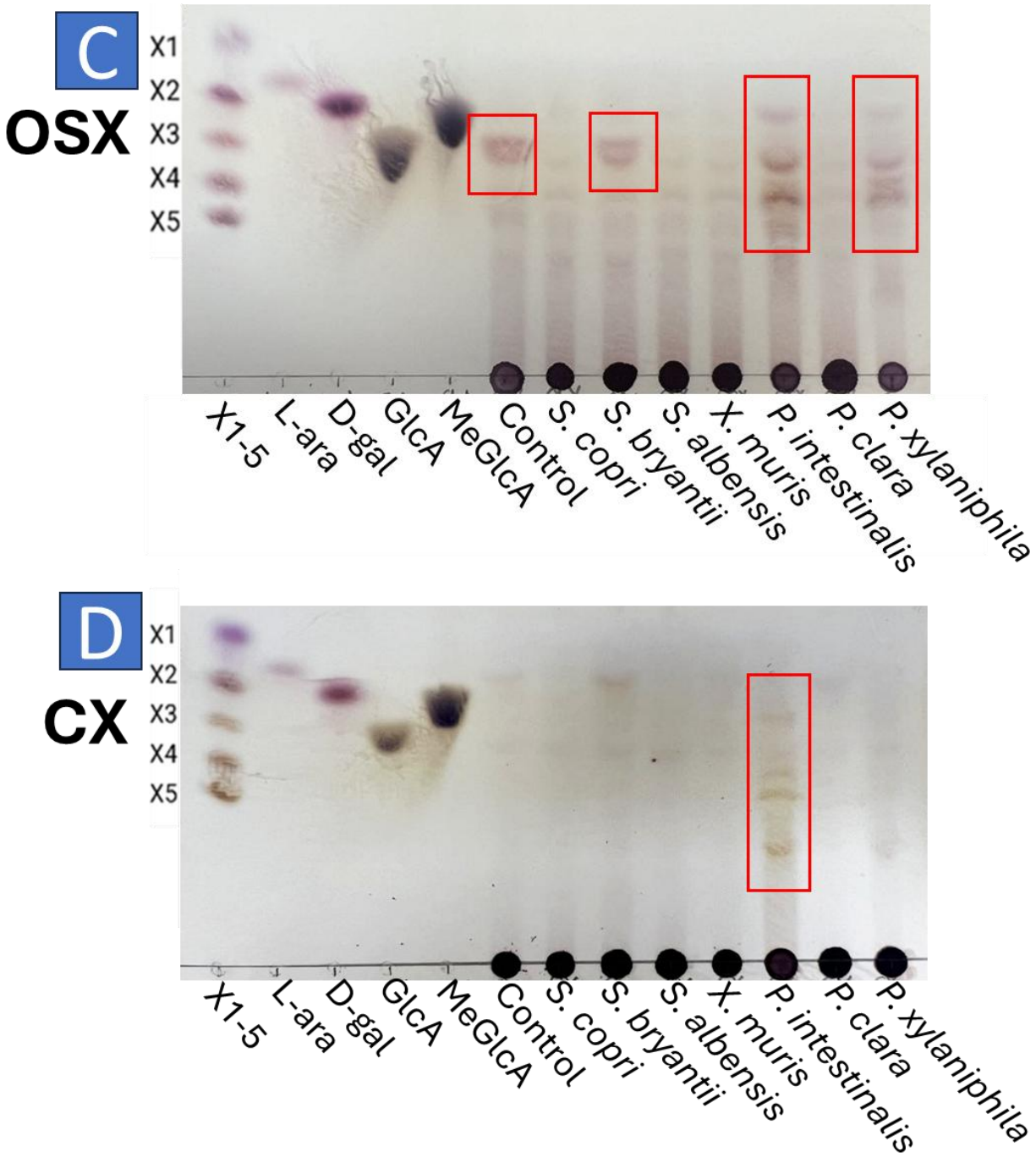


Figure 3-4: Breakdown products remaining after growth of *Prevotellaceae* spp. on different xylans. Seven individual strains were grown for 48 hours on YCFA media + 5% rumen fluid using xylans as a sole carbon source; (A) wheat arabinoxylan (WAX), (B) birchwood xylan (BWX), (C) oat spelt xylan (OSX) and (D) corn xylan (CX). Red boxes show xylan degradation products in the media.

3.2.4 Growth of human gut *Bacteroides* on different xylans

In order to assess capability of individual strains to degrade xylans and utilise them to support growth, each strain was grown on *Bacteroides* minimal media supplemented

with 5 mg mL^{-1} of either glucose, xylose, CX, wheat arabinoxylan (WAX), oat spelt arabinoxylan (OSX) or birchwood glucuronoarabinoxylan (BWX).

Growth kinetics for individual strains are shown in **Figure 3-3.** Data show that all *Bacteroides* spp. with predicted xylan PULs had the ability to grow on arabinoxylan and glucuronoxylan (WAX and BWX). Interestingly, all strains with predicted xylan PULs could utilise CX, reaching a maximum optical density comparable to that achieved during growth on simpler xylans WAX and BWX. Based on data from Rogowski *et al.* (Rogowski *et al.*, 2015) demonstrating that a knockout construct of Bo98 lacking the GH98 is incapable of growth on CX as a sole carbon source, in some *Bacteroides* spp. this growth is likely enabled by initial CX degradation by the surface located GH98 or GH5_21 enzymes. However, other *Bacteroides* may be able to utilise complex GAX, albeit by a different mechanism. These data suggest that *P. plebeius*, which does not encode either a GH98 or a GH5_21 enzyme, possesses alternative methods for the degradation of GAX for use as a growth substrate. This likely involves prior debranching of CX via the concerted activity of PUL encoded GH43 arabinofuranosidases, GH115 and GH67 α-glucuronidases and GH95 α-L-galactosidase (**Figure 3-2**), resulting in a less decorated xylan chain which can be degraded by the PUL-encoded GH10 endoxylanase. Furthermore, the 5 strains with no predicted xylan PULs, *B. caccae, B. clarus, B. finegoldii, B. fluxus* and *B. salyersiae,* did not show growth on any xylan. This indicates that, in a selection of intestinal *Bacteroides* spp., we can reliably identify and characterise xylan PULs and utilise this to determine whether strains have xylan utilisation capacity.

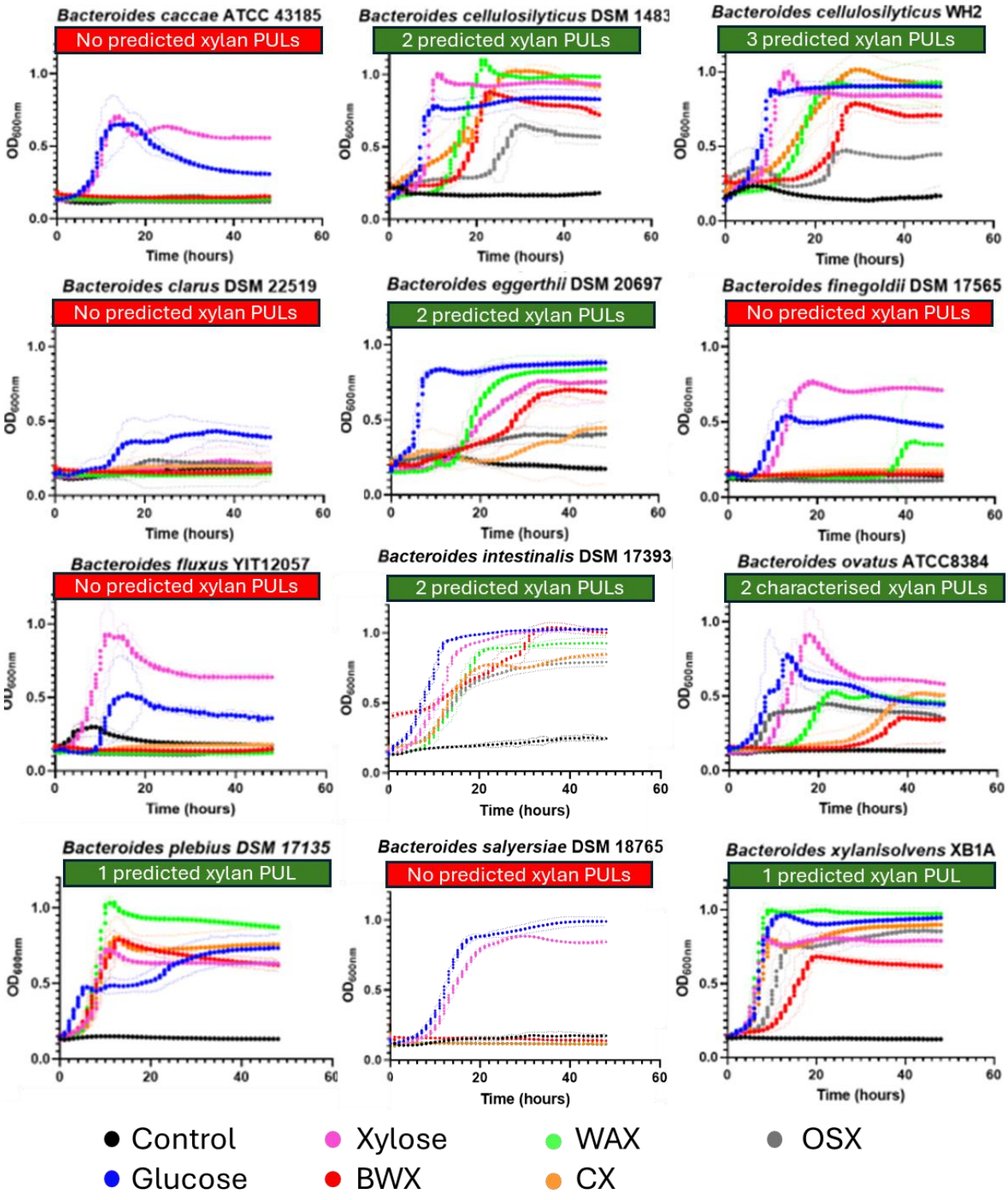


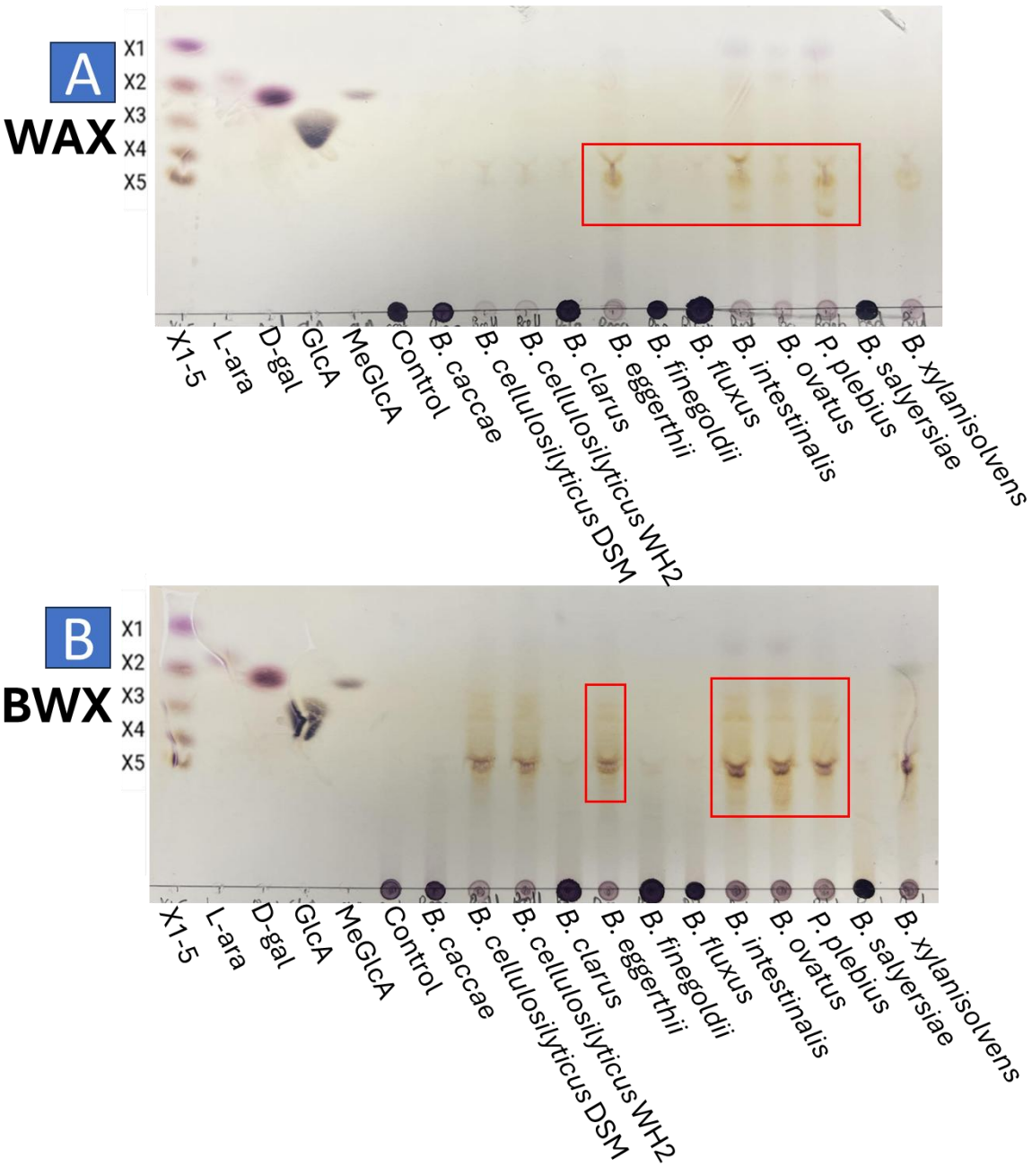
Figure 3-5: Growth of human gut *Bacteroides* strains on different xylans as the sole carbon source. 12 strains were grown on minimal media supplemented with 5 mg mL^{-1} of relevant xylan for 48 hours. Each culture was run in triplicate. Optical density at 600 nm was recorded every 10 minutes.

3.2.5 Breakdown products remaining after growth of *Bacteroides* spp. on different xylans.

In order to investigate the strategies of xylan degradation by different gut *Bacteroides* species, we analysed the profiles of xylan breakdown products in spent media after 48 hours of growth by thin layer chromatography (TLC) (***Figure 3-4***). When examining media following *Bacteroides* spp. growth on WAX, for the seven species with predicted xylan PULs, who could grow utilise xylans as growth substrates, the majority of the WAX was utilised as shown by lack of sugar on the origin of the TLC plate (**Figure 3-4A**). For *B. cellulosilyticus* DSM 14838, *B. cellulosilyticus* WH2, *B. ovatus* and *B. xylanisolvens,* very little oligo- or monosaccharide product was seen in the media (**Figure 3-4A**), indicating that these strains may utilise WAX via a "selfish mechanism", taking up all breakdown products into the cells, although further time course experiments would be required to ascertain this. Alternatively, smaller degradation products may have been released into the media during earlier stages of growth, but these were fully utilised by 48 hours. A time-course experiment would be necessary to gain a better understanding of this. On the other hand, *B. eggerthii, B. intestinalis* and *B. plebeius* all seem to leave oligosaccharide products in the media, running around close to xylotetraose and xylopentaose oligosaccharide standards on the TLC (**Figure 3-4A**), which may be imported and utilised by other species in the intestinal niche. On the contrary, during growth on BWX and OSX, all seven xylan-utilising species leave a range of small oligosaccharide products in the culture media (**Figure 3-4B + C**), indicating that the glycan importation systems of Bacteroides spp. may have some preference for arabinose- decorated xylo-oligosaccharides over glucuronoxylo-oligosaccharides, or undecorated xylo-oligosaccharides. Interestingly, during growth on CX there seems to be greater accumulation of degradation products

in the media, compared to other types of xylan, as demonstrated by the darker smear on the TLC (**Figure 3-4D**), in particular by *B. eggerthii, B. intestinalis* and the two *B. cellulosilyticus* strains. In comparisons to OSX or BWX degradation products seen in spent media, some of these CX degradation products are larger oligosaccharides (**Figure 3-4B, C + D**). This accumulation of breakdown products may be due to issues with binding and import of these oligosaccharides using SusC/D- like systems, particularly as they may be highly decorated and complex. *B. eggerthii* shows poor growth on CX (**Figure 3-3**) despite possessing a GH5_21 enzyme (***Figure 3-2***) predicted to display endoxylanase activity against complex GAX such as CX. The very dark smear shown on the TLC (**Figure 3-4D**) indicates that this enzyme is likely functional, but the products cannot be effectively imported and hence remain in the media. On the contrary, the other 6 xylan-utilising *Bacteroides* strains must be sufficiently capable of importation of highly decorated xylo-oligosaccharides to allow growth. Interestingly, *P. plebeius,* which, as discussed earlier does not encode any enzymes with predicted GAX degrading capabilities, shows the release of more small oligosaccharides into the culture media, with relatively fewer large oligosaccharides as shown by a much fainter smear (**Figure 3-4D**). This suggests less endo-activity of *P. plebeius* against CX and supports the hypothesis that this strain may be capable of complex GAX utilisation via the concerted activity of debranching enzymes, producing monosaccharides which can be imported into cells and potentially used as a carbon source to support growth. Furthermore, GH115, GH67, GH43 and GH95 enzymes, predicted to be involved in debranching of CX, possess predicted SPI signal peptides, suggesting that they are secreted into the periplasm as a lipoprotein. It is unclear whether these proteins are further secreted extracellularly, and as such it is challenging to ascertain where they function in xylan degradation.

As expected, the five species with no predicted xylan PULs, *B. caccae, B. clarus, B. finegoldii, B. fluxus* and *B. salyersiae* (**Figure 3-2**), shown to be incapable of xylan utilisation for growth (**Figure 3-3**), did not show any degradation of WAX, BWX, or CX, such that all the large xylan polysaccharide remained on the origin on the TLC (**Figure 3-4A,B + D**). However, when these strains were incubated with OSX, the TLC appears to show some xylan degradation, with the production of small oligosaccharide products, particularly by *B. clarus* and *B. salyersiae* (**Figure 3-4C**). Despite this, these strains seem incapable of the uptake and utilisation of these products, and it is likely that this low level of degradation would be insufficient to allow growth of these species.


A
WAX
X1
X2
X3
X4
X5
X1-5
L-ara
D-gal
GlcA
MeGlcA
Control
B. caccae
B. cellulosilyticus DSM
B. cellulosilyticus WH2
B. clarus
B. eggerthii
B. finegoldii
B. fluxus
B. intestinalis
B. ovatus
P. plebius
B. salyersiae
B. xylanisolvens
B
BWX
X1
X2
X3
X4
X5
X1-5
L-ara
D-gal
GlcA
MeGlcA
Control
B. caccae
B. cellulosilyticus DSM
B. cellulosilyticus WH2
B. clarus
B. eggerthii
B. finegoldii
B. fluxus
B. intestinalis
B. ovatus
P. plebius
B. salyersiae
B. xylanisolvens

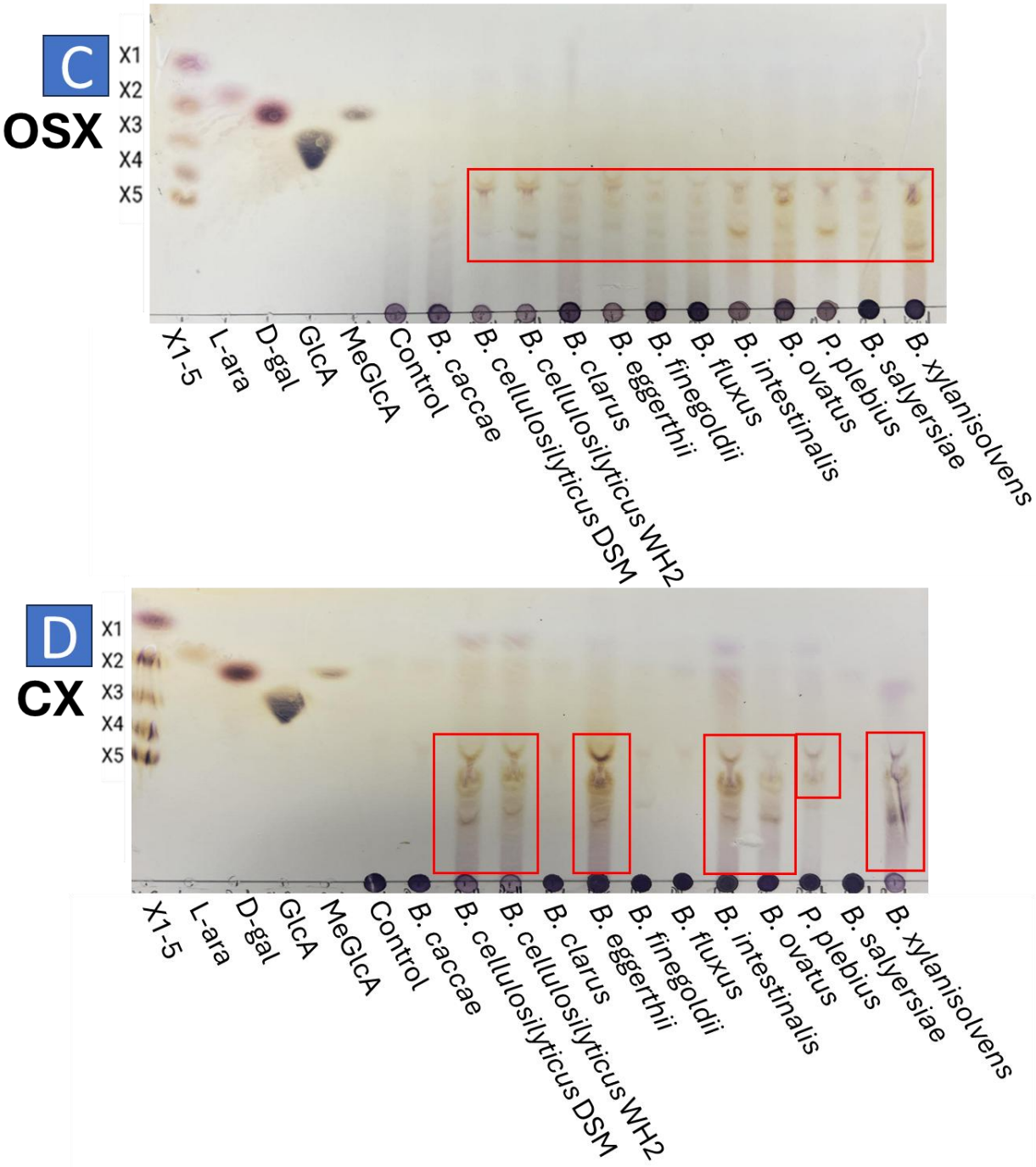


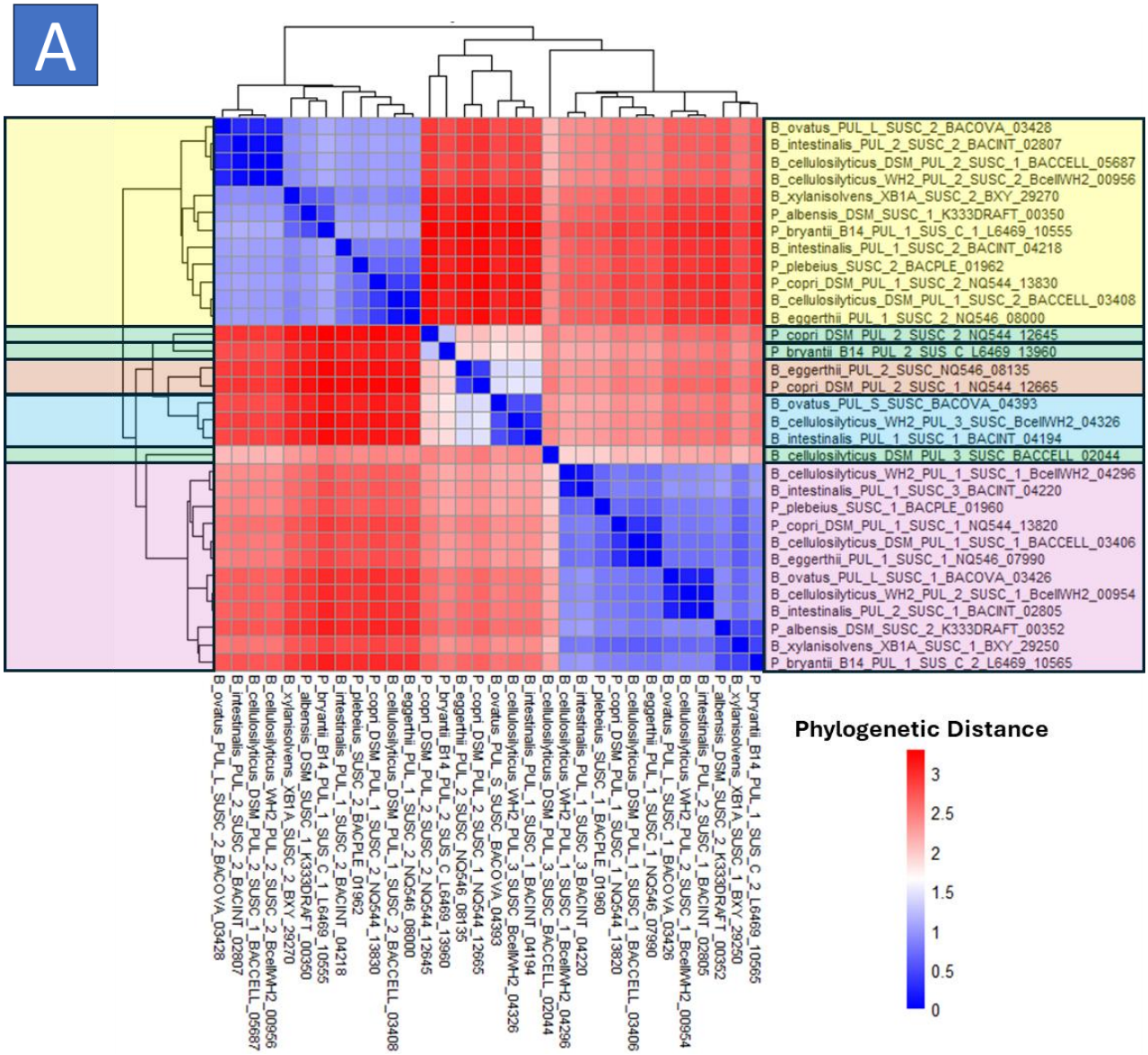
Figure 3-6: Breakdown products remaining after growth of *Bacteroides* spp. on different xylans. The different species were grown for 48 hours on minimal media using xylans as the sole carbon source and the spent media was spotted on the TLC to examine the sugar products released; (A) wheat arabinoxylan (WAX), (B) birchwood xylan (BWX), (C) oat spelt xylan (OSX) and (D) corn xylan (CX). Xylo-oligosaccharide as well as monosaccharide standards are spotted on the left of each TLC. MeGlcA – methyl glucuronic acid. Red boxes highlight xylan degradation products in the media.

3.2.6 Bioinformatic analysis of SusC- and SusD- like sequences from Bacteroidota xylan PULs

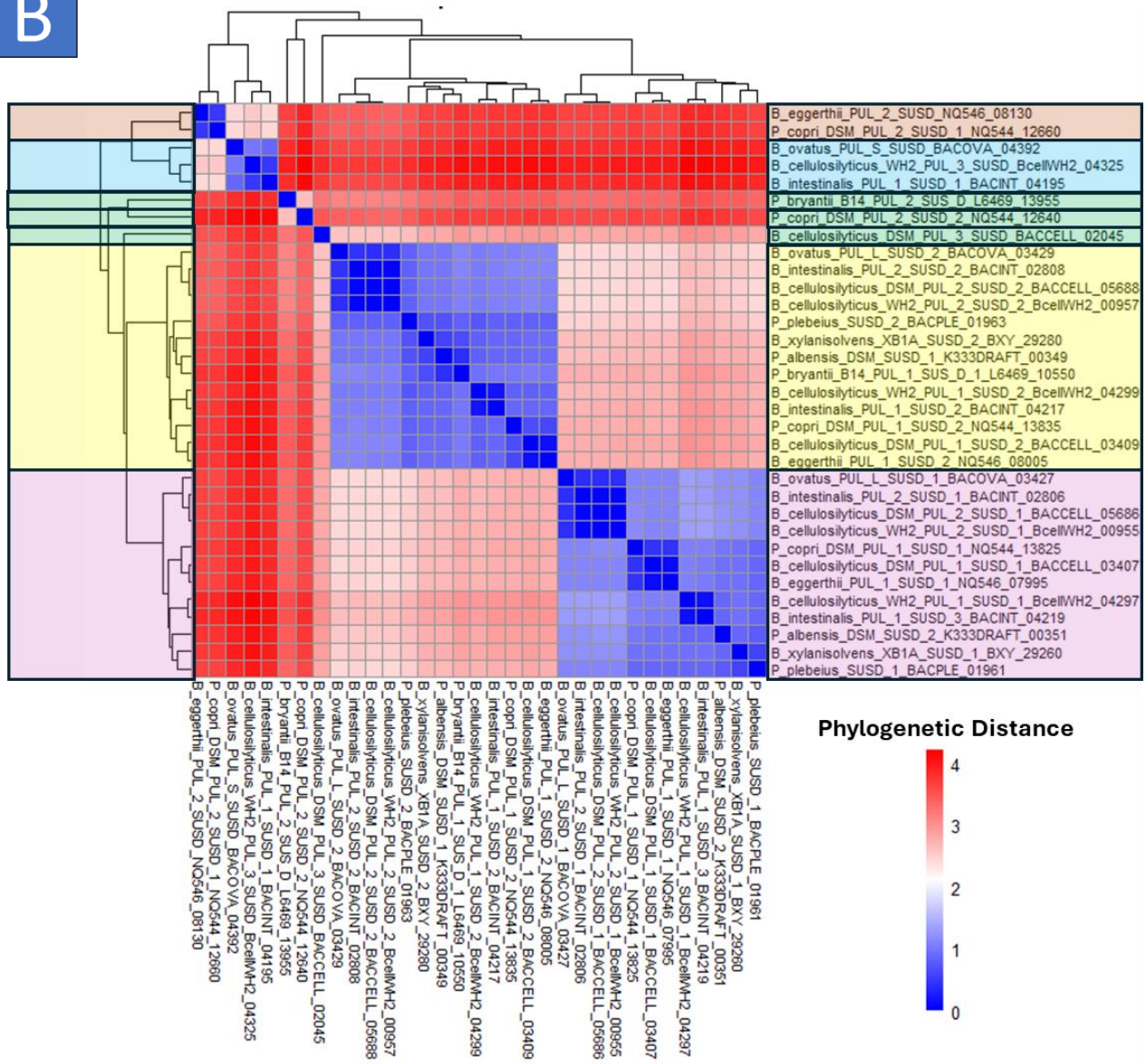
As discussed previously, *Bacteroides* PULs are characterised by the possession of at least one pair of genes SusC/D-like proteins. Many putative xylan PULs contain 2 such pairs (**Figure 3-2**). In fact, PUL analysis of *Bacteroides* and *Prevotellaceae* spp. presented in this study shows that, for these species, every species with putative xylan PUL possesses at least one with 2 SusC/D-like pairs, indicating that Bacteroidetes xylan PULs encode two tandem SusC/D-like pairs except where a strain possesses two xylan-targeting PULs, in which case the second may only have 1 SusC/D-like pair. Rogowski *et al.* (2015) showed that BoXylPUL-L (2 SusC/D pairs) was upregulated in response to growth on more complex AX or GAX, in comparison to BoXylPUL-S (1 SusC/D pair) which was upregulated during growth on simpler xylans, indicating that different SusC/D pairs may play a role in the recognition of different xylans.

In order to assess the role of different SusC/D-like pairs in xylan recognition, and to elucidate the necessity for multiple pairs within 1 PUL, bioinformatic analysis of these sequences was undertaken to assess for patterns in sequence homology and xylan utilisation capacities. Phylogenetic analysis of SusC- and SusD-like protein sequences demonstrated the ability to distinctly group closely related SusC or SusD sequences (**Figure 3-7A + B**) such that they can be separated into 4 groups of clustered sequences. These were designated group 1 through to 4, where group 1 sequences show high similarity/identity to the first SusC/D pair from *B. ovatus* XylPUL-L, and group 2 sequences show high similarity/ identity to the second pair (**Figure 3-7C**). Group 3 sequences were rare but showed homology to the SusC/D pair from *B. ovatus* XylPUL-S. Analysis shows that SusC/D-like sequences occur in distinct pairs, such that a group 1 SusC-like protein encoding gene is always found in tandem with a group

1 SusD-like gene within PULs, and the same is seen for groups 2/3. The majority of xylan PULs contain 1 SusC/D-like pair from group 1 and one pair from group 2, such that almost all analysed *Bacteroides* and *Prevotellaceae* species capable of utilising xylans possessed a PUL with a SusC/D-like pair from each of groups 1 and 2, indicating that this may play an important role in the ability to degrade xylans. Interestingly, the single SusC/D-like pair from XylPUL-S in *B. ovatus* only possesses close sequence similarity/ identity to a third SusC/D-like pair from the large xylan PUL of *B. intestinalis*, and a pair from the third xylan-targeting PUL from *B. cellulosilyticus* WH2. The third putative xylan PUL from *B. cellulosilyticus* DSM 14838, and the second from *P. bryantii,* contain SusC/D-like pairs which lack homology to any of the others studied (**Figure 3-7C**).


A
B_ovatus_PUL_L_SUSC_2_BACOVA_03428
B_intestinalis_PUL_2_SUSC_2_BACINT_02807
B_cellulosilyticus_DSM_PUL_2_SUSC_1_BACCELL_05687
B_cellulosilyticus_WH2_PUL_2_SUSC_2_BcellWH2_00956
B_xylanisolvens_XB1A_SUSC_2_BXY_29270
P_albensis_DSM_SUSC_1_K333DRAFT_00350
P_bryantii_B14_PUL_1_SUS_C_1_L6469_10555
B_intestinalis_PUL_1_SUSC_2_BACINT_04218
P_plebeius_SUSC_2_BACPLE_01962
P_copri_DSM_PUL_1_SUSC_2_NQ544_13830
B_cellulosilyticus_DSM_PUL_1_SUSC_2_BACCELL_03408
B_eggerthii_PUL_1_SUSC_2_NQ546_08000
P_copri_DSM_PUL_2_SUSC_2_NQ544_12645
P_bryantii_B14_PUL_2_SUS_C_L6469_13960
B_eggerthii_PUL_2_SUSC_NQ546_08135
P_copri_DSM_PUL_2_SUSC_1_NQ544_12665
B_cellulosilyticus_WH2_PUL_3_SUSC_BcellWH2_04326
B_ovatus_PUL_S_SUSC_BACOVA_04393
B_intestinalis_PUL_1_SUSC_1_BACINT_04194
B_cellulosilyticus_DSM_PUL_3_SUSC_BACCELL_02044
B_cellulosilyticus_WH2_PUL_1_SUSC_1_BcellWH2_04296
B_intestinalis_PUL_1_SUSC_3_BACINT_04220
P_plebeius_SUSC_1_BACPLE_01960
P_copri_DSM_PUL_1_SUSC_1_NQ544_13820
B_cellulosilyticus_DSM_PUL_1_SUSC_1_BACCELL_03406
B_eggerthii_PUL_1_SUSC_1_NQ546_07990
B_ovatus_PUL_L_SUSC_1_BACOVA_03426
B_cellulosilyticus_WH2_PUL_2_SUSC_1_BcellWH2_00954
B_intestinalis_PUL_2_SUSC_1_BACINT_02805
P_albensis_DSM_SUSC_2_K333DRAFT_00352
B_xylanisolvens_XB1A_SUSC_1_BXY_29250
P_bryantii_B14_PUL_1_SUS_C_2_L6469_10565
B_ovatus_PUL_L_SUSC_2_BACOVA_03428
B_intestinalis_PUL_2_SUSC_2_BACINT_02807
B_cellulosilyticus_DSM_PUL_2_SUSC_1_BACCELL_05687
B_cellulosilyticus_WH2_PUL_2_SUSC_2_BcellWH2_00956
B_xylanisolvens_XB1A_SUSC_2_BXY_29270
P_albensis_DSM_SUSC_1_K333DRAFT_00350
P_bryantii_B14_PUL_1_SUS_C_1_L6469_10555
B_intestinalis_PUL_1_SUSC_2_BACINT_04218
P_plebeius_SUSC_2_BACPLE_01962
P_copri_DSM_PUL_1_SUSC_2_NQ544_13830
B_cellulosilyticus_DSM_PUL_1_SUSC_2_BACCELL_03408
B_eggerthii_PUL_1_SUSC_2_NQ546_08000
P_copri_DSM_PUL_2_SUSC_2_NQ544_12645
P_bryantii_B14_PUL_2_SUS_C_L6469_13960
B_eggerthii_PUL_2_SUSC_NQ546_08135
P_copri_DSM_PUL_2_SUSC_1_NQ544_12665
B_ovatus_PUL_S_SUSC_BACOVA_04393
B_cellulosilyticus_WH2_PUL_3_SUSC_BcellWH2_04326
B_intestinalis_PUL_1_SUSC_1_BACINT_04194
B_cellulosilyticus_DSM_PUL_3_SUSC_BACCELL_02044
B_cellulosilyticus_WH2_PUL_1_SUSC_1_BcellWH2_04296
B_intestinalis_PUL_1_SUSC_3_BACINT_04220
P_plebeius_SUSC_1_BACPLE_01960
P_copri_DSM_PUL_1_SUSC_1_NQ544_13820
B_cellulosilyticus_DSM_PUL_1_SUSC_1_BACCELL_03406
B_eggerthii_PUL_1_SUSC_1_NQ546_07990
B_ovatus_PUL_L_SUSC_1_BACOVA_03426
B_cellulosilyticus_WH2_PUL_2_SUSC_1_BcellWH2_00954
B_intestinalis_PUL_2_SUSC_1_BACINT_02805
P_albensis_DSM_SUSC_2_K333DRAFT_00352
B_xylanisolvens_XB1A_SUSC_1_BXY_29250
P_bryantii_B14_PUL_1_SUS_C_2_L6469_10565
Phylogenetic Distance
3
2.5
2
1.5
1
0.5
0

B


B_eggerthii_PUL_2_SUSD_NQ546_08130
P_copri_DSM_PUL_2_SUSD_1_NQ544_12660
B_ovatus_PUL_S_SUSD_BACOVA_04392
B_cellulosilyticus_WH2_PUL_3_SUSD_BcellWH2_04325
B_intestinalis_PUL_1_SUSD_1_BACINT_04195
P_bryantii_B14_PUL_2_SUS_D_L6469_13955
P_copri_DSM_PUL_2_SUSD_2_NQ544_12640
B_cellulosilyticus_DSM_PUL_3_SUSD_BACCELL_02045
B_ovatus_PUL_L_SUSD_2_BACOVA_03429
B_intestinalis_PUL_2_SUSD_2_BACINT_02808
B_cellulosilyticus_DSM_PUL_2_SUSD_2_BACCELL_05688
B_cellulosilyticus_WH2_PUL_2_SUSD_2_BcellWH2_00957
P_plebeius_SUSD_2_BACPLE_01963
B_xylanisolvens_XB1A_SUSD_2_BXY_29280
P_albensis_DSM_SUSD_1_K333DRAFT_00349
P_bryantii_B14_PUL_1_SUS_D_1_L6469_10550
B_cellulosilyticus_WH2_PUL_1_SUSD_2_BcellWH2_04299
B_intestinalis_PUL_1_SUSD_2_BACINT_04217
P_copri_DSM_PUL_1_SUSD_2_NQ544_13835
B_cellulosilyticus_DSM_PUL_1_SUSD_2_BACCELL_03409
B_eggerthii_PUL_1_SUSD_2_NQ546_08005
B_ovatus_PUL_L_SUSD_1_BACOVA_03427
B_intestinalis_PUL_2_SUSD_1_BACINT_02806
B_cellulosilyticus_DSM_PUL_2_SUSD_1_BACCELL_05686
B_cellulosilyticus_WH2_PUL_2_SUSD_1_BcellWH2_00955
P_copri_DSM_PUL_1_SUSD_1_NQ544_13825
B_cellulosilyticus_DSM_PUL_1_SUSD_1_BACCELL_03407
B_eggerthii_PUL_1_SUSD_1_NQ546_07995
B_cellulosilyticus_WH2_PUL_1_SUSD_1_BcellWH2_04297
B_intestinalis_PUL_1_SUSD_3_BACINT_04219
P_albensis_DSM_SUSD_2_K333DRAFT_00351
B_xylanisolvens_XB1A_SUSD_1_BXY_29260
P_plebeius_SUSD_1_BACPLE_01961
Phylogenetic Distance
4
3
2
1
0

C

Species and Xylan-Targetting Polysaccharide Utilisation Loci (PUL)	Group 1 SusC/D-like sequences	Group 2 SusC/D-like sequences	Group 3 SusC/D-like sequences	Group 4 SusC/D-like sequences	Outliers
B. ovatus ATCC 8483 XylPUL-L	SusC/D 1	SusC/D 2			
B. ovatus ATCC 8483 XylPUL-S			SusC/D		
B. cellulosilyticus DSM 14838 PUL 1	SusC/D 1	SusC/D 2			
B. cellulosilyticus DSM 14838 PUL 2	SusD 1	SusC/D 2			
B. cellulosilyticus DSM 14838 PUL 3					SusC/D
B. cellulosilyticus WH2 PUL 1	SusC/D 1	SusD 2			
B. cellulosilyticus WH2 PUL 2	SusC/D 1	SusC/D 2			
B. cellulosilyticus WH2 PUL 3			SusC/D		
B. eggerthii DSM 20697 PUL 1	SusC/D 1	SusC/D 2			
B. eggerthii DSM 20697 PUL 2				SusC/D	
B. xylanisolvens XB1A	SusC/D 1	SusC/D 2			
P. plebius DSM 17135	SusC/D 1	SusC/D 2			
B. intestinalis DSM 17393 PUL 1	SusC/D 3	SusC/D 2	SusC/D 1		
B. intestinalis DSM 17393 PUL 2	SusC/D 1	SusC/D 2			
P. copri DSM 18205 PUL 1	SusC/D 1	SusC/D 2			
P. copri DSM 18205 PUL 2				SusC/D 1	SusC/D 2
P. albensis DSM 11379	SusC/D 2	SusC/D 1			
P. bryantii B14 PUL 1	SusC 2	SusC/D 1			
P. bryantii B14 PUL 2					SusC/D

Figure 3-7: Grouping of SusC- and SusD-like sequences from *Bacteroides* and *Prevotella* PULs according to protein sequence clustering. SusD (A) and SusC (B) sequences were phylogenetically analysed using the maximum likelihood approach in the phangorn package in R. Phylogenetic distances were visualised using the pheatmap package in R with a diverging colour scale from blue to red, where blue shows closest phylogenetic distances. SusC and D sequences were then characterised into groups based on phylogenetic clustering.

As almost all xylan utilising Bacteroidota species studied here possess a xylan PUL containing both a Group 1 and a Group 2 SusC/D-like pair (**Figure 3-1 + Figure 3-2**), it is possible that these differences in sequence identity/ similarity equate to differences in function, such that group 1 SusC/D-like pairs target the binding and import of slightly different xylo-oligosaccharides to those from other groups. This means that Bacteroidota species would have the capacity to utilise xylans of many

different complexities and branch types encountered in the intestinal niche, ultimately improving fitness in that environment.

3.2.7 Knockout of SusC- and SusD- like pairs from *Bacteroides ovatus* xylan PULs

In order to assess the functional role of these distinct SusC/D-like pairings in xylan utilisation, we aimed to perform knockouts of each of the three pairs in *B. ovatus* ATCC 8483, and analysis the growth capabilities of knockout strains on different xylan types.

Unfortunately, due to issues with performing conjugations, this work is yet to be completed. However, by assessing the growth profiles of *B. ovatus* knockout strains lacking SusC/D-like pairs on various xylan classes we could elucidate the role of these distinct group types on xylan utilisation capabilities. This is undoubtedly an important area of study for the future.

3.3 Discussion

Whilst the xylan-utilisation strategies of some *Bacteroides* spp. including *B. ovatus* (Rogowski *et al.*, 2015), and *B. xylaniphila* (Despres *et al.*, 2016) have been well characterised, we understand very little regarding xylan degradation capabilities of other Bacteroidota including those of the *Prevotellaceae* family.

Here we show the ability to bioinformatically identify xylan-targeting PULs via genome-wide searches for GH enzymes with known xylan-degrading activities. This builds upon the ability to automatically predict PULs in the genomes of Bacteroidota species by the identification of SusC/D homologues using the PULDB algorithm. Whilst the complete biochemical characterisation of individual PULs using recombinant techniques will remain the gold standard to ascertain the target glycan of that system (McKee *et al.*, 2021), this is not feasible for all PULs from all systems, and hence

bioinformatic predictions could provide huge advances in knowledge. For each of the 19 Bacteroidota spp. studied here, the presence of a predicted xylan PUL corresponded to the ability to utilise xylans as a growth substrate. Species without the presence of a xylan PUL lack the capability to utilise xylans as a sole carbon source for growth. These principles can be applied to all characterised Bacteroidota, as well as metagenomic assembled genomes (MAGs) from a range of ecological niches. This could prove beneficial for application in functional annotation of genomes, which will ultimately allow us to infer whole-system xylan utilisation via the annotation of metagenomic or metatranscriptomic datasets, allowing us to assess the xylan utilisation capacity of bacterial ecosystems such as the gut microbiome. Furthermore, there is potential for these strategies to be applied to many different glycan types, including, but not limited to mannans, β-glucans and xyloglucans, presenting the potential for automatic annotation of PULs in the near future, on top of existing capabilities of DBcan (Zheng *et al.*, 2023).

However, there are likely to be challenges with this, for example some PULs target multiple glycan types, or may be involved in the degradation and acquisition of non-carbohydrate nutrients, increasing complexity in prediction of target substrates (McKee *et al.*, 2021). Furthermore, Bacteroidota may possess CAZymes which are not encoded within PULs, such as the GH98 enzyme in some *Prevotellaceae* species, meaning that we cannot fully predict glycan utilisation capabilities by looking at PULs alone. Despite this, when considering the impact of the gut microbiome on human or animal health, the metagenomic prediction of glycan breakdown capabilities via the automatic annotation of PULs could allow for improved understanding of influence of factors such as diet, exercise, disease and health status on the glycan-degrading capacity of the intestinal microbiota, for example the correlation between diseases

such as Crohn's disease or colitis and increased abundance of mucous degrading bacteria in the HGM (Png *et al.*, 2010).

Although these studies suggest that we can accurately predict the ability of a given strain to degrade and utilise some types of xylan, the prediction of exactly which types of xylans are targeted may be more challenging. Based on the activities of known GH families with endoxylanase activity (Rogowski *et al.*, 2015; Biely *et al.*, 2023), and the loss of GAX-degrading capability in a *B. ovatus* GH98 knockout strain, we hypothesised that Bacteroidota strains which do not encode a GH5_21 or GH98 enzyme would have no ability to utilise complex GAX, such as CX, as growth substrates. However, here we show that 1 strain, *Phoecaeicola plebeius*, is capable of growth on CX (***Figures 3+5***) despite lacking both GH5_21 and GH98 enzymes, indicating that they must employ different strategies to degrade complex GAX. This likely involves primary extracellular debranching of the xylan chain via the concerted effort of GH43 arabinofuranosidases, GH67 α-glucuronidases and, in the case of *P. plebeius,* a GH115 α-glucuronidase and a GH95 α-galactosidase. The monosaccharide products released by these extracellular GH enzymes are likely imported into the cell where they can be utilised as a carbon source. Both strains encode GH10 enzymes which likely possess endo-activity against debranched CX chains. For both of these strains, a range of oligosaccharides are left in the media during growth, as shown by the smear on the TLC (**Figures 3-4D + 3-6D**) indicating incomplete utilisation of CX. Overall, these data suggest that the ability to degrade complex GAX in the intestinal environment may be more widespread than previously thought, however the debranching mechanism of CX utilisation described here is not ubiquitous amongst xylan utilising Bacteroidota*,* as *B. ovatus* cannot utilise CX for growth when lacking the GH98 enzyme (Rogowski *et al.*, 2015).

PUL-encoded glycan utilisation systems sometimes allow nutrient acquisition via a selfish mechanism, allowing the strain to sequester nutrients via the concerted activity of surface glycan binding proteins, extracellular endo-degradation by GH enzymes and import of relatively large xylooligosaccharides via SusC/D-like transport systems. This means that for some species and some glycan types relatively small amounts of oligosaccharide product are available to other species in the microbial community, limiting cross-feeding and providing a competitive advantage to the PUL-encoding strain (Cuskin *et al.*, 2015). There is also evidence of PUL-mediated cross-feeding by Bacteroidota, known as the sharing mechanism, which often depends on the comparative rates of oligosaccharide production and intracellular transport, for example xylooligosaccharides products from *B. ovatus* xylan degradation can cross-feed other gut commensals such as *Bifidobacteria* spp. (Rogowski *et al.*, 2015). In this study we looked at xylan-degradation products in the media following 48 hours of growth as a proxy for whether 19 Bacteroidota strains may utilise xylan derived nutrients via a selfish or sharing mechanism.

Interestingly, the uptake of xylan-derived products into the cells of intestinal Bacteroidota is not entirely species dependent but rather seems to vary between types of xylan. WAX breakdown products seem to be entirely taken up into the periplasm of Bacteroidota cells, so are unlikely to be shared with other species in the microbial community, indicating that these may be a preferential nutrient source for members of the gut microbiome. Evolution of effective mechanisms for the selfish utilisation of AXs may have occurred due to the prevalence of AX in the diet of humans and domesticated animals based on the abundance in important food crops such as cereal grains. Furthermore, AXs can be degraded by endoxylanase enzymes commonly found in intestinal bacteria, including those from GH families 10, 11 and 5_21, so can

be degraded by species lacking complex xylan degradation systems, increasing competition within the intestinal niche. Based on the experimental evidence presented here, we predict that *Prevotellacaea*, with the exception of *P. intestinalis,* show a similar selfish uptake of xylooligosaccharides during growth on the more complex and highly decorated CX, however further time course experiments are required to prove this. On the contrary, following growth of *Bacteroides* on CX, a range of oligosaccharides remain in the media after 48 hours of growth, indicating that in the gut niche these breakdown products may be available for cross-feeding with other members of the HGM. Interestingly, during growth on OSX and BWX, xylan utilising Bacteroidota spp. seem to release a range of monosaccharides or small oligosaccharides, which can be visualised on TLC following 48 hours of growth. This suggests that Bacteroidota do not use these small degradation products, perhaps due to preferential import of longer oligosaccharides for further periplasmic processing, leaving them available for cross-feeding with other members of the gut microbiota.

Some species, including *B. clarus* and *B. salyersiae,* seem to have some capability to degrade OSX, resulting in the presence of oligosaccharides in the growth media, despite the lack of a xylan PUL or any known xylan-active enzymes encoded elsewhere in the genomes. However, these strains lack the ability to utilise these breakdown products for growth. One potential explanation for this may be non-specific activity against OSX due to promiscuity of other enzymes with endo-activity. This might mean that they can degrade xylan chains, which are largely undecorated, but not those of more highly decorated AXs such as WAX, or GX or GAX. It is likely that such activity would release insufficient suitable products for use as a carbon source to support growth.

Prediction of xylan active PULs within *Bacteroides* and *Prevotellaceae* has allowed for further characterisation of the constituent proteins encoded with them. As discussed previously, all PULs contain at least 1 SusC/D-like pair which function in the binding of glycans and their transport across the outer membrane and into the periplasm. Based on this it is possible that differences between SusC/D-like pairs may constitute differences in the specificity of glycan recognition and translocation.

Many of the xylan-utilising Bacteroidota species studied here possess a large xylan PUL with 2 SusC/D-like pairs in tandem. Here we show that Bacteroidota SusC/D-like sequences from xylan PULs can be clustered into 3 main groups, and we predict that we may see structural conservation within these groups. Here, we attempted to compare the structural Alphafold2 predictions for members of each of these groups, however, structural models did not make sense in accordance with known structures of SusC/D-like complexes (White *et al.* 2023) and as such were not included in this work. It is possible that the presence of multiple SusC/D-like systems within xylan PULs may allow for the binding and subsequent transport of different xylan types, allowing for broad specificity with respect to xylan utilisation.

Detailed studies into xylan utilisation by *B. ovatus,* demonstrate that the larger xylan PUL, XylPUL-L, possesses one group 1 SusC/D-like pair, and one belonging to group 2, as is typical of many of the xylan-targeting PULs from gut Bacteroidota. Expression of this PUL was upregulated during growth on WAX and CX, but not on the glucuronoxylan, BWX (Rogowski *et al.*, 2015). A XylPUL-L knockout strain was incapable of growth on complex GAX but could still utilise simpler AX and GX (Rogowski *et al.*, 2015), presumably due to XylPUL-S mediated utilisation of simpler xylans. This suggests that the 2 SusC/D- like pairs from Groups 1 and 2 together

function in the import of various xylan types including AXs and GAXs, but do not recognise GXs.

The smaller xylan-targeting PUL from *B. ovatus*, XylPUL-S, was shown to target simpler xylans (Rogowski *et al.*, 2015), and a genetic knockout strain lacking XylPUL-S showed growth defects on AX and GX. In this case, lack of growth of ΔXylPUL-S strain on AX is likely due to this PUL possessing the only GH10 endoxylanase of *B. ovatus,* such that its removal prevents effective degradation. Based on previous studies it is possible that Group 3 SusC/D pairs, as found in BoXylPUL-S, may function in the recognition, and import of simpler, less highly decorated xylans (Rogowski *et al.*, 2015).

Overall, despite differences between sequences of different SusC/D- groups we do not understand how these differences equate to function, and ultimately the ability of the individual Bacteroidota strains to utilise breakdown products of different xylan types. Further characterisation of different SusC/D groups will be necessary to elucidate this. As mentioned previously, this study aimed to individual knockout Group 1, Group 2 and Group 3 SusC/D-like pairs from *B. ovatus* and assess the impact on the ability to utilise different xylan types for growth. This will provide further information of the role of each pair in xylan utilisation. Furthermore, studies into the binding kinetics of SusD-like proteins with different xylan types by isothermal titration calorimetry will provide information on the substrate specificity of different groupings. Co-crystallisation or crystal soaks to solve the structure of SusC/D-xylan complexes will allow visualisation of the interactions between xylans and SusC/D proteins, consequently improving understanding of xylan utilisation.

In this chapter we have shown that we can bioinformatically predict the ability of Bacteroidota spp. to utilise xylans, via the identification of PUL encoding GH10, GH5_21, GH30_8 or GH98 proteins. There also appear to be divergent mechanisms for degradation of highly decorated GAX, such as CX, involving endo-activity by GH5_21 or GH98 endoxylanases, or involving extracellular debranching and subsequent import of monosaccharide products, and/or endo-type cleavage of the resulting less highly decorated xylan chains. Whilst most xylan-utilising *Bacteroides* spp. studied seem to have broad xylan-degrading capacities, members of the *Prevotellacaea* family have a preference for utilisation of AX or GAX over GX such as BWX.

SusC/D-like sequences from within these xylan-targeting PULs occur in distinct pairs, which can be grouped according to sequence similarity and identity. Almost all xylan-utilising strains possess a PUL encoding a group 1 SusC/D- like pair and a group 2 SusC/D-like pair, indicating that these likely play an important role in xylan-utilisation systems. It is not yet clear how these differences in sequence, and likely structure, may equate to changes in function with respect to binding and import of xylan-degradation products, and further studies are required to elucidate this.

4. Structural and functional analysis of complex xylan degradation by GH98 family endoxylanases

4.1 Introduction

Bacteroides ovatus, a prominent member of the HGM, is considered a generalist with respect to polysaccharide utilisation, encoding a broad repertoire of CAZymes conferring the ability to degrade a wide range of hemicelluloses, including xylans (Flint, Scott, Duncan, *et al.*, 2012; Martens *et al.*, 2011). Xylans are a major structural component of plant cell walls and are prevalent in the human diet, particularly from cereal grains such as oat or corn (Curry *et al.*, 2023). Xylans are highly variable between species and tissues but are classified into three major classes reflecting the type of branching of the conserved β-1,4 linked xylose backbone: glucuronoxylans (GXs), arabinoxylans (AXs) and glucuronoarabinoxylans (GAXs) (**Figure 1-6**) (Agger *et al.*, 2010; Xu *et al.*, 2022). GAXs, exemplified by corn xylan (CX) are thought to be more recalcitrant to breakdown by members of the HGM than other classes of xylans, likely due to increased number and complexity of side chains (Agger *et al.*, 2010).

Many endo-xylanase enzymes, such as those belonging to GH families 10, 11 and 43, cannot degrade complex GAXs without prior debranching (Rogowski *et al.*, 2015; Agger *et al.*, 2010). In fact, only two enzyme families, GH98 and GH5 subfamily 21, have been found to have any endo-xylanase activity against complex GAX, CX (Rogowski *et al.*, 2015; Aspeborg *et al.*, 2012). Whilst the GH5_21 endo-xylanases display activity against a range of AXs, due to the recognition of arabinose side chains, GH98 activity is specific to complex GAX (Rogowski *et al.*, 2015), a unique and intriguing specificity.

One mechanism of GAX utilisation involves cell surface endo-degradation by GH98 or GH5_21 enzymes to produce smaller oligosaccharides which can be imported into the periplasm via SusC/D- like systems for further processing. With reference to chapter 3, BoXylPUL-L and many other *Bacteroides* xylan PULs encode enzymes capable of debranching GAXs such as CX (**Figure 3-3, Figure 4-1**), including GH43 arabinofuranosidases, specifically targeting double (for example Bacova_03417) or single (for example Bacova_03425) xylose and arabinose decorations. Alpha-glucuronidases from GH families 67 and 115 (Bacova_03449) respectively remove terminally and internally substituted, GlcA and MeGlcA decorations from xylan or xylooligosaccharide chains (**Figure 4-1**) (Rogowski *et al.*, 2015). Rogowski *et al.* also showed that a member of the GH95 family, Bacova_03438, removes L-galactose from CX. In *B. ovatus* these debranching enzymes are periplasmic (Rogowski *et al.*, 2015), therefore functioning on imported xylooligosaccharides, produced from GH98 degraded xylan chains. Other Bacteroidota, such as *P. plebieus* possess extracellular xylan debranching enzymes, allowing growth on CX without a GAX-specific endo-xylanase enzyme, as shown in Chapter 3.

Rogowski *et al.* demonstrated that knocking out the gene encoding the GH98 enzyme in *B. ovatus*, Bo98, results in loss of ability of the bacterium to grow using CX as the sole carbon source (Rogowski *et al.*, 2015), and that this phenotype can be rescued by addition of recombinant Bo98 to the media, demonstrating the importance of this enzyme for extracellular utilisation of complex GAX in *B. ovatus.*

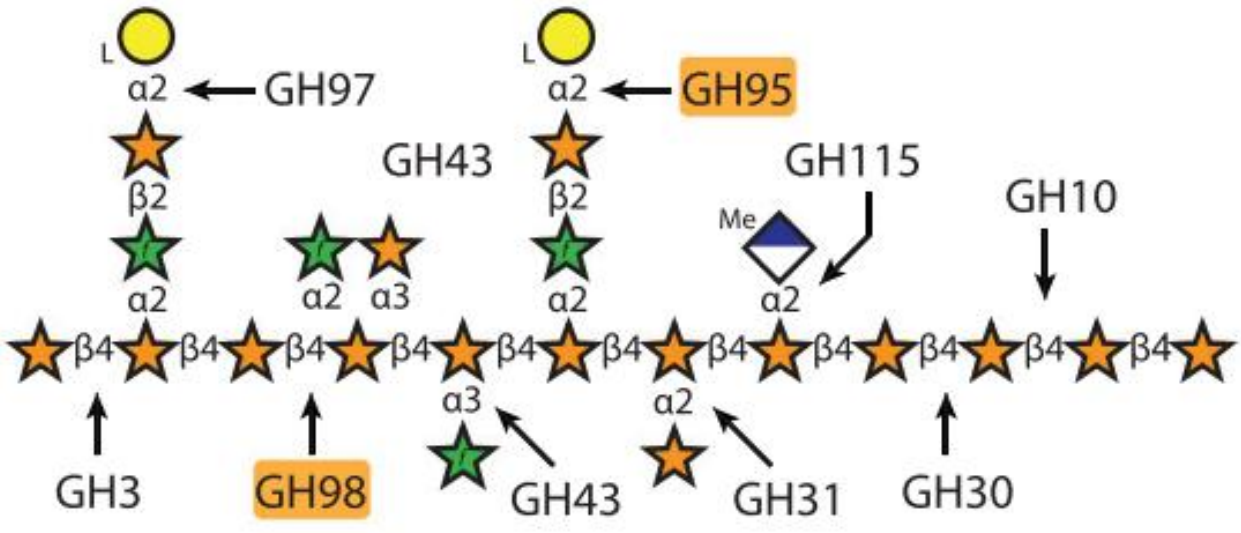


Figure 4-1: Glycoside hydrolase enzymes targeting different linkages within complex glucuronoarabinoxylans (GAX).

Interestingly, the first characterised member of the GH98 family was an endo-β1,4-galactosidase, from pathogen *Clostridium perfringens* (*Cp*GH98) (Anderson *et al.*, 2005). This enzyme displayed catalytic activity against blood group sugars A and B, with an inverting catalytic mechanism, cleaving the β1,4 linked galactose-GlcNAc bond to generate a trisaccharide product (**Figure 4-2**).

Subsequent characterisation of other GH98 β-galactosidases demonstrated that there were differences in specificity between different members of the family. For example, a GH98 β-galactosidase encoded by *Streptococcus pneumoniae* TIGR4 (*Sp4*GH98) possesses activity specific to Lewis Y antigen (Higgins *et al.*, 2009) and a recently characterised enzyme from the mucus-degrading human gut symbiont *Ruminococcus gnavus, Rg*GH98, is specific for blood group A tetrasaccharide type II (Wu *et al.*, 2021).

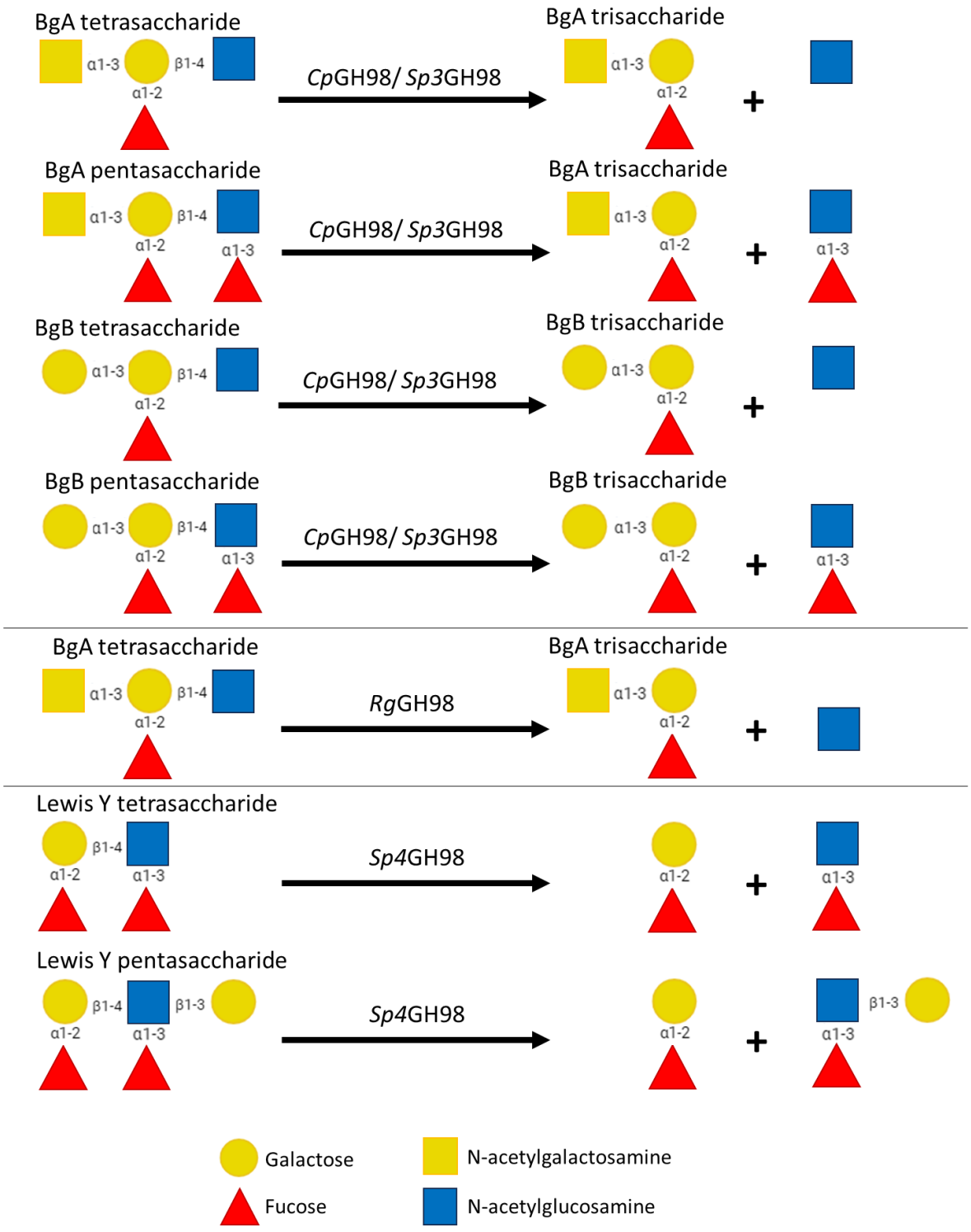


Figure 4-2: Structures of blood group sugars and their degradation by characterised GH98 endo-galactosidase enzymes. GH98 enzymes from *Clostridium perfringens* (*Cp*GH98) (Anderson *et al.*, 2005), and *Streptococcus pneumoniae* Sp3-BS71 (*Sp3*GH98) (Higgins *et al.*, 2009) possessed activity against blood group A and B (BgA/B) sugars, whereas *Rg*GH98 from *Ruminococcus gnavus* displayed specific activity against BgA (Wu *et al.*, 2021). *S. pneumoniae* TIGR4 *Sp4*GH98 was active solely against Lewis Y tetra- and pentasaccharides (Higgins *et al.*, 2009).

Activity of these 4 characterised GH98 β-galactosidase enzymes against blood group sugars is classed as endo- activity, as the enzymes recognise multiple sugars in the negative subsites (**Figure 4-3 A+B**). Despite this, the activity, and active site morphology, can be viewed as exo-type, cleaving towards the non-reducing end of a glycan chain. Consequently, for both *Sp3*- and *Sp4*GH98 structures we see a pocket-like active site, as expected in exo-acting GHs (**Figure 4-3 A+B).** On the contrary, Bo98 seems to possess strict endo-activity (Rogowski *et al.*, 2015), and due to the length and complexity of GAX chains we would expect to see significant differences in the active site conformation, with an elongated and open cleft for binding to a highly decorated xylan chain.

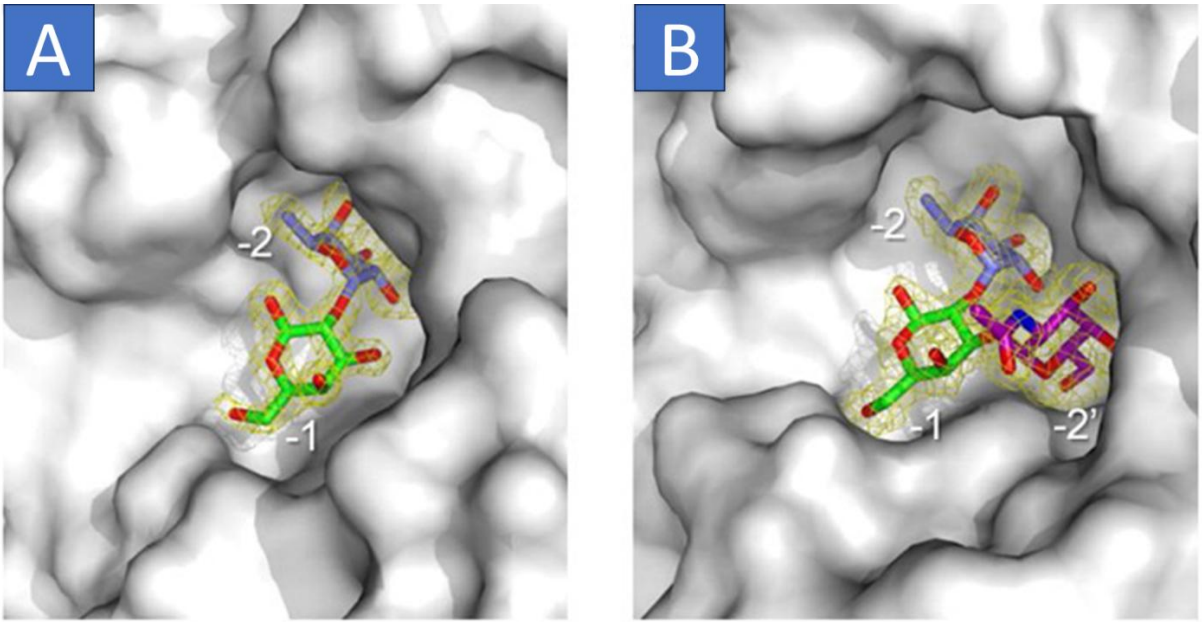


Figure 4-3: **Active sites of S. pneumoniae GH98 endo-galactosidase enzymes bound to product.** (A) *S. pneumoniae* TIGR4 (*Sp4*GH98) pocket-like active site containing H disaccharide from the degradation of Lewis Y tetrasaccharide. (B) *S. pneumoniae* Sp3-BS71 (*Sp3*GH98) containing A trisaccharide. Green – galactose; grey – fucose; magenta – N-acetylgalactosamine.

Bo98, the first, and to date only, characterised endo-xylanase member of GH98, is encoded within the large xylan-targeting PUL of *B. ovatus* (**Figure 1-20**). As mentioned

previously, Bo98 possesses a C-terminal carbohydrate binding module (CBM), belonging to family 35, presumably involved in binding to ligand to improve catalytic efficiency of the enzyme, although whether this is a functional CBM is yet to be explored (Rogowski *et al.*, 2015; Boraston *et al.*, 2004). Bo98 is presented on the surface of *B. ovatus*, and likely functions in the breakdown of long chain GAX into large oligosaccharides which can be transported into the periplasm (Rogowski *et al.*, 2015). Furthermore, it is possible that Bo98 could form part of the complex xylan utilisome, an outer membrane complex involved in surface GAX binding, degradation and uptake, composed of the SusCD-like pair, as well as an SGBP and the GH98 enzyme, however there is not yet any experimental evidence to prove this.

This production of large oligosaccharides suggests that Bo98 recognises infrequently occurring decorations within GAX, however little is known regarding the exact specificity determinants of this enzyme. Pre-treatment of CX with a cocktail of xylan-active enzymes to remove all branching, prior to incubation with Bo98, unsurprisingly created a substrate inaccessible to Bo98 (Rogowski *et al.*, 2015). Removal of the double-specific arabinofuranosidase from this cocktail created a product which could be degraded by Bo98. This suggests that xylose residues decorated at O2 and O3 with arabinose maybe critical for Bo98 activity. However, as these would be present in WAX, against which Bo98 is inactive, it has been proposed that decoration of O2 with arabinose and O3 with xylose may be the crucial specificity determinant of Bo98 (Rogowski *et al.*, 2015).

Overall, although the specificity of BoGH98 endo-xylanase for GAX has been shown, the precise structural features that drive this unique specificity are not known. Furthermore, it is not known if this novel specificity is common in other members of the GH98 family. Here, we show the first structure for a GH98 endoxylanase bound to

a complex decorated xylooligosaccharide, which reveals the enzyme likely requires double xylose-arabinose substitutions as well as the presence of glucuronic acid for catalytic activity. Characterisation of several other putative endo-xylanase GH98s, from bacteria residing in different environmental niches, demonstrates strict specificity for complex GAX like the Bo enzyme, as well as conservation of the C-terminal CBM35 domain within the endoxylanses from GH98.

4.2 Results

4.2.1 Distribution of GH98 enzymes

Two known activities exist within the GH98 family, endo-β1,4-xylanase against complex GAX, and endo-β1,4-galactosidase activity against blood group sugars (**Figure 4-4**). In order to investigate the environmental and taxonomic distribution of GH98 enzymes, phylogenetic analysis of all GH98 sequences within the CAZy database was performed (Cantarel *et al.*, 2009) (**Figure 4-4**). Based on activities of characterised enzymes and phylogenetic analysis, distinct clustering could be seen allowing tentative putative classification of sequences into endoxylanase and endo-galactosidase activity types.

As expected, predicted blood group targeting endo-galactosidases are commonly found in pathogenic bacteria, with high prevalence within the *Streptococcus* genus (Higgins *et al.*, 2009; Anderson *et al.*, 2005), but also in host associated gut mutualists such a *R. gnavus (***Figure 4-4**) (Wu *et al.*, 2021). Bacteria encoding GH98 enzymes with predicted endo-xylanase activity are found in a wide range of environmental niches, including soil, seawater and plant pathogens, as well as human gut or rumen symbionts. In all these niches it is clear that the capacity to degrade complex xylans from within plant cell walls may confer a competitive advantage. Notably, GH98

enzymes appear to be particularly well distributed within members of the plant pathogen genus *Xanthomonas*, more so than in gut symbionts such as *Bacteroides,* indicating the potential importance of degradation of complex xylans in the plant cell wall for invasion and pathogenicity.

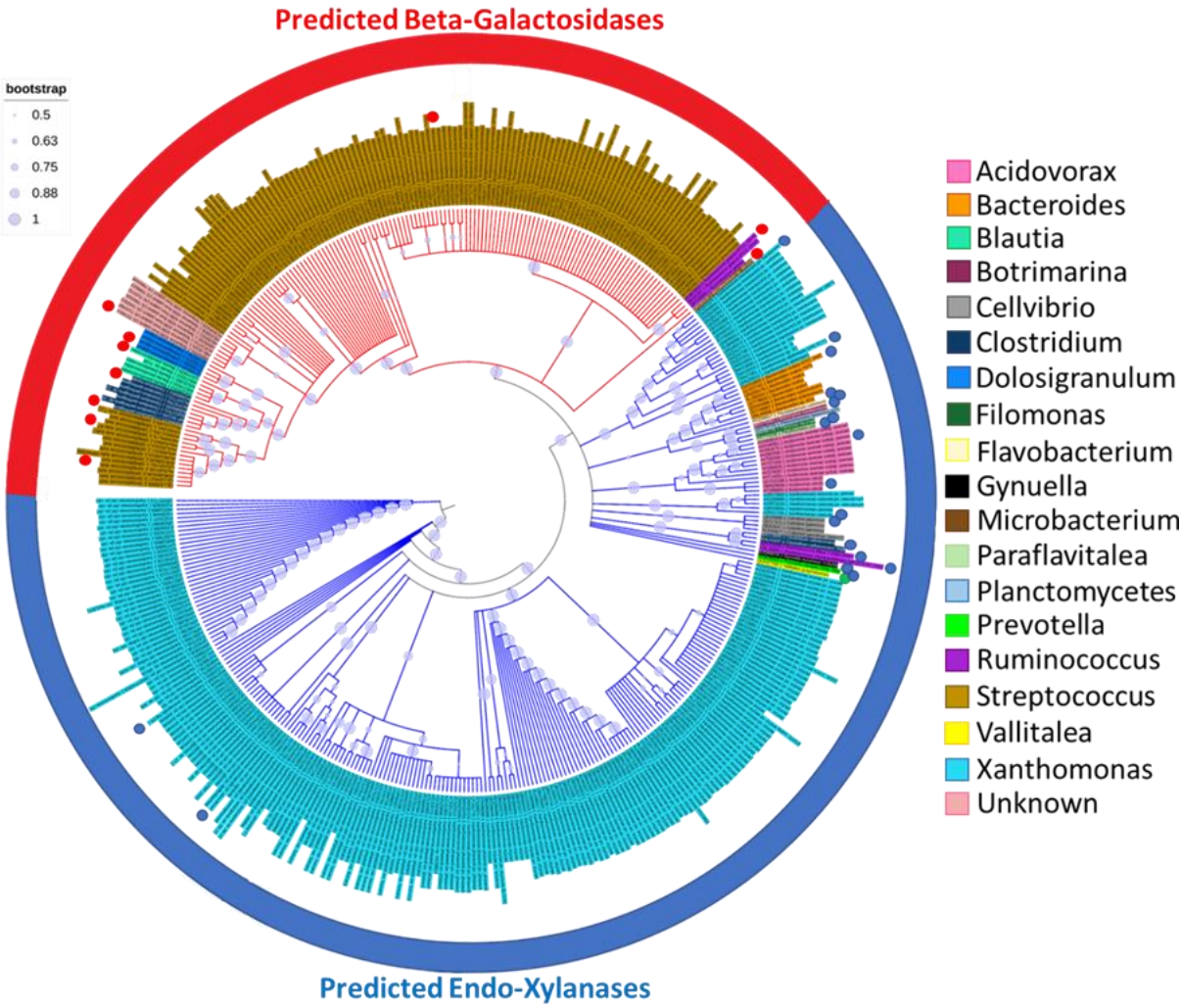


Figure 4-4: Phylogenetic tree of GH98 sequences found within the CAZy Database. 540 GH98 protein sequences from the CAZy database (Cantarel *et al.*, 2009) were used to construct the tree utilising the maximum likelihood method, calculated in MEGA11 (Tamura *et al.*, 2021). Red outer circle indicates predicted endo-galactosidase enzymes, whereas blue indicated predicted endo-xylanase activity. Clustering of sequences using CDhit (Fu *et al.*, 2012) with a 90% similarity threshold returned 30 representative sequences, indicated by blue or red dots. Sequences with blue dots possess a CBM35, those with red dots possess CBMs belonging to other families. Based on this we predict that all putative GH98 endoxylanases possess a CBM35 domain.

In order to improve understanding of domain composition and associated functional diversity within the GH98 family, GH98 sequences from within the CAZy database were clustered at 90% using CDhit, resulting in 30 clusters at this similarity threshold. For each of these clusters, one consensus GH98 sequence was provided by CDhit. The domain composition of these 30 sequences was analysed using DBcan, Pfam and Interpro (**Figure 4-5**). Of these 30 representative sequences, 19 possessed a CBM35 domain following the C-terminal of the catalytic domain, as seen in Bo98. These 19 protein sequences all clustered with Bo98 in the phylogenetic tree, distinct to characterised endo-galactosidases (**Figure 4-4**). Based on phylogeny and the conserved presence of the CBM35 domain in all enzymes clustering with characterised endoxylanases in the phylogenetic tree, it was predicted that these enzymes possess endoxylanase activity.

A further 7 of these 30 representative GH98 sequences contained at least one N-terminal CBM belonging to family 51, and the remainder possessed C-terminal CBMs from family 32 or 47 (**Figure 4-5**). These were all predicted to possess beta-galactosidase activity against blood group sugars, as any enzymes lacking the CBM35 domain were phylogenetically clustered more closely with characterised blood-group sugar active enzymes, *Cp*GH98, *Rg*GH98, *Sp3*GH98 and *Sp4*GH98. This provided further evidence for the functional characterisation predictions presented in this work.

Interestingly, only GH98 enzymes encoded by members of the *Bacteroides* genus possess the N-terminal domain of unknown function (DUF) 4988 (**Figure 4-5**), as in *B. ovatus,* hypothesised to be a spacer domain, putatively involved in the correct

positioning of the enzyme within the outer membrane utilisome complex (White *et al.*, 2023; Larsbrink *et al.*, 2014), although this has not been experimentally proven as yet.

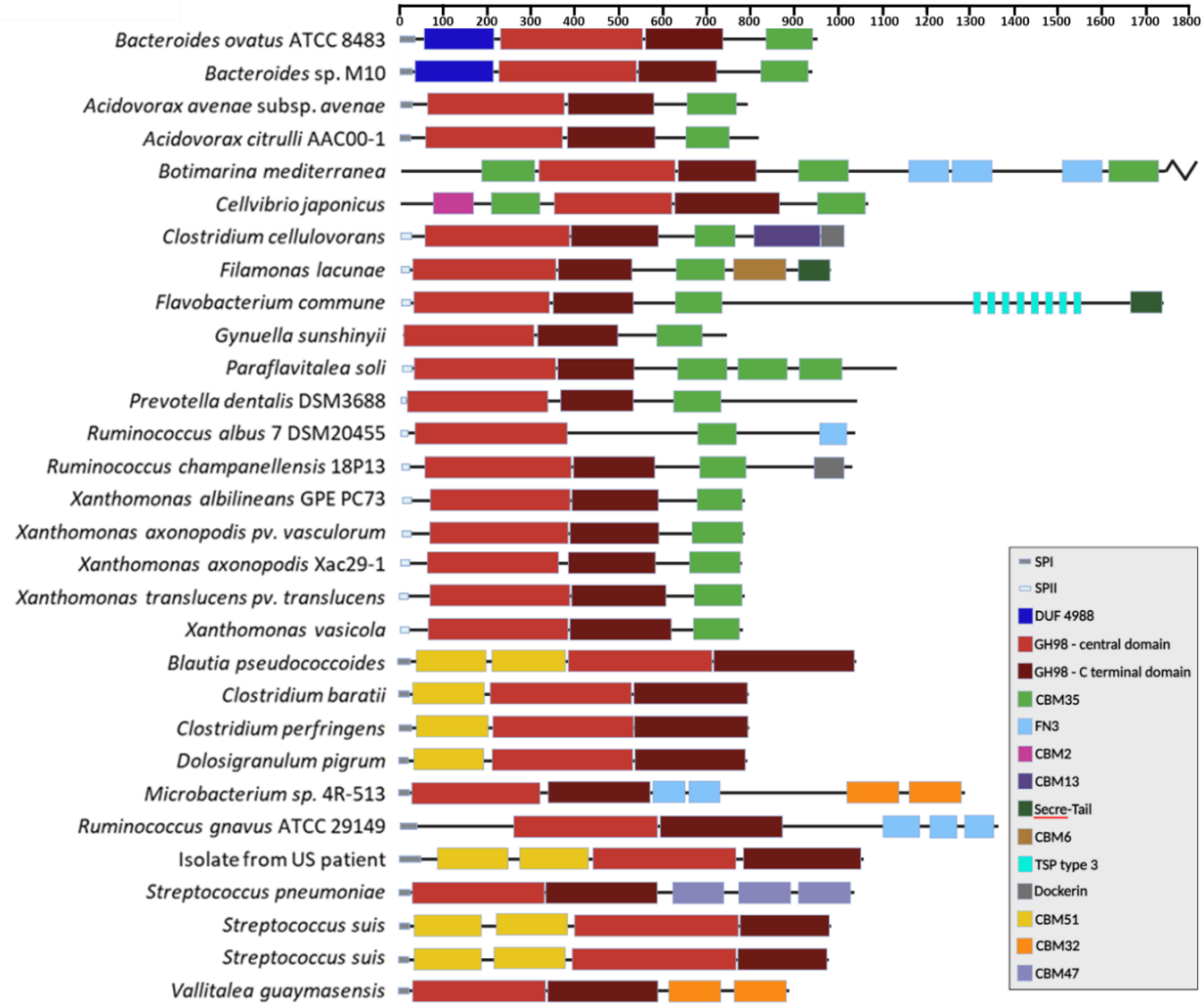


Figure 4-5: Domain composition of 30 representative GH98 enzymes. All GH98 sequences within the CAZy database (Cantarel *et al.*, 2009) were clustered at 90% using CDhit (Fu *et al.*, 2012), demonstrating 30 representative sequences. DBcan and Interpro were utilised to classify domains present in each of the 30 sequences. Comparison of phylogenetic clustering with conservation in domain compositions allowed putative annotation of function. Predicted endoxylanase enzymes possess a C-terminal CBM35, whereas predicted endo-galactosidase enzymes possess multiple CBMs belonging to families 51, 32 or 47. Only sequences from the genus *Bacteroides* possess an N-terminal domain of unknown function (DUF4988).

As well as Bo98, 4 further predicted GH98 enzymes from various ecological niches were chosen for further analysis, to test the hypothesis that all CBM35 containing GH98 enzymes possess endoxylanase activity. These were from the human gut *R.*

champanellensis 18P13 (Rc98), the soil bacterium *Cellvibrio japonicus* UEDA107, (Cj98), the sheep ruminal species *Segotella albensis* M384 (Pa98) and economically relevant rice pathogen *Xanthomonas oryzae pv. Oryzae* KXO85 (Xo98*)*. These enzymes possess 42%, 51%, 38% and 46% identity to Bo98, respectively. Cj98, Pa98 and Xo98 were chosen to increase understanding of GH98 activity in a range of environments, and Rc98 due to the interest in this cellulosome-producing strain in the overall project.

Sequence similarity network (SSN) analysis of members of the GH98 family demonstrated three major clusters, each containing functionally characterised enzymes (**Figure 4-6**). A fourth smaller cluster contains the blood group A active β-galactosidase encoded by *R. gnavus* (Wu *et al.*, 2021). Bo98 sits within the first cluster with all other putative endoxylanase containing GH98 enzymes, including those more phylogenetically distant sequences encoded by *Xanthomonas* spp. and the four enzymes, Rc98, Pa98, Cj98 and Xo98 chosen for further analysis. Furthermore, all 19 representative GH98 sequences shown to possess a C-terminal CBM35 lie in this cluster (**Figure 4-6**), providing further evidence for two distinct activities within the GH98 family, and the conservation of C-terminal CBM35 domains within endoxylanase GH98 enzymes.

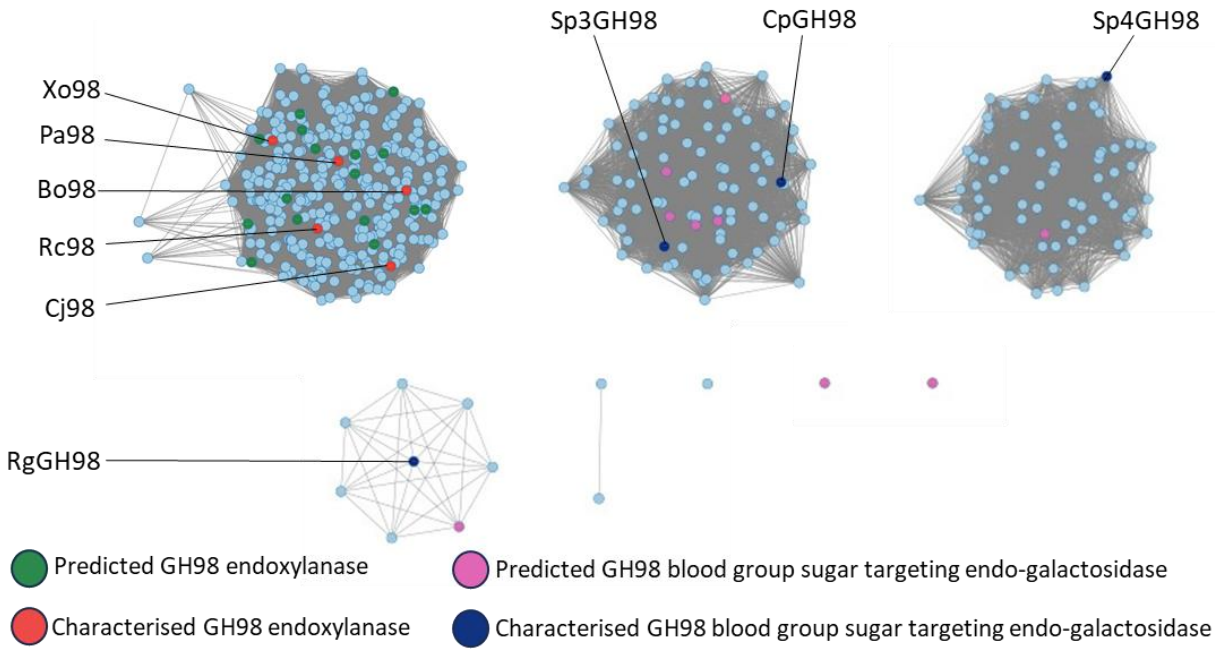


Figure 4-6: Sequence similarity network analysis of the GH98 family. 540 GH98 protein sequences from the CAZy database (www.cazy.org) (Cantarel *et al.*, 2009) were analysed by SSN. An alignment score of 120 was utilised. Each node represents one protein sequence, shown as a pale blue circle. Functionally characterised enzymes are coloured. GH98 endoxylanases to be characterised in this study: Bo98, Rc98, Cj98, Pa98 and Xo98, are shown in red. Predicted endoxylanase enzymes: representative sequences from CDhit analysis possessing a CBM35 domain are shown in green, and all fall in the same cluster as characterised endoxylanase enzymes. Characterized blood group targeting β-galactosidases, Sp3GH98 and Sp4GH98 (Higgins *et al.*, 2009), CpGH98 (Anderson *et al.*, 2005) and RgGH98 (Wu *et al.*, 2021), are coloured in dark blue and fall in separate clusters to Bo98. Pink coloured circles show representative sequences with predicted β-galactosidase activity.

Of note, Sp3GH98 and CpGH98, which display the same substrate specificity for Type A or B blood group antigens (**Figure 4-1**), are seen in the same cluster with SSN analysis, whereas Sp4GH98 and RgGH98, which possess different specificities, lie separately (**Figure 4-6**).

Although domain composition, phylogenetic and SSN analysis has led to the hypothesis that GH98 enzymes which cluster with Bo98 and possess a C-terminal CBM35 will show endoxylanase activity, further enzyme characterization studies are required to confirm this.

4.2.2 Investigating the activity of predicted GH98 endo-xylanase enzymes

To explore the activity of enzymes in the predicted endo-xylanase cluster of the GH98 phylogenetic tree, several other GH98 enzymes were cloned and recombinantly expressed (**Table 4-1**). Bo98, Cj98 and Rc98 gene sequences were amplified using primers described in **Supplementary Table 8-2** and cloned into expression vectors using relevant restriction enzymes. Pa98 and Xo98 encoding genes were synthesised in a pET29a vector by Twist Bioscience (www.twistbioscience.com).

Bo98, as well as Cj98, Pa98 and Rc98 expressed at high levels in a soluble form in an *E. coli* expression system (**Figure 4-7**). Unfortunately, Xo98 enzyme did not express under tested conditions, despite several attempts at optimising its expression.

Table 4-1: Cloning strategy and details of gene fragments amplified and analysed in this study.

Code name for recombinant protein	Locus Tag	Region cloned from DNA	Cloning vector	Cloning sites in vector	Theoretical molecular weight (kDa)	Extinction coefficient ($M^{-1}cm^{-1}$)
Bo98	BACOVA_03433	157-2844	pET21a	NheI/XhoI	101	189385
Bo98ΔDUF	BACOVA_03433	691-2844	pET21a	NheI/XhoI	81	160965
Rc98	RUM_20560	67-2773	pET21a	NheI/XhoI	86	173135
Cj98	CJA_3286	901-3049	pET21a	NheI/XhoI	80	151860
Pa98*	K333DRAFT_01908	65-3105	pET29b	NdeI/XhoI	114	297055
Xo98*	Ga0399197_01_992664_994847	1-2184	pET29b	NdeI/XhoI	79	146235
Bo03417ł	BACOVA_03417	113-1603	pET22b	EcoRI/XhoI	60	102260
Bo03425ł	BACOVA_03425	157-1801	pRSET A	BamHI/HindIII	68	113845
Bo03438ł	BACOVA_03438	51-2435	pET21a	BamHI/HindIII	91	154505
Bo03449ł	BACOVA_03449	91-2571	pET28b	EcoRI/XhoI	98	194970

Theoretical molecular weights (kDa) and extinction coefficients ($M^{-1}cm^{-1}$) were calculated using the Expasy Protparam tool at web.expasy.org/cgi-bin/protparam/protparam. Gene fragments marked ł were cloned in previous study by Rogowski *et al.* (2015) and those marked * were synthesised by Twist Bioscience, https://www.twistbioscience.com.

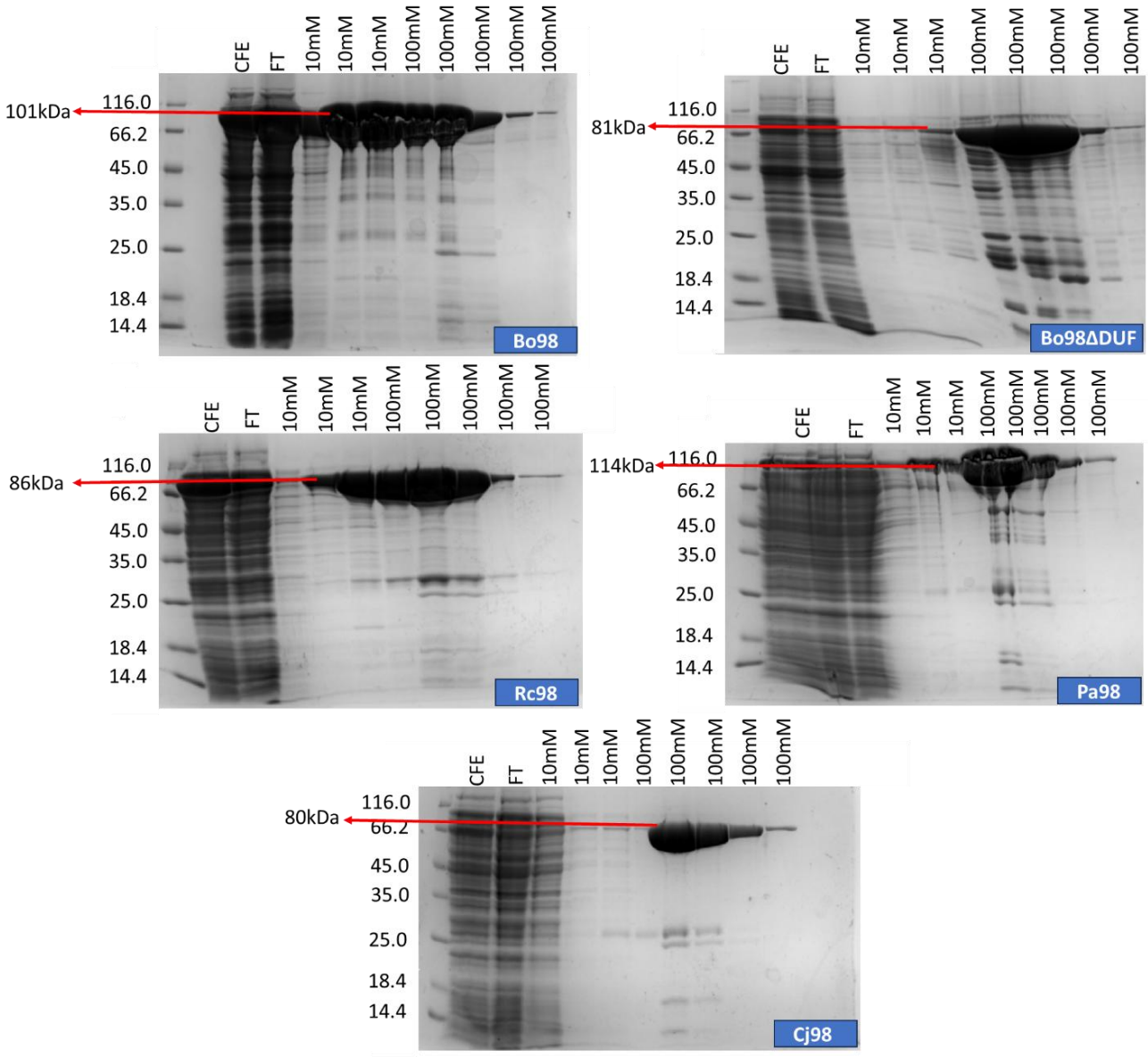


Figure 4-7: Expression and purification of recombinantly expressed GH98 proteins. GH98 enzymes were cloned and recombinantly expressed in *E. coli* Tuner cells His-tagged protein was subsequently purified and analysed by SDS PAGE with a 14.4 to 116kDa molecular weight ladder (labelled on the left). Cell free extract (CFE), flow through (FT) and 8 imidazole-eluted fractions were collected: 3 x 10 mM, followed by 5 x 100 mM. Bo98 and Bo98ΔDUF from *B. ovatus* ATCC8384, Rc98 from *Ruminococcus champanellensis* 18P13, Pa98 from *Segotella albensis* M384 (previously named *Prevotella albensis*) and Cj98 from *Cellvibrio japonicus* UEDA107 had calculated molecular weights of 101, 81, 86, 114 and 80 kDa respectively.

Analysis of enzyme activity by TLC demonstrated that all 4 successfully expressed GH98 enzymes (Bo98, Rc98, Pa98 and Cj98) possessed GAX specific endo-xylanase activity, producing a range of large oligos from corn xylan (CX), visualised as a smear

on the TLC, similar to the activity reported previously for the Bo enzyme (**Figure 4-8**). When assessing GH98 endoxylanase activity against GAX, TLC showed production of a range of different oligosaccharides, judged to be larger than 5 sugars. There always seems to be lots of material left on the origin, which may be uncleaved chains, or very large, cleaved xylan chains. No activity against the other xylans tested (birchwood glucuronoxylan or wheat arabinoxylan) was detected, as seen previously for Bo98.

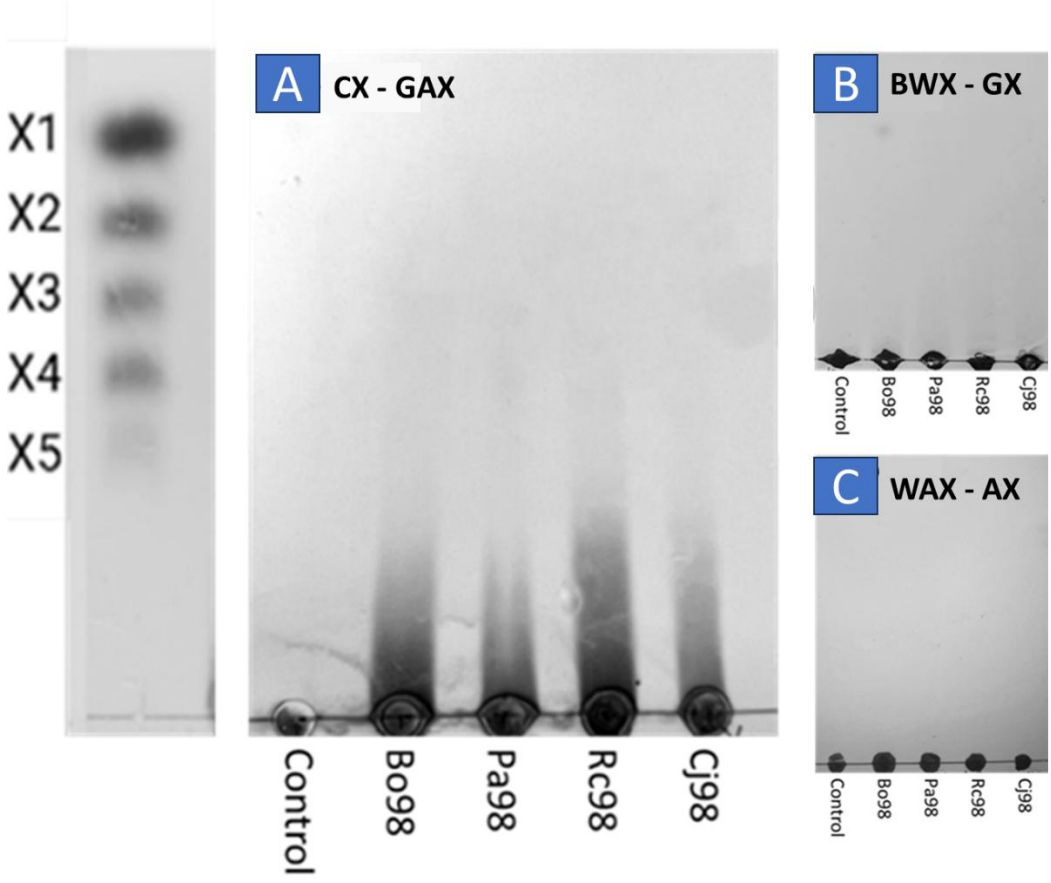


Figure 4-8: Activity of different GH98 enzymes against xylans. Bo98 and the 3 uncharacterised GH98 enzymes were tested for activity against complex GAX from corn (CX) (A) as well as simpler xylans from Birchwood (BWX) (B) and Wheat (WAX) (C). All enzymes possessed activity against CX, as can be seen from the smear of large products but lacked activity against the other xylans tested. Bo98 (*Bacteroides ovatus* ATCC8384), Rc98 from *Ruminococcus champanellensis* 18P13, Pa98 from *Segotella albensis* M384 and Cj98 from *Cellvibrio japonicus* UEDA107. All four enzymes lacked activity against simpler AXs (wheat arabinoxylan - WAX, rye arabinoxylan - RAX, oat spelt xylan - OSX) or GXs such as birchwood xylan – BWX. Linear xylo-oligosaccharide standards from X1-X5 are shown on the left.

To investigate the role of the *Bacteroides* specific N-terminal DUF4988 in Bo98, a construct of the enzyme lacking this domain was produced (Bo98ΔDUF). Initial studies demonstrated Bo98ΔDUF activity against CX (**Figure 4-9**), indicating that this domain is not essential for function, supporting the proposed role of this domain in spatial orientation of the enzyme in the OM utilisome rather than directly contributing to catalytic activity.

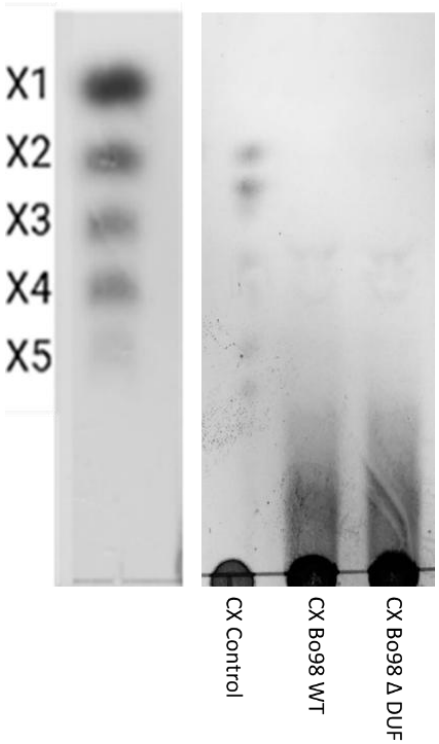


Figure 4-9: Activity of full length Bo98 and a construct lacking the N-terminal DUF against CX. Bo98WT and Bo98ΔDUF were assayed at against the complex GAX from corn (CX). Bo98ΔDUF showed activity comparative to WT at a single time point. Linear xylo-oligosaccharide standards from X1-X5 are shown on the left.

4.2.3 Application of reducing sugar assays to quantify GH98 endoxylanase activity

Following initial characterisation of GH98 activity via TLC, bicinchoninic acid (BCA) reducing sugar assays were trialled to assess their suitability as a quantitative assay for analysis GH98 endoxylanase activity against complex GAX (**Figure 4-10**). This would allow effective comparison of activity of Bo98, Rc98, Pa98 and Cj98, as well as

quantification of the impact of single point mutations on enzyme activity. Due to limitations in the amount of CX substrate available, we first validated the application of BCA assays for quantifying the endo- degradation of xylans using a BWX-GH10 endo-xylanase model. Recombinantly expressed GH10, Bacova_04390, from BoXylPUL-S, with known activity against BWX, was utilised for these studies (**Figure 4-10**). Five concentrations of Bacova_04390 were incubated with 10 mg mL^{-1} BWX for 1, 5, 15 and 20 minutes. After following the BCA assay protocol, the absorbance of samples at 560nm was measured. For BWX-Bacova_04390 reactions we found that BCA assays were successful for the quantification of enzyme activity over time (**Figure 4-10**). At an enzyme concentration of 1 μM, after just 1 minute a high absorbance was seen, with only a small increase at subsequent timepoints, indicating that the enzymatic reaction neared completion after just 1 minute. At a concentration of 0.1 μM the reaction neared completion at 60 minutes, whereas at lower concentrations the absorbance was a lot lower at all timepoints. When the same methodology was applied to CX degradation by Bo98, for a given enzyme concentration we did not see increases in absorbance over time. Furthermore, a very high background absorbance was seen for CX, likely due to a shorter average chain length in comparison to BWX meaning a higher concentration of reducing ends at the starting point. Furthermore, because Bo98 creates large oligos by breaking β1,4 linkages at specific points in the xylan chain, the increase in the number of reducing ends may be limited, even if the reaction goes to completion. Based on these data we deemed BCA assays to be ineffective to quantify Bo98 activity against CX and instead used semi-quantitative TLC analysis to assess GH98 activity, meaning that it was challenging to study enzyme kinetics.

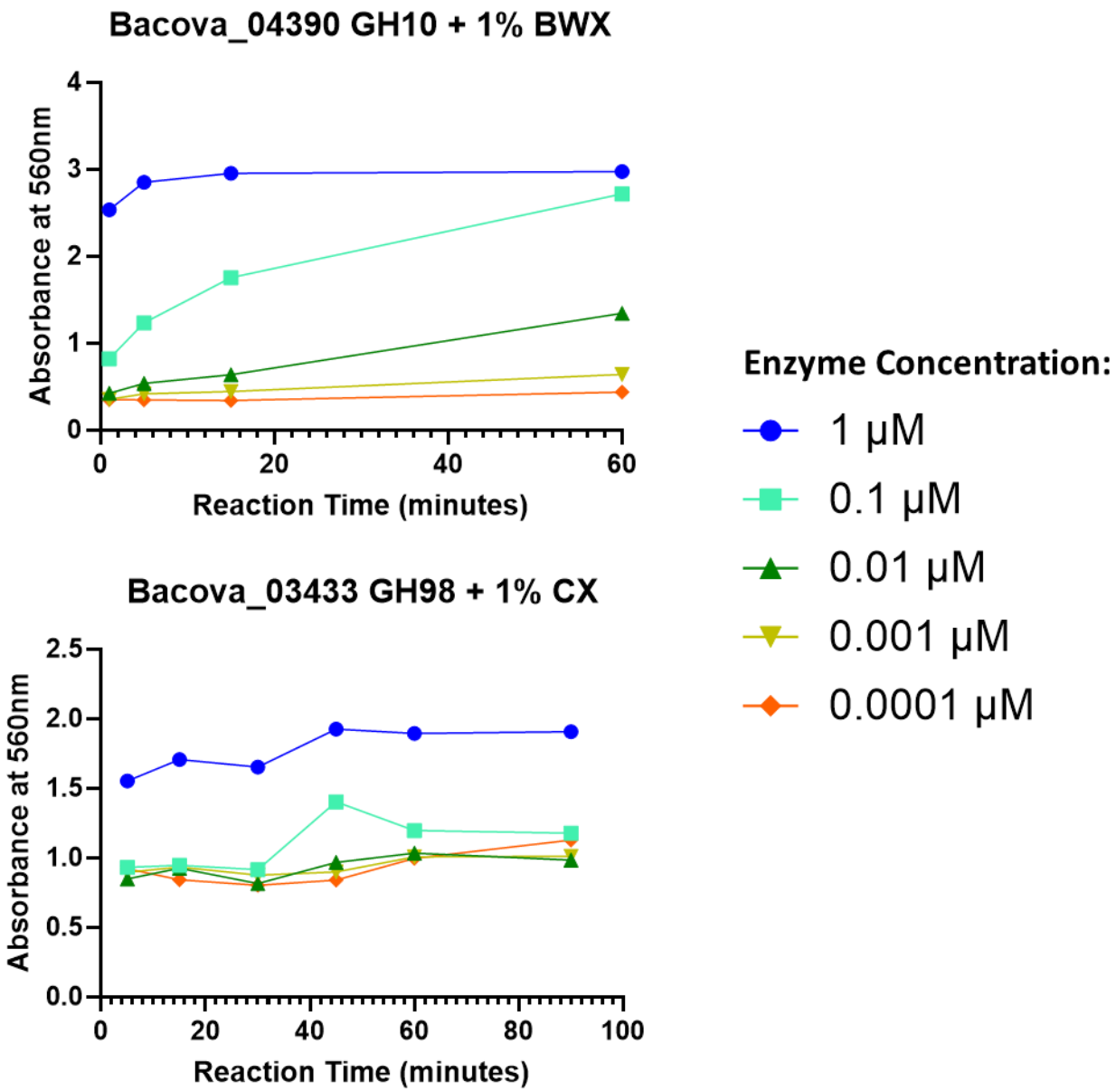


Figure 4-10 : Bicinchoninic acid (BCA) reducing sugar assays to study kinetics of GH98 endoxylanase activity. Preliminary studies showed that BCA assays were effective for quantification of Bacova_04390 GH10 activity against birchwood xylan (BWX), top panel. Reactions with a higher enzyme concentration resulted in a higher absorbance reading following the BCA assay and reached completion at an earlier timepoint. For Bo98 (Bacova_03433) reactions against corn xylan (CX) the impact of enzyme concentration was not as clear, bottom panel. Similarly, there was not a meaningful increase in absorbance over time for reactions at any given concentration.

4.2.4 Structural studies into Bo98 substrate specificity

X-ray crystallography structural studies were performed to investigate the structural differences basis for endo-β1,4-galactosidase and endo-β1,4-xylanase activities in the GH98 family, as well as to elucidate the critical specificity determinants of Bo98. Firstly, an inactive mutant of Bo98 was created by site directed mutagenesis, where the glutamate nucleophile residue at position 361 was mutated to an alanine. The position of this nucleophile was predicted based on alignment with acid nucleophile from the characterised blood group- active enzymes *Sp4*GH98, Glu^{158}, and *Sp3*GH98 Glu^{558}, as the location of the key catalytic residues is normally conserved within a GH family. Studies showed that Bo98 E361A lacked activity against CX (**See Figure 4-19A**). Five commercial screens were utilised for crystallisation of selenomethionine (SeMet) incorporated Bo98 WT and SeMet Bo98 E361A. As larger crystals formed using the E361A mutant protein, this was utilised for subsequent studies. Crystals of SeMet incorporated Bo98 E361A protein were optimised and grown using conditions of 0.2 M sodium thiocyanate and 20% w/v PEG3350. Spine-like crystals of approximately 450 µm in length formed after approximately 14 days at 20 ºC (**Figure 4-11**).

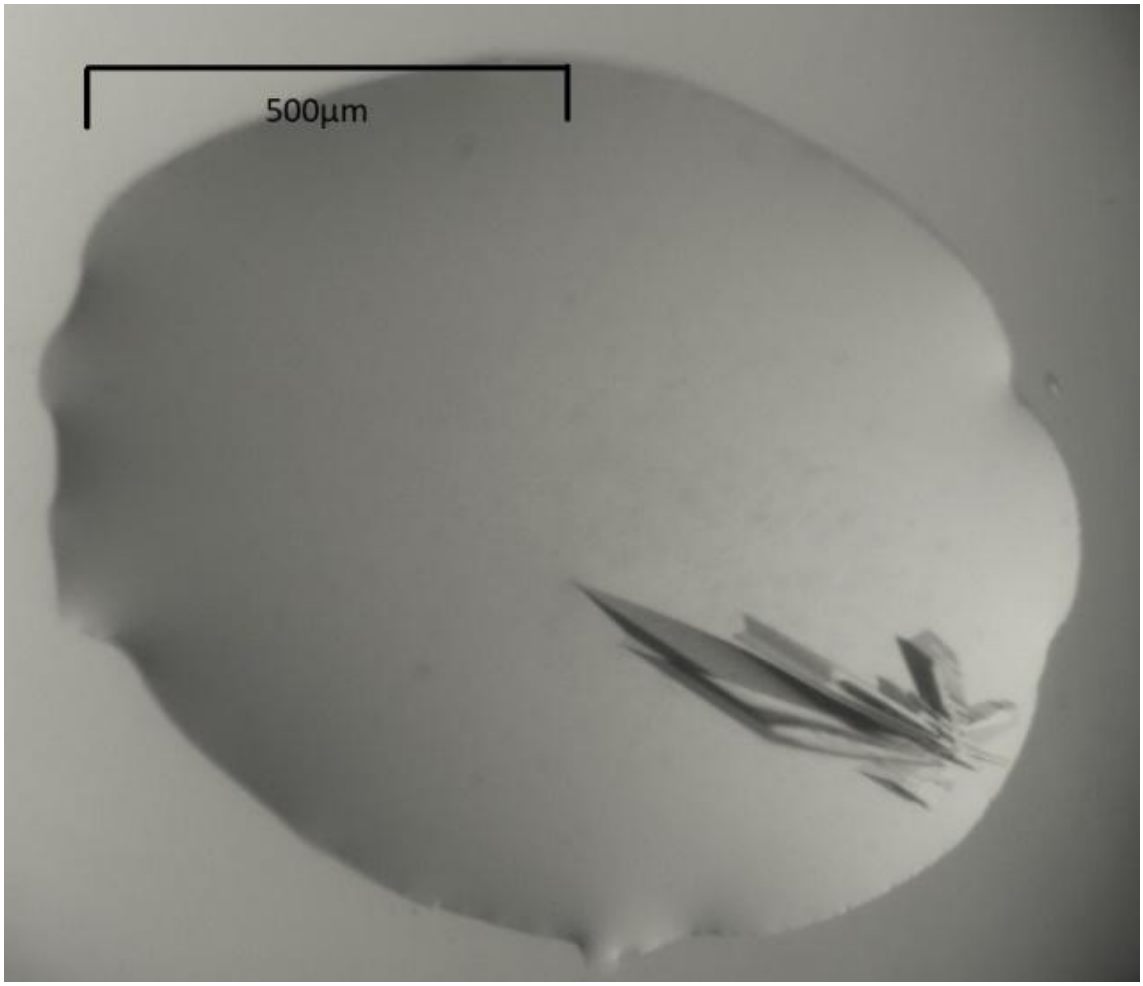


Figure 4-11: Crystals of SeMet Bo98 E361A. Crystals formed via the sitting drop vapour diffusion method under conditions of conditions of 0.2 M sodium thiocyanate and 20% w/v PEG3350.

Subsequently, the structure was solved by single-wavelength anomalous diffraction (SAD) to 2.1Å resolution (**Figure 4-12, Table 4-2**). The structure of Bo98 reveals overall similarity in structure of catalytic domains with that of Sp4GH98 (**Figure 16A**) as previously described by Higgins *et al.* (2009): a classical TIM $(\alpha/\beta)_8$-barrel followed by a β sandwich domain consisting of 11 strands. A long, linear unstructured linker sequence then connects the catalytic domain with the C-terminal CBM. The N-terminal DUF consists of a central α-helix, flanked by 2 small β-sheets.

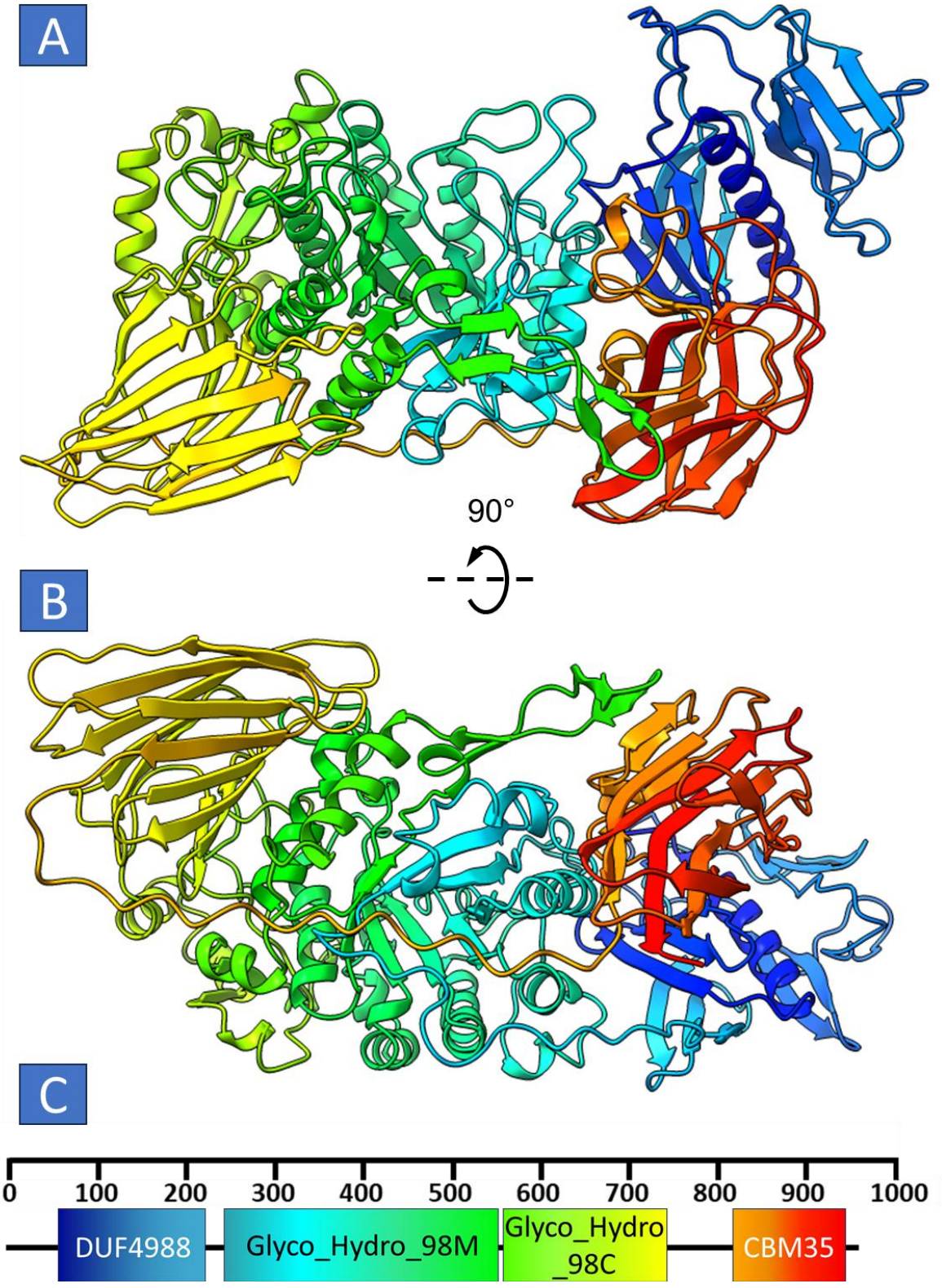


Figure 4-12: Schematic representation of the overall structure of Bo98 GAX-specific endoxylanase. (A + B) Cartoon depicting the overall structure of Bo98 colour ramped from N terminus (blue) to C terminus (red) showing the relative position of the N terminal domain of unknown function (DUF4988) and the C terminus CBM35. A long linear, unstructured linker region is seen between the C terminus of the catalytic domain, and the start of the CBM35, which appears in dark yellow. (C) Schematic demonstrating the domain composition of Bo98.

Table 4-2: Bo98 crystal structure data collection, statistics and refinement.

Data Statistics	**SeMet E361A**	**Native E361A**
Beamline	I03	I03
Date	07/07/21	15/10/21
Wavelength (Å)	0.898	0.898
Resolution (Å)	2.10	2.80-53.02
Space Group	C 1 2 1	C 1 2 1
Unit-cell parameters:		
a (Å)	183.91	186.05
b (Å)	62.45	63.48
c (Å)	178.71	180.90
α, β, γ (°)	90.00. 118.06, 90.00	90.00, 118.44, 90.00
Unit-cell volume ($Å^3$)		
Solvent content (%)		
No. of measured reflections		46244
No. of independent reflections		2323
Completeness (%)	99.9	99.90
Redundancy		
Rmerge (%)		0.24
<I>/<σ(I)>		6.30
Refinement Statistics		
Rwork (%)	0.2068	0.191
Rfree# (%)	0.2443	0.285
No. of non-H atoms:		
No. of protein atoms	14191	13735
No. of ion atoms	4	
No. of ligand atoms	-	136
R.m.s. deviation from ideal values:		
Bond length (Å)	0.0077	0.0052
Bond Angle (°)	1.3980	1.2330
Average B factor ($Å^2$):		
Protein	29.0	44.4
Ligand	-	91.1
Ions	36.1	
Ramachandran plot+, residues in allowed and most favoured regions (%)	96.07	94.5

4.2.6 Structure of the BoGH98-arabinoxylo-oligosaccharide complex

To investigate the unique specificity of the GH98 endoxylanase in more detail, we aimed to solve the crystal structure of a substrate-bound complex. He we show the first structure of a GH98 enzyme in complex with a CX derived oligosaccharide substrate. Modelling of the xylooligosaccharide substrate was performed by Alan Cartmell. To obtain a small enough oligosaccharide substrate suitable for crystal soaks, CX was first partially digested with Bo98. The resulting oligosaccharide products were separated by size exclusion chromatography using Bio-Rad Bio-Gel P2 gel, allowing isolation of fractions of different sized oligosaccharides (**Figure 4-13A**). Fractions of a similar size were selected to ensure that a relatively homogeneous mixture was used for crystal soaks. Fractions C7 to D6 were chosen for use in crystal soaks, as some degradation of CX can be seen via TLC, however the reaction had not gone to completion, as shown by the lack of smaller products on the TLC. These fractions were combined, freeze-dried, resuspended and concentrated to 350 mg mL^{-1}, to produce an oligosaccharide mixture for use in crystal soaks, referred to as fraction C. This fraction was subsequently incubated with Bo98 to show that it was still a substrate for the enzyme (**Figure 4-13B**).

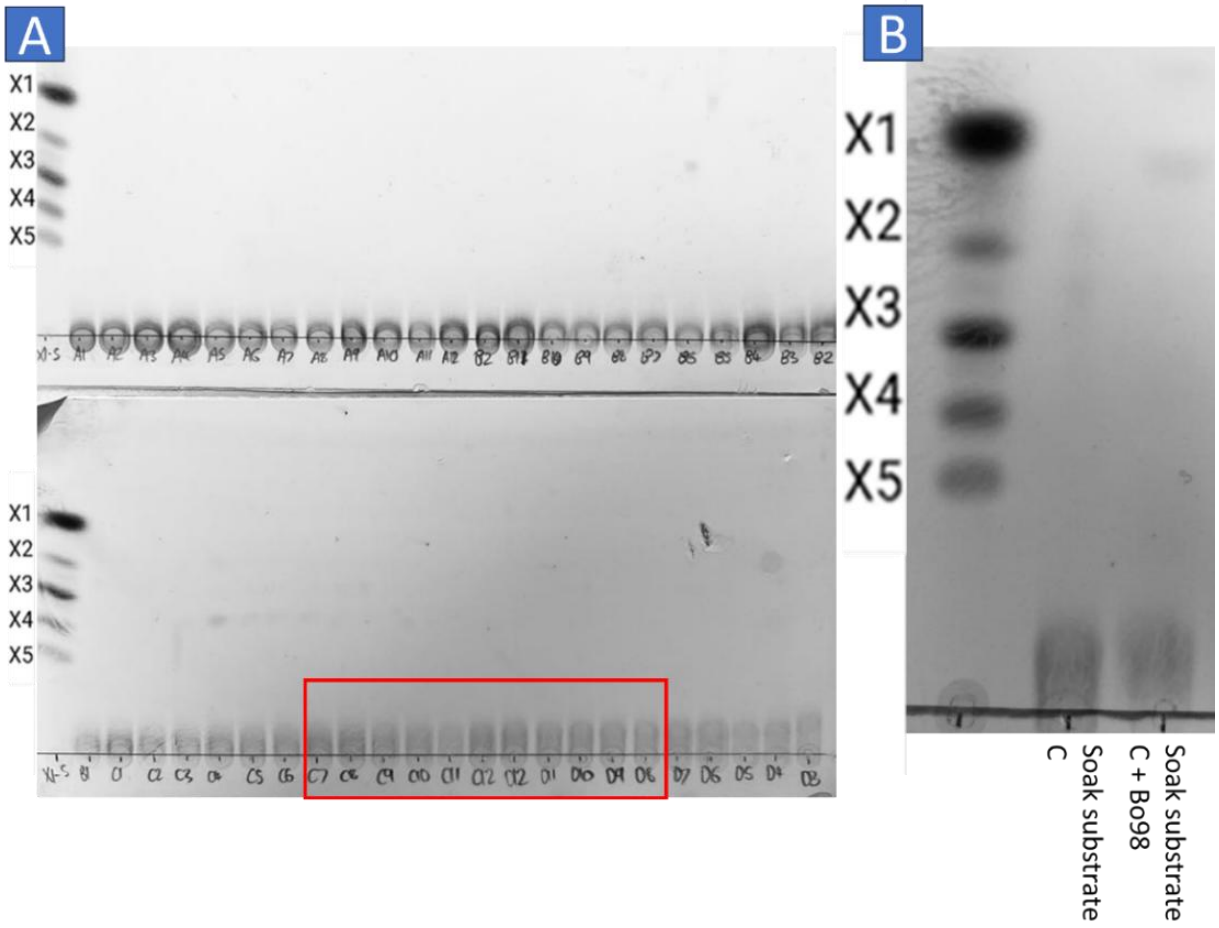


Figure 4-13: Isolation of CX derived oligosaccharide fractions for crystal soaks. (A) Bo98 partially-digested CX was separated by size exclusion chromatography to produce fractions of varying sized oligosaccharides. Fractions C7 to D6 were combined to provide a substrate used for soaking with crystals of catalytically dead Bo98 mutant E361A, referred to as soak substrate (B). Subsequent digestion with Bo98 demonstrated that this was still a substrate for the enzyme, as less sugar could be seen on the origin following Bo98 digestion.

Crystals of native Bo98 E361A were formed under conditions of 0.2M potassium thiocyanate, 0.1M Bis Tris propane and 20% w/v PEG3350, based on well H4 of the JCSG+ screen. These crystals were then soaked with the fraction C oligosaccharides for 16 hours prior to freezing and shooting. The initial selenomethionine structure was subsequently used to determine the native ligand bound structure of the E361A inactive mutant of Bo98 to 2.8Å resolution (**Table 4-2**). The relatively poor resolution of this model presented challenges, particularly when modelling the ligand. The

resulting density revealed a highly decorated xylo-oligosaccharide substrate in the active site (**Figure 4-14**).

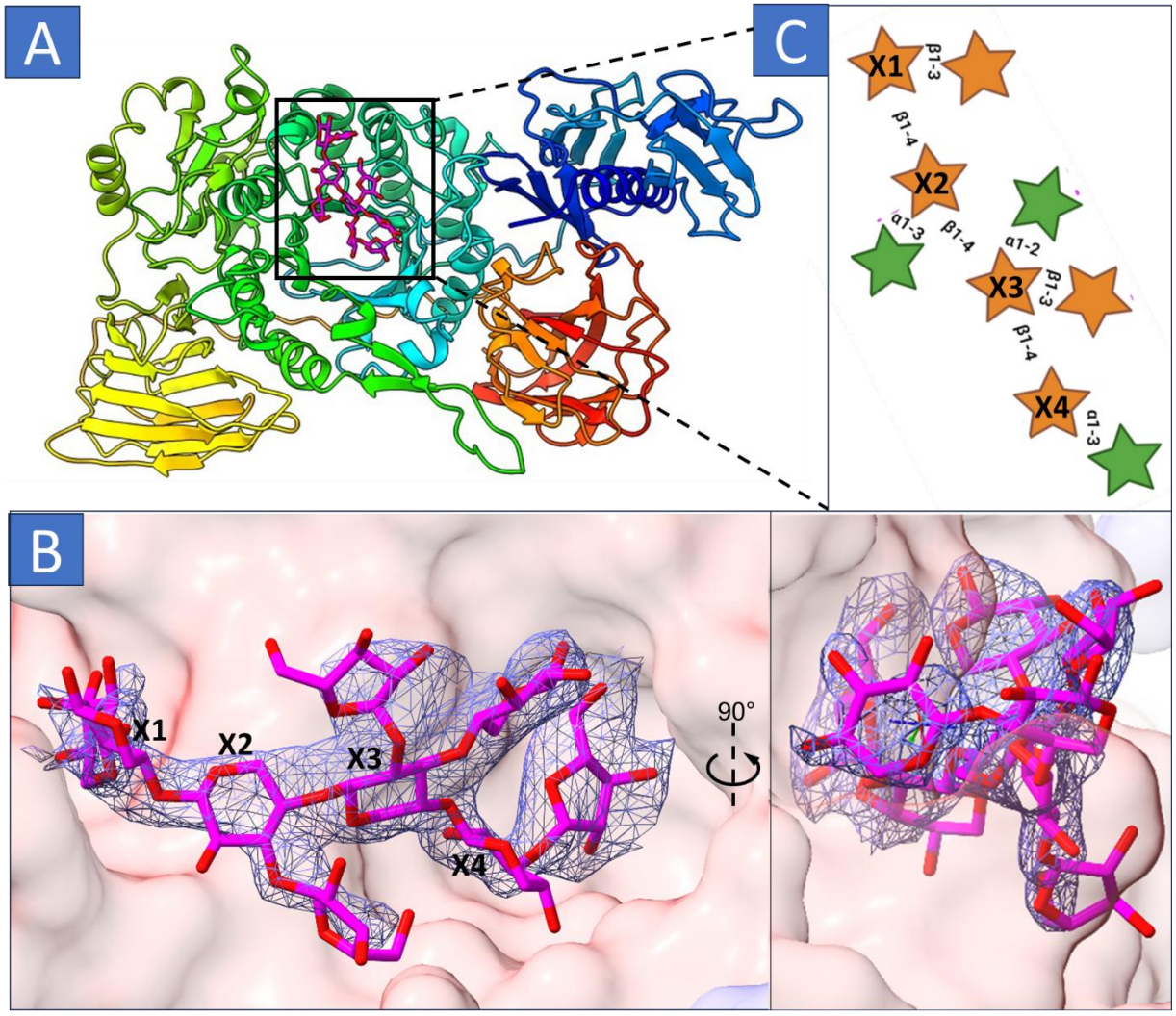


Figure 4-14: Schematic representation of the overall structure of an inactive mutant of Bo98 endoxylanase with a highly decorated arabinoxylo-oligosaccharide substrate bound in the active site. (A) Cartoon depicting the overall structure of Bo98 colour ramped from N- (blue) to C-terminus (red), bound to a nonasaccharide ligand (magenta). (B) Close up of the Bo98 active site. The 2Fo − Fc density for deca glycine is shown as a blue mesh at 1.5σ. Orientation differs from that in parts A and C. The four xylose residues forming the backbone of the xylooligosaccharide chain are labelled X1 to X4 from the non-reducing towards the reducing end. (C) Schematic representation of the xylo-oligosaccharide ligand modelled from the electron density map. This ligand consists of a highly decorated β1-4 linked xylotetrasaccharide backbone, shown as X1 to X4 from the non-reducing end to the reducing end. Based on ligand density, the β1-3 xylose decoration of X3 could potentially have been modelled as an α1-3 arabinose.

The Bo98-bound ligand modelled into the electron density was a nonasaccharide, consisting of a highly decorated β1-4 linked xylotetraose backbone (**Figure 4-14B**). Based on the positioning of this xylooligosaccharide in the active site, we predict that the sugar labelled as X4 sites in the -1 enzyme site, and X3, X2 and X1 in the +1, +2 and +3 sites respectively. The xylose at the non-reducing end, X1, was decorated with a single β1-3 linked D-xylose. The next xylose in the chain, X2 possessed an α1-3 L-arabinose. The electron density at X3 showed a double decoration, modelled as an α1-2 linked L-arabinose and a β1-3 D-xylose, however the arabinose decoration could have been built as an α1-3 xylose, an issue with the low resolution of the model (**Figure 4-14C**). The double arabinose-xylose decoration seen at the X3 sugar matches prediction by Rogowski *et al.* (Rogowski *et al.*, 2015) that this feature may be an important recognition moiety of Bo98. The reducing-end xylose moiety, X4, is decorated with a single α1-3 arabinose. Based on the density shown, there is potential for the X4 sugar may be distorted (**Figure 4-14C**), but at a resolution of 2.8 Å this cannot be ascertained. The xylooligosaccharide structure displayed in here was built and refined by Dr. Alan Cartmell, University of York.

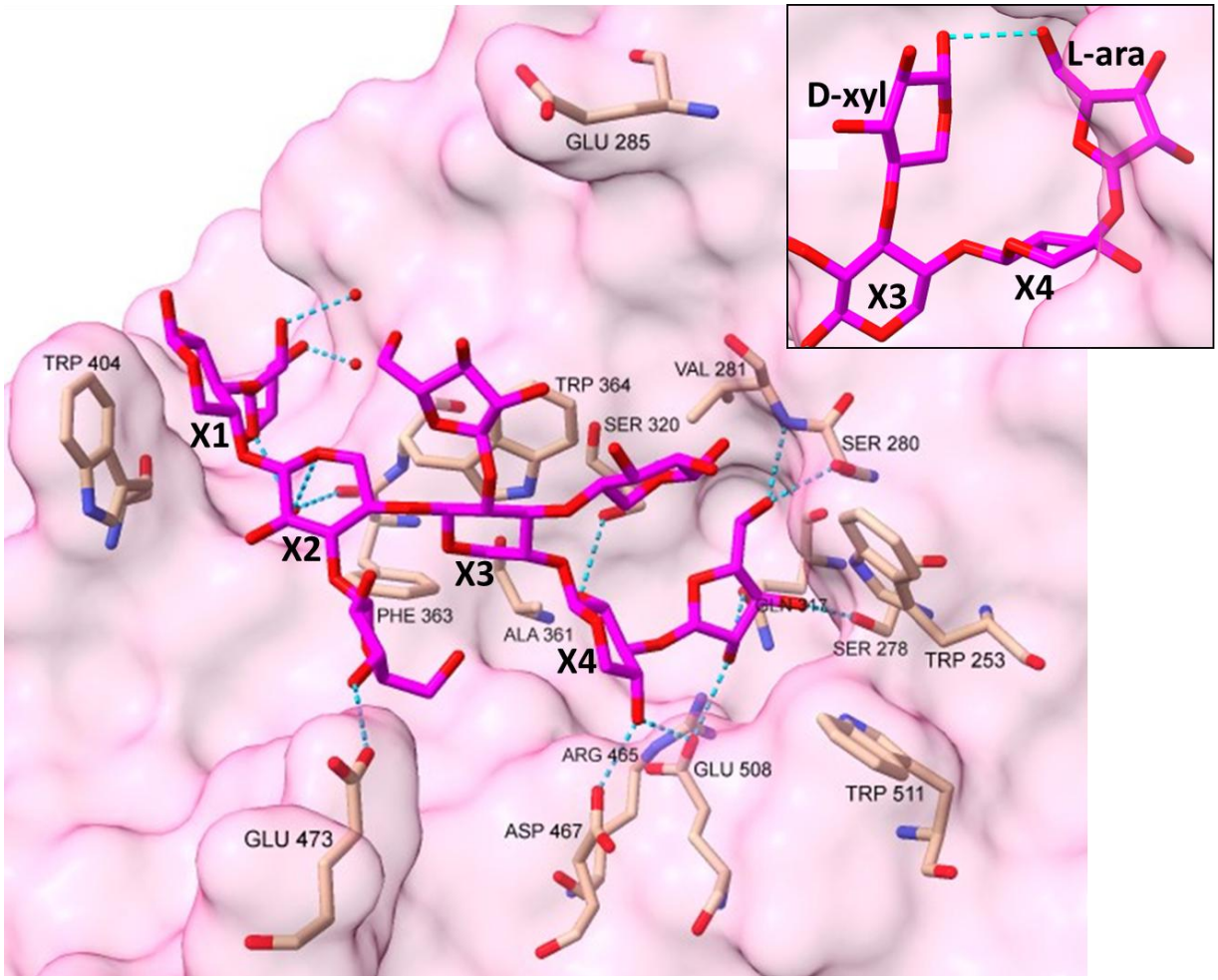


Figure 4-15: Bo98 active site bound with xylo-oligosaccharide ligand. Active site residue side chains are shown in pale pink stick format. Xylo-oligosaccharide ligand is shown in magenta, with the backbone xylose backbone residues labelled X1-X4 from the non-reducing, towards reducing end. Water molecules are shown as red dots Predicted hydrogen bonds between the ligand and enzyme structure or water molecules are shown as blue dotted lines. Insert shows hydrogen bond between O4 of the xylose decoration of the X3 sugar and O5 of the arabinose decorating the X4 sugar.

Despite modelling the expected double arabinose-xylose decorations of the X3 sugar of the xylo-oligosaccharide, based on predicted interactions between ligand and active site residues it seems unlikely that this double decoration is the sole specificity determinant of the enzyme (**Figure 4-15**). Positioning of the ligand in the active site suggests that the X3 xylose sugar forms a pi stacking interaction with Trp364 of Bo98, however, the model suggests that no hydrogen bonds are likely to form between either the arabinose or xylose substitution, and any active site residues of Bo98 (**Figure 4-**

15). Based on this, it is unlikely that the enzyme could dictate specificity for an arabinose or xylose at either position, and hence it seems unlikely that this site is the sole specificity determinant of the enzyme.

The X1 sugar of the modelled xylo-oligosaccharide forms a pi stacking interaction with the tryptophan at position 404 of Bo98 (**Figure 4-15**). This interaction may be important in the recognition and orientation of xylo-oligosaccharides with a degree of polymerisation greater than 3 in the active site. The xylose decoration of the X1 sugar appears to form hydrogen bonds with water molecules (**Figure 4-15**). The X2 sugar possesses a single arabinose decoration. Hydrogen bonding between this arabinose and glutamate 473 indicates that recognition of the sugar decoration in this position may play a role in substrate specificity (**Figure 4-15**). Interestingly, the arabinose decoration of the X4 xylose sits in a tight pocket and forms hydrogen bonds with many active site residues: Val281, Ser280, Ser278 and Glu508. The spatial constriction of this arabinose decoration suggests that it is likely important for substrate recognition, and ultimately enzyme function. The reducing end xylose (X4) forms a hydrogen bond with serine 320 (**Figure 4-15**), however, based on the modelled density the oxygen at C4 at the non-reducing end of the xylo-oligosaccharide seems to face into the body of the enzyme, forming hydrogen bonds with Asp467 and Glu508. If this was the true positioning of the X4 sugar in the -1 enzyme site, this would dictate a terminal sugar, such that no subsequent backbone xylose residues could be recognised, indicating an exo-acting enzyme. However, as demonstrated previously, Bo98 is endo-acting and as such would need to bind to other sugars in the xylan backbone, in the negative subsites past -1. Inspection of the density at X4 reveals that the xylose could be modelled as a distorted sugar, rather than the ground state shown here. In this distorted form the O4 points towards the two opposing tryptophan residues at positions

511 and 253, which appear to form a channel for recognition of a longer chain-substrate (**Figure 4-15**). Based on the positioning of these residues we propose that these tryptophan residues may play an important role in recognition of a xylan chain past the -1 subsite.

Of particular interest when examining the ligand modelled here is the hydrogen bond between the O3 xylose decoration of the X3 sugar, in the predicted +1 site, and the O2 arabinose decoration of the X4 sugar, the predicted -1 site (**Figure 4-15 Insert**). This interaction may be important for recognition by the Bo98 enzyme and would explain lack of activity against AX, as AXs possess no xylose side chains and a hydrogen bond such as that shown here would not form between two arabinose residues.

4.2.7 Structural comparisons between Bo98 and Sp4GH98 β-galactosidase

Comparison of the Bo98-xylo-oligosaccharide structure with Sp4GH98, a blood-group sugar active β-galactosidase GH98 from *Streptococcus pneumoniae* TIGR4 in complex with Lewis Y tetrasaccharide (Higgins *et al.*, 2009) (PDB 2WMG) demonstrated, as expected, conservation of the GH98 catalytic domains, including the central TIM barrel (**Figure 4-16A**), with an RMSD of aligned regions of 1.16Å. Structural alignment demonstrates the lack of C-terminal CBM35 and N-terminal DUF in Sp4GH98. Ligand-bound structures demonstrate similarities in the position and orientation of the active site, however, as predicted based on the differences in substrates, there is divergence in the overall active site conformation. Whilst Sp4GH98 active site forms a shallow pocket, that of Bo98 appears elongated, with exposed aromatic residues positioned for stacking interactions with longer xylan chains (**Figure 4-16B + C**). Where the pair of tryptophan residues, Trp511 and Trp253 in Bo98 appear

to form an elongated channel for interactions with a xylan chain, (**Figure 5-15**), this position in Sp4GH98 appears to be covered by a loop. This demonstrates how Bo98 can possess endo- activity against a large polymeric substrate like corn xylan, whereas the Sp4GH98 recognises the non-reducing ends of Lewis Y tetra- or pentasaccharide glycans.

Furthermore, where the surface of Bo98 and Sp4GH98 active sites are coloured by hydrophobicity (**Figure 4-16B + C**), a clear difference in the positioning and spread of hydrophobic side chains, more likely to interact with sugar moieties, can be seen. In Sp4GH98, hydrophobicity is greatest in and around the shallow pocket of the active site, whereas in Bo98, hydrophobicity is shown along the length of the active site, from Trp404, which forms a pi stack with the xylose backbone residue labelled X1 in this study, through to the pair of tryptophan residues forming a putative binding site for sugar X5 (**Figure 4-15**).

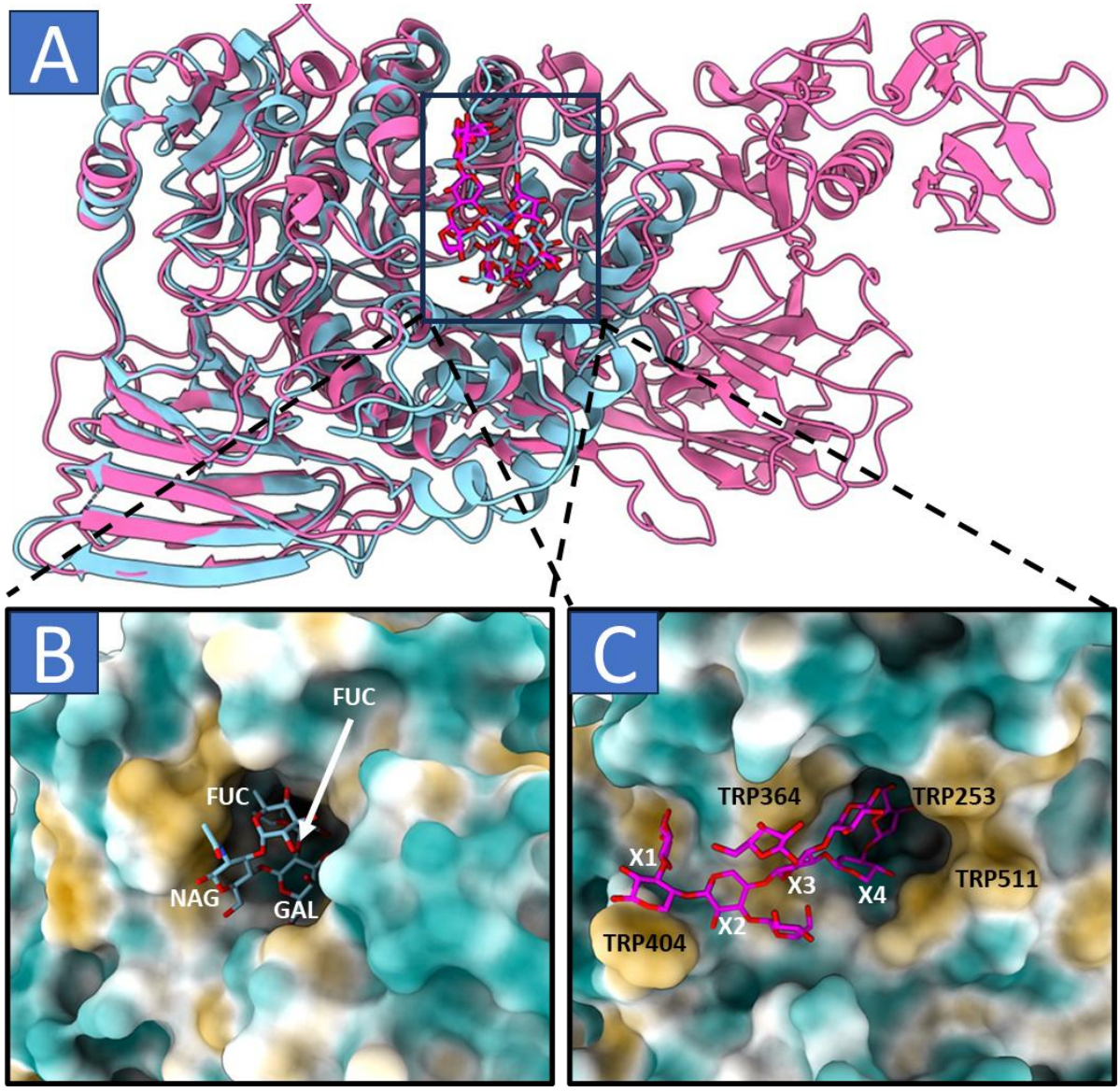


Figure 4-16: Structural comparison of Bo98 endoxylanase with Sp4GH98 β-galactosidase. (A) Overlay of crystal structures of xylo-oligosaccharide bound Bo98 (pink) with Lewis Y tetrasaccharide bound Sp4GH98 (blue) (Higgins *et al.*, 2009), demonstrating overall conservation of catalytic domain morphology within the GH98 family, with an RMSD of 1.16. Active site morphology of Sp4GH98 (B) and Bo98 (C) are shown, demonstrating surface hydrophobicity. Surface regions of higher hydrophobicity are shown in gold, demonstrating the elongated binding site of Bo98, largely formed by 4 tryptophan residues, allowing recognition of long xylan chains, in comparison to the pocket of Sp4GH98. Xylo-oligosaccharide backbone xylose residues are labelled X1-X4 from the non-reducing end. FUC = fucose, NAG = N-Acetylglucosamine, GAL = galactose.

Overall, the position and orientation of the key residues, defined by Higgins *et al.* (Higgins *et al.*, 2009; Rogowski *et al.*, 2015), forming enzyme active site in Sp4GH98, are conserved in Bo98: Asp251/467, Glu301/508, Lys220/436, Glu158/361 and

Trp161/364 (labelled Sp498/Bo98) (**Figure 4-17A + B**) (Higgins *et al.*, 2009). As expected in members of the same GH family, as well as the general catalytic acid, Glu158/361, this includes the candidate general base residues Glu301/508 and Asp251/467 (Higgins *et al.*, 2009; Anderson *et al.*, 2005), demonstrating that like other members of the GH98 family, Bo98 possesses an inverting catalytic mechanism. The conserved tryptophan residue, Trp161 in Sp4GH98, forms a stacking interaction with the pyranose ring of the GlcNAc residue of the Lewis Y tetrasaccharide, whereas Trp364 from Bo98, with similar orientation, forms a stacking interaction with X3, the third xylose from the non-reducing end of the xylo-oligosaccharide chain (**Figure 4-17A + B**).

This conservation of the key catalytic apparatus of the two enzymes is critical for cleavage of β1-4 glycosidic linkages by members of the same family, however differences in surrounding active site residues and overall morphology will contributes to specificity of substrate recognition for each member of the family.

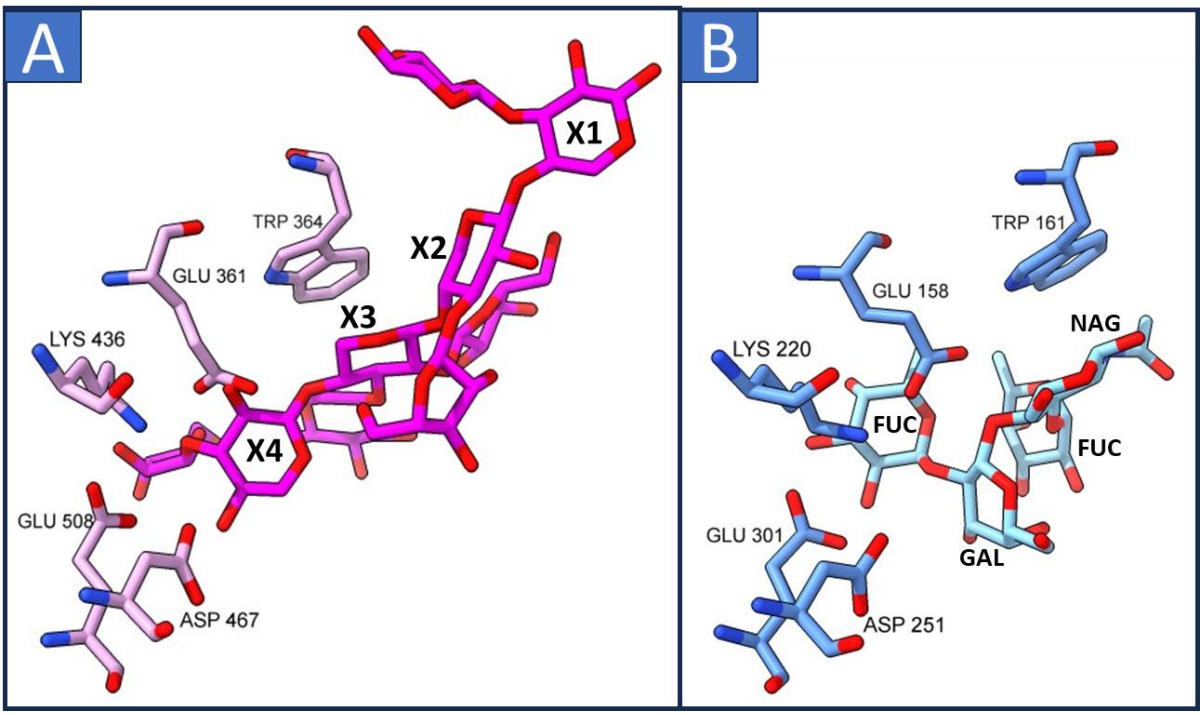


Figure 4-17: Conservation of active site residues in Bo98 endoxylanase and Sp4GH98 galactosidase. (A) Xylo-oligosaccharide bound to Bo98, labelled X1-X4 from non-reducing end. Relative position of the catalytic acid GLU361 in this Bo98 E361A structure was modelled using Alphafold2. (B) Lewis Y tetrasaccharide- bound Sp4GH98. Position and residue of Sp4GH98 key catalytic apparatus, as determined by Higgins *et al.* (Higgins *et al.*, 2009) is fully conserved in Bo98. FUC = fucose, NAG = N-Acetylglucosamine, GAL = galactose.

When examining the relative configuration of xylo-oligosaccharide bound to Bo98 with that of Lewis Y tetrasaccharide bound to Sp4GH98, there is similarity in the position and orientation of sugars in the enzyme active sites (**Figure 4-18A**). The pyranose ring of the GlcNAc residue of Lewis Y sits in a similar configuration to that of the X3 residue of the xylo-oligosaccharide chain. This is to be expected due to stacking interactions with the conserved tryptophan residue Trp161/364 (**Figure 4-18A**). Furthermore, the O3 linked xylose decoration at X3 of the xylo-oligosaccharide overlays with the fucose of Lewis Y, and the O2 linked arabinose decoration lies in the same orientation as the acetyl group of the GlcNAc (**Figure 4-18A**). Based on this conservation, and the previously defined enzyme subsites of Sp4GH98, we could designate the position of X3 as the +1 enzyme site, such that the O3 xylose lies in +1',

and O2 arabinose in +1" (**Figure 4-18B**). Conservation in the -1 site is not as stringent, although the position of the Lewis Y galactose is similar to that of X4, the reducing end xylose residue seen in the xylo-oligosaccharide chain. Divergence is greater in the -1' site, with the fucose residue of Lewis Y sitting in a different orientation to the O3 arabinose decoration (**Figure 4-18A**). Furthermore, we hypothesize that the 2 tryptophan residues 511 and 253 form the -2 subsite, although this site is unoccupied in the substrate complex solved here.

In both the endoxylanase and galactosidase structure, the +1 and -1 sugars possess decoration. As mentioned previously, the interaction between the arabinose decoration of X4, now known to be in the -1' site, and the xylose decoration of X3 at the +1' site suggests that both these residues may be important for GH98 substrate recognition, and conservation of this dual decoration in the +1' and -1' fucose residues of Lewis Y pentasaccharide supports this hypothesis. Furthermore, it is not common for enzymes to bind to substrates with decoration of +1 and -1 subsites, further highlighting the potential importance of this for GH98 enzymes.

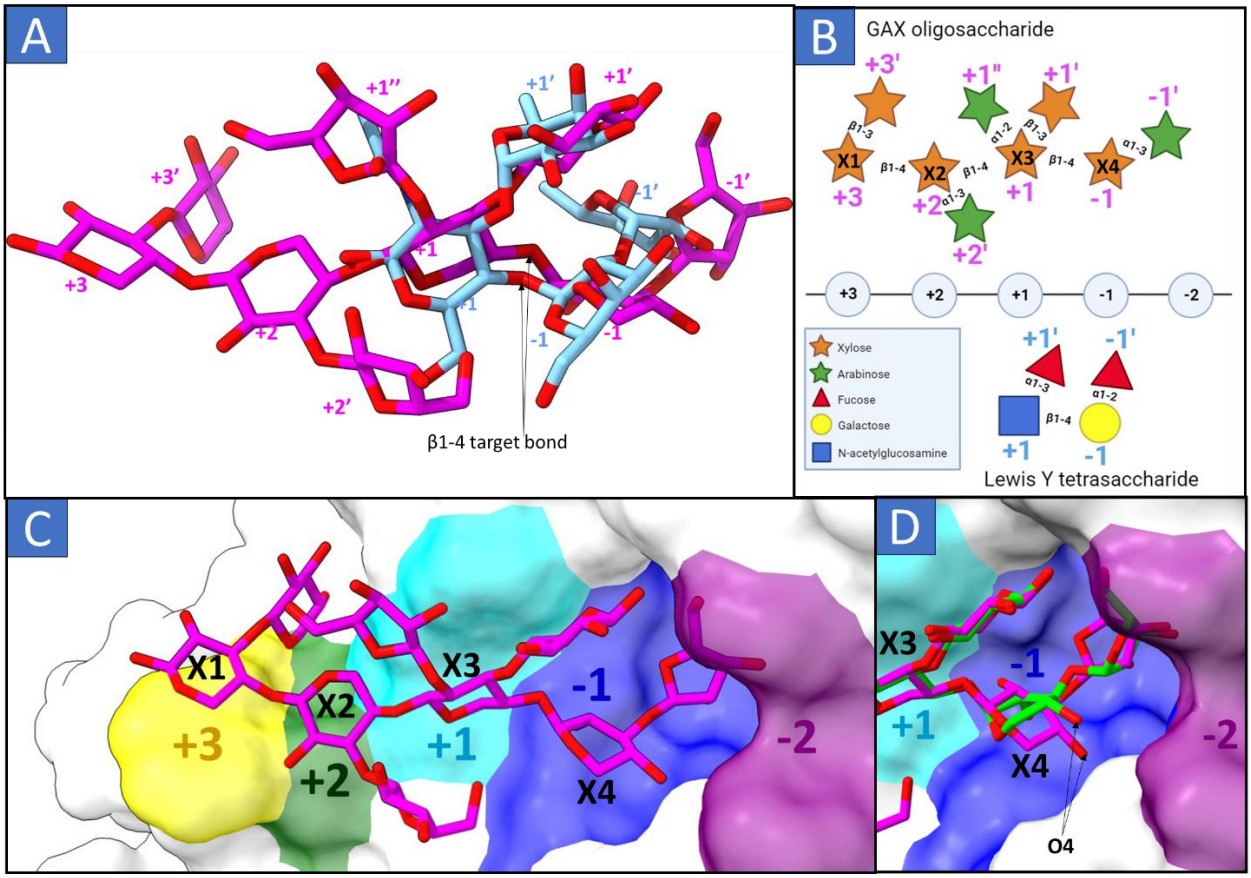


Figure 4-18: Bo98 active site subsites. (A) Overlay of Lewis Y tetrasaccharide (pale blue) and xylo-oligosaccharide (magenta) ligands bound to Sp4GH98 and Bo98 respectively. Overlay demonstrates some conservation in the position and orientation of the bound sugars. The positions of the +1 sugar and the predicted β1-4 glycosidic bond cleaved by the enzyme are highly conserved. (B) Schematic demonstrating the positioning of sugars within the enzyme subsites of Bo98 (top) and Sp4GH98 (bottom). (C) Positioning of enzyme subsites of Bo98 with xylo-oligosaccharide bound. (D) Shift in position of the X4 xylose residue is modelled as a distorted sugar, with the O4 pointing towards the cleft forming the predicted -2 subsite. The xylan chain with distorted X4 sugar is shown in green,

Overall, the position and orientation of the β1-4 target bond are similar between the two substrates (**Figure 4-18A**), indicating that the predicted enzyme subsites for Bo98 are correct, and the β1-4 bond between X3 and X4 is the true site of enzyme cleavage. However, if the xylo-oligosaccharide seen in crystal soaks were a true substrate for Bo98, then we would expect the enzyme would possess exo-type activity, resulting in the release of an α1-3 linked xylose-arabinose disaccharide. We know from TLC and HPAEC-PAD analysis (Rogowski *et al.*, 2015) that Bo98 solely demonstrates endo-

activity resulting in the production of a range of oligosaccharides. This indicates that the ligand shown here must be an artefact of crystallography, or a result of low occupation in the -2 and subsequent sites resulting in insufficient density. Alternatively, the E361A catalytic mutant, with a smaller alanine residue in the position of the glutamate, may have had some impact on the spatial arrangement of ligands within the active site. As discussed previously, from the Bo98 structure it seems logical that the -2 subsite would be formed by Trp253 and Trp511 (**Figure 4-18C**), however based on modelling of the X4 sugar in the -1 subsite in ground state, it is not apparent how the β1-4 linked chain could continue. This suggests that the recognition of larger polymeric xylan chains that potentially bind at the –2 subsite and beyond would require somewhat different configuration at the –1 site than is observed here. This is somewhat solved by modelling the X4 xylose residue as a distorted sugar (**Figure 4-18C**), which as discussed previously, can be fit within the density observed at this site. If this X4 sugar was in this position, the O4 faces the direction of the predicted -2 subsite, which would allow continuation of the xylan chain between the two tryptophan residues. Further crystal structures at a better resolution would be required to fully elucidate this.

Overall, similarities in enzyme subsites and positioning and orientation of the target bond demonstrates that despite the differences in sugar composition between xylan and blood group structures, the overall configuration of the glycans is similar such that members of the same family can exhibit activity against these divergent substrates.

4.2.8 Investigating the relative importance of residues in the Bo98 active site for enzyme activity

In order to probe the importance of residues around the active site for catalytic activity, site directed mutagenesis (SDM) was utilised to create 15 different single point

mutations to alanine. Mutant proteins were recombinantly expressed and assayed against corn xylan. Relative activity of mutants in comparison to Bo98 was assessed via TLC (**Figure 4-19A**).

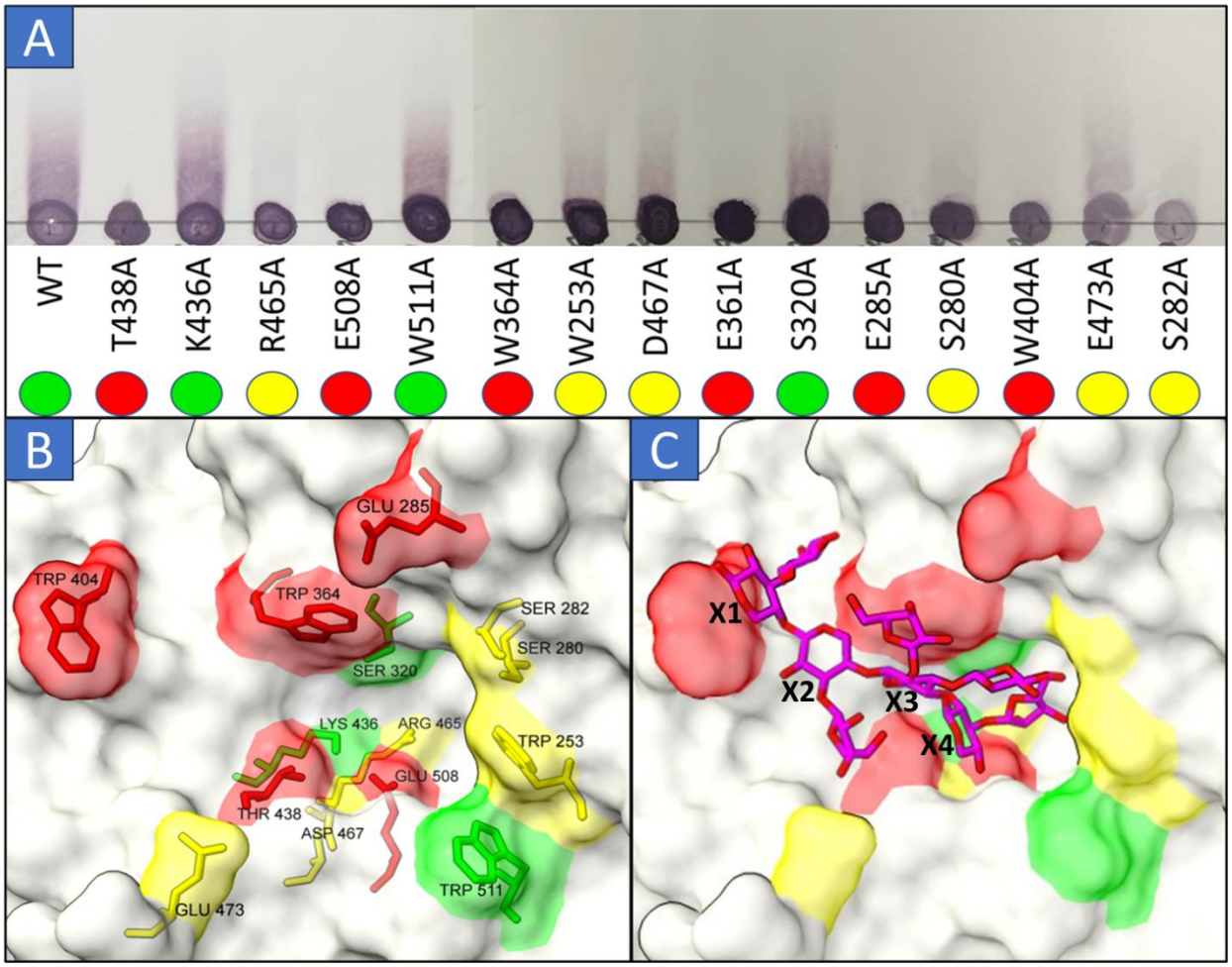


Figure 4-19: Alanine scanning mutagenesis of Bo98 active site. (A) Activity of mutants against CX, in comparison to WT Bo98, was assessed via TLC. 1 µM protein was incubated with 1% CX for 1 hour at 37 °C. (B) Impact of mutation on activity was mapped onto Bo98 active site surface structure. Residues whose mutation knocked out Bo98 activity are shown in red, those in yellow had a moderate impact on activity, whereas those in green showed similar activity to WT. (C) Bo98 active site colour coded according to impact of mutation on activity vs CX, relative to positioning of xylo-oligosaccharide substrate.

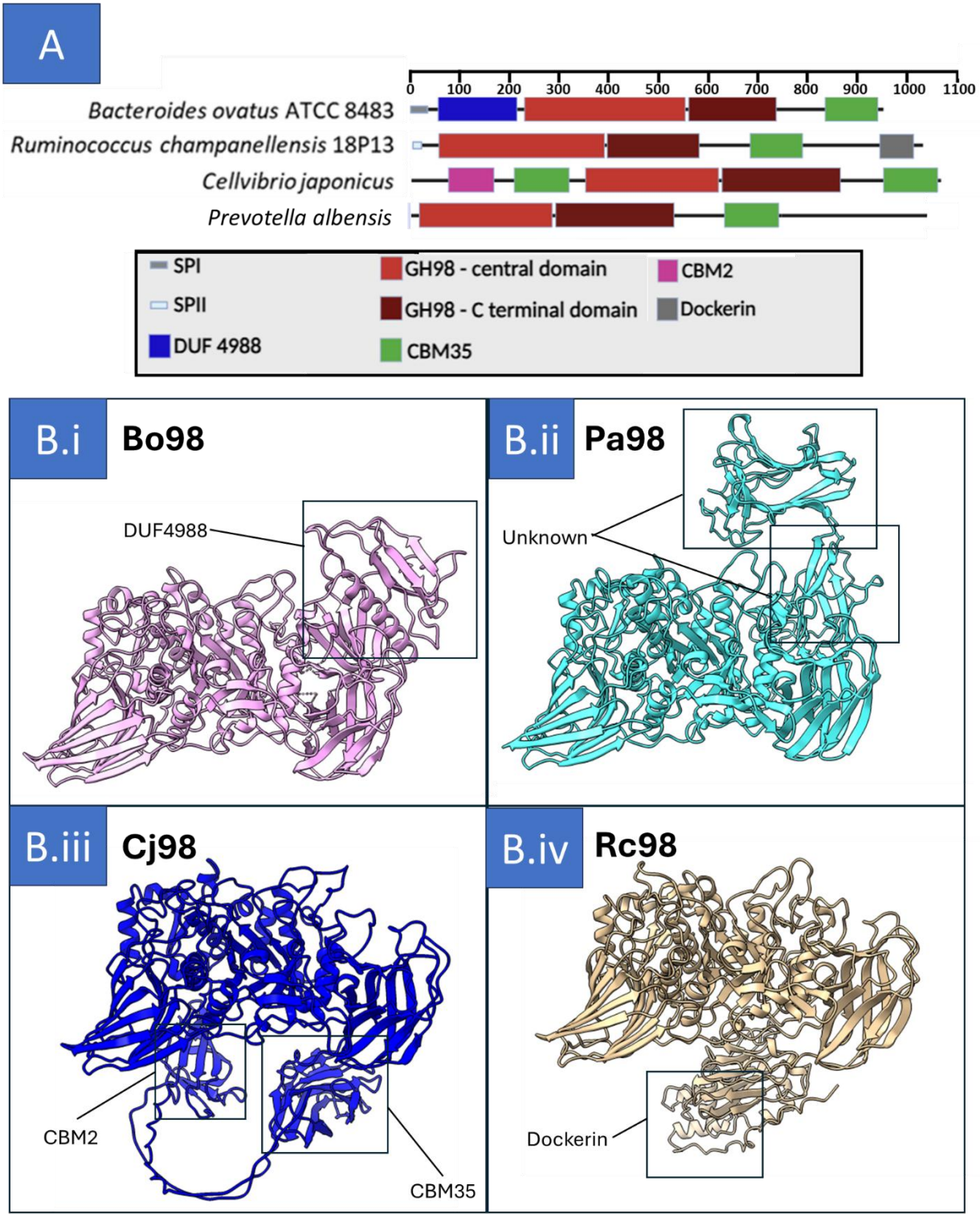
Lack of quantitative assays for Bo98 activity posed difficulty in data interpretation, so inactive mutants were deemed to be those where no smear could be visualised on TLC following 1 hour of reaction with an enzyme concentration of 1µM (**Figure 4-19A**).

Where the TLC smear appeared similar to WT activity, mutants were considered fully active, and for any level of activity in between the mutant was said to have a moderate impact on activity.

As already discussed, mutation of the predicted acid catalytic residue to alanine, E361A, resulted in complete loss of activity, and this residue was completely conserved amongst endoxylanase GH98 enzymes (**Figure 4-18A**). Interestingly, alanine mutation of one of the candidate general bases (based on comparison with Sp4GH98 **Figure 4-17** (Higgins *et al.*, 2009)), D467A, only had a minor impact on catalytic activity despite its complete conservation, however the other, E508A, resulted in no activity, indicating that the glutamate at position 508 may be the catalytic base in the Bo98 enzyme (**Figure 4-18**). Mutation of conserved tryptophan at position 364, which forms a stacking interaction with the backbone xylose sugar at the +2 subsite also knocks out Bo98 activity, as does the alanine mutation of Trp 404, which forms a pi stack at the +3 subsite, in this case with the X1 sugar, potentially important in the recognition of long xylan chains. Mutation of Lys436, which is conserved in Sp4GH98 (Lys220) (**Figure 4-17A+B**) and thought to be involved in stabilisation of the catalytic group (Higgins *et al.*, 2009) does not seem to impact on Bo98 activity.

Overlay of the Bo98 structure with Alphafold2 (Bryant *et al.*, 2022) models of the other GH98 endoxylanases characterised here; Rc98, Cj98 and Pa98, showed that although different domains were present (**Figure 4-20 A+B**), the core GH98 and CBM35 domains were highly conserved in all 4 enzymes (**Figure 4-20 C**). As discussed previously, Bo98 possesses the DUF hypothesised to function in incorporation of the enzymes into a utilisome complex, whereas Rc98 possesses a C-terminal dockerin domain for interaction with scaffoldins within the cellulosome. Interestingly, DBCAN does not predict the CBM35 domain in Pa98 from *S. albensis,* as in other examined

GH98 endoxylanases, suggesting that this region is more divergent in sequence. Despite this, structural overlay shows very high conservation, such that we hypothesise that this may be a more divergent member of the CBM35 family, although further studies would be required to ascertain this.



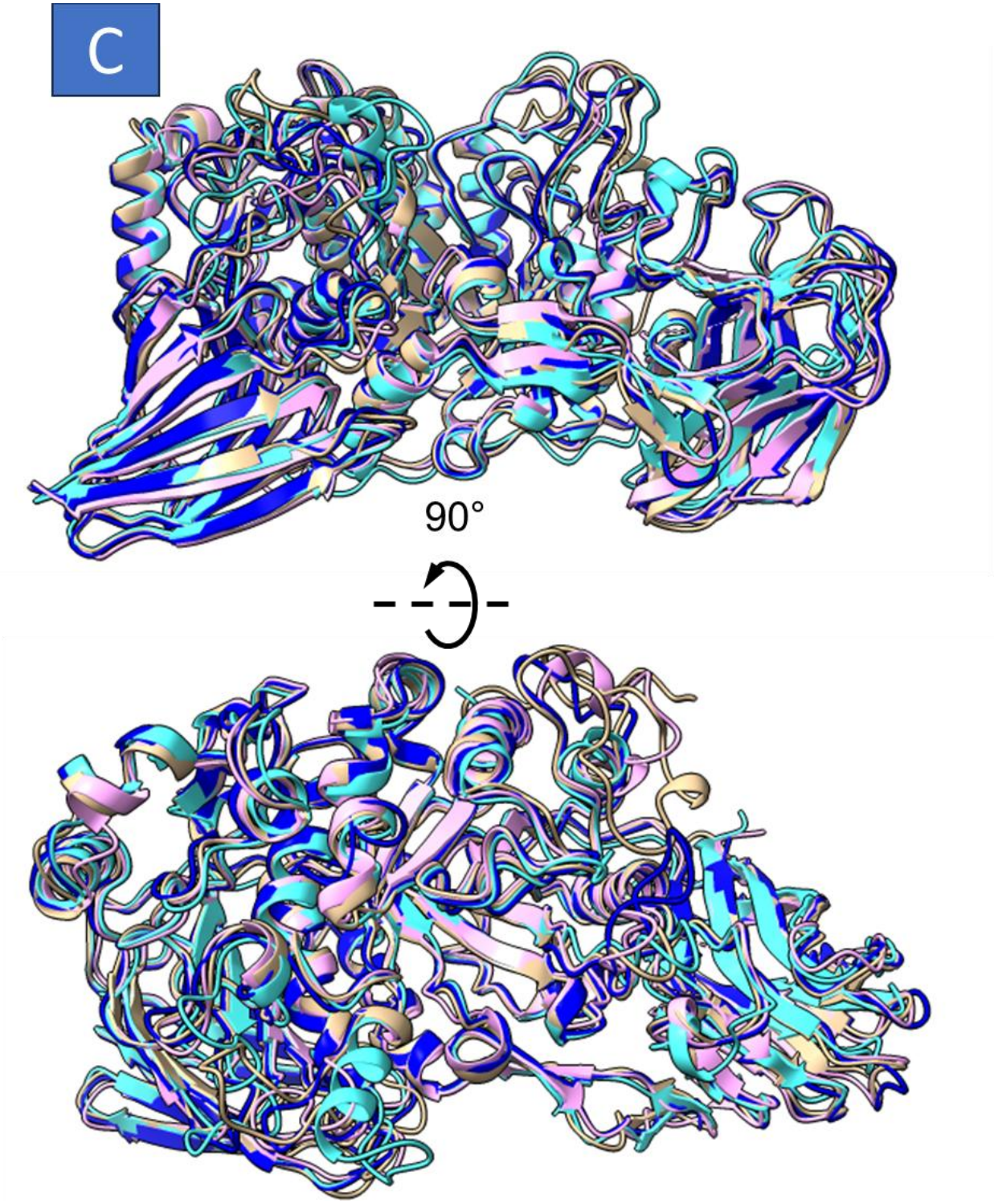
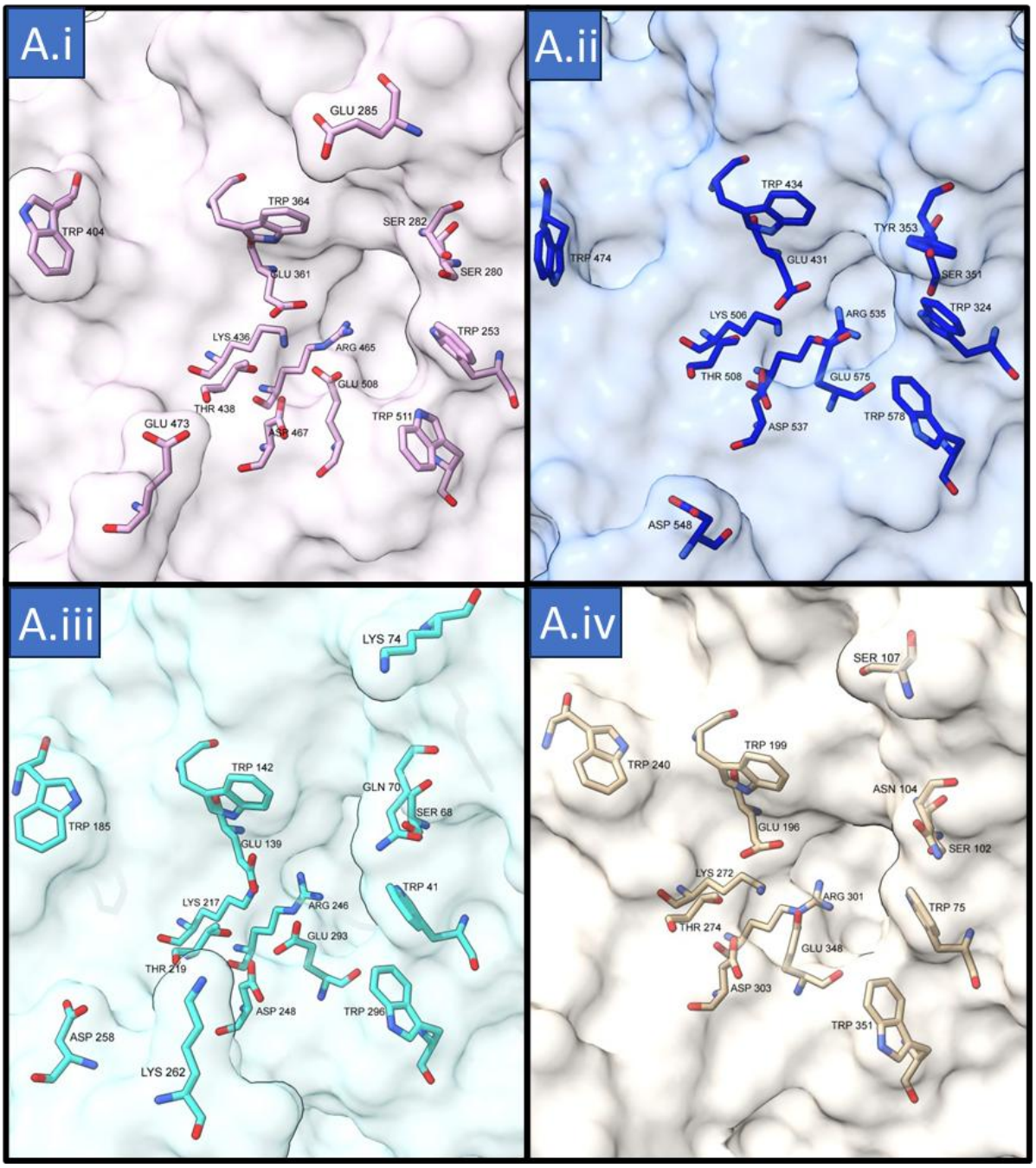


Figure 4-20: Structure of four functional GH98 endoxylanase enzymes demonstrating high conservation of core catalytic domains and the C-terminal CBM35. (A) Domain composition of GH98 endoxylanases demonstrating the core catalytic domains and CBM35, but divergence in other domains. (B) AlphaFold 2 models of Cj98, Pa98 and Rc98, as well as Bo98 experimental structure demonstrating overall architecture and highlighting additional domains in the different proteins. (C) Overlay of the 4 GH98 endoxylanase structures demonstrating high conservation of the catalytic domains and the CBM35.

Further to overall conservation in tertiary structure between the four characterised GH98 endoxylanases, overlay of Alphafold2 models showed complete conservation and orientation of core predicted catalytic apparatus (**Figure 4-21**), comprised of residues corresponding to Bo98 Glu361, Asp467, Trp404, Trp364, Ser280, Lys436, Thr438, Arg465, Glu508, Trp253 and Trp511, indicating that these residues are likely to be of functional importance for GH98 endoxylanase activity.

Interestingly, on inspection of the Bo98-ligand structure it was hypothesised that the opposing tryptophan pair Trp253 and Trp511 create a channel for the xylan chain at the proposed -2 enzyme subsite (**Figure 4-16C + 4-18C**) However, despite their conservation in the 4 endoxylanase GH98 enzymes studied (**Figure 4-21A**), impact of mutation of these residues in Bo98 was limited; W511A possessed activity akin to WT Bo98, whereas W253A showed reduced but still measurable activity (**Figure 4-19B**), suggesting that neither tryptophan residue is critical for enzyme function or substrate recognition. This suggests that these tryptophan residues may not be very important for substrate specificity, such that binding occurs at a low affinity which may be why we do not see sugar bound in the negative subsites past -1.


A.i
GLU 285
TRP 364
TRP 404
SER 282
SER 280
GLU 361
LYS 436
ARG 465
TRP 253
GLU 508
THR 438
GLU 473
ASP 467
TRP 511
A.ii
TRP 434
TYR 353
TRP 474
GLU 431
SER 351
LYS 506
ARG 535
TRP 324
THR 508
GLU 575
ASP 537
TRP 578
ASP 548
A.iii
LYS 74
TRP 142
GLN 70
TRP 185
SER 68
GLU 139
TRP 41
LYS 217
ARG 246
GLU 293
THR 219
ASP 248
TRP 296
ASP 258
LYS 262
A.iv
SER 107
TRP 240
TRP 199
ASN 104
GLU 196
SER 102
LYS 272
ARG 301
TRP 75
THR 274
GLU 348
ASP 303
TRP 351

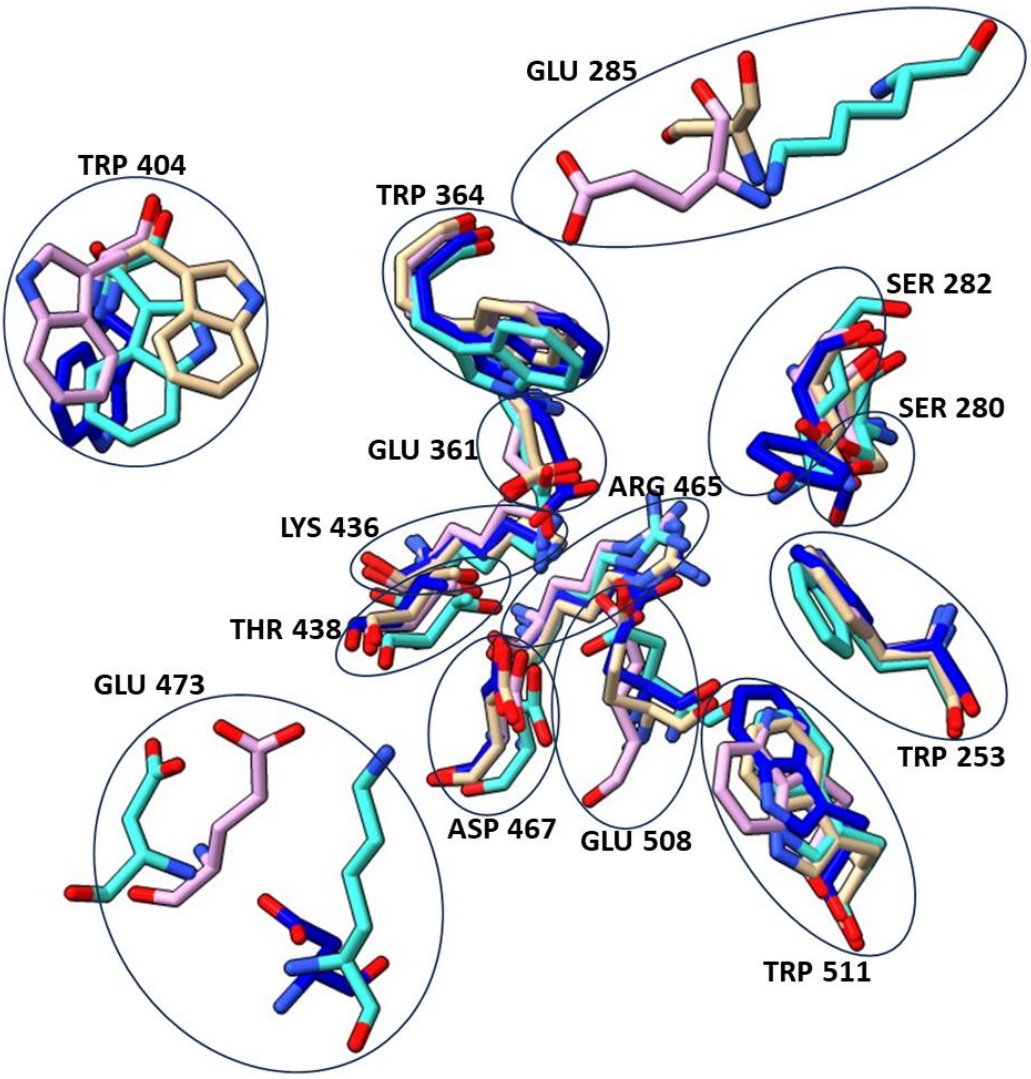


Figure 4-21: Conservation of active site residues in GH98 endoxylanases. (A) Analysis of predicted active site residues from crystal structure of (A. i) Bo98 (pink) and those from Alphafold2 (Bryant *et al.*, 2022) models of (A. ii) Cj98 (navy blue), (A. iii) Rc98 (beige) and (A. vi) Pa98 (teal). (B) Overall active site morphology and residues are highly conserved, particularly in positions corresponding to Glu 361, Asp 467, Trp 404, Trp 364, Ser 280, Lys 436, Thr 438, Arg 465, Glu 508, Trp 253 and Trp 511 residues of Bo98.

Submission of GH98 sequences to weblogo3 for creation of a sequence logo(Crooks *et al.*, 2004) (**Figure 4-22**) demonstrated that active site encoding regions were much more conserved in putative endoxylanase sequences, compared those regions in the GH98 family as a whole. Those GH98 enzymes possessing a C-terminal CBM35 domain were predicted as endoxylanases. In all GH98 sequences we found complete

conservation of the predicted acid nucleophile, E361, and candidate base residues E508 and D467, as well as the substrate stacking residue W364, and the other functionally important residue T438.

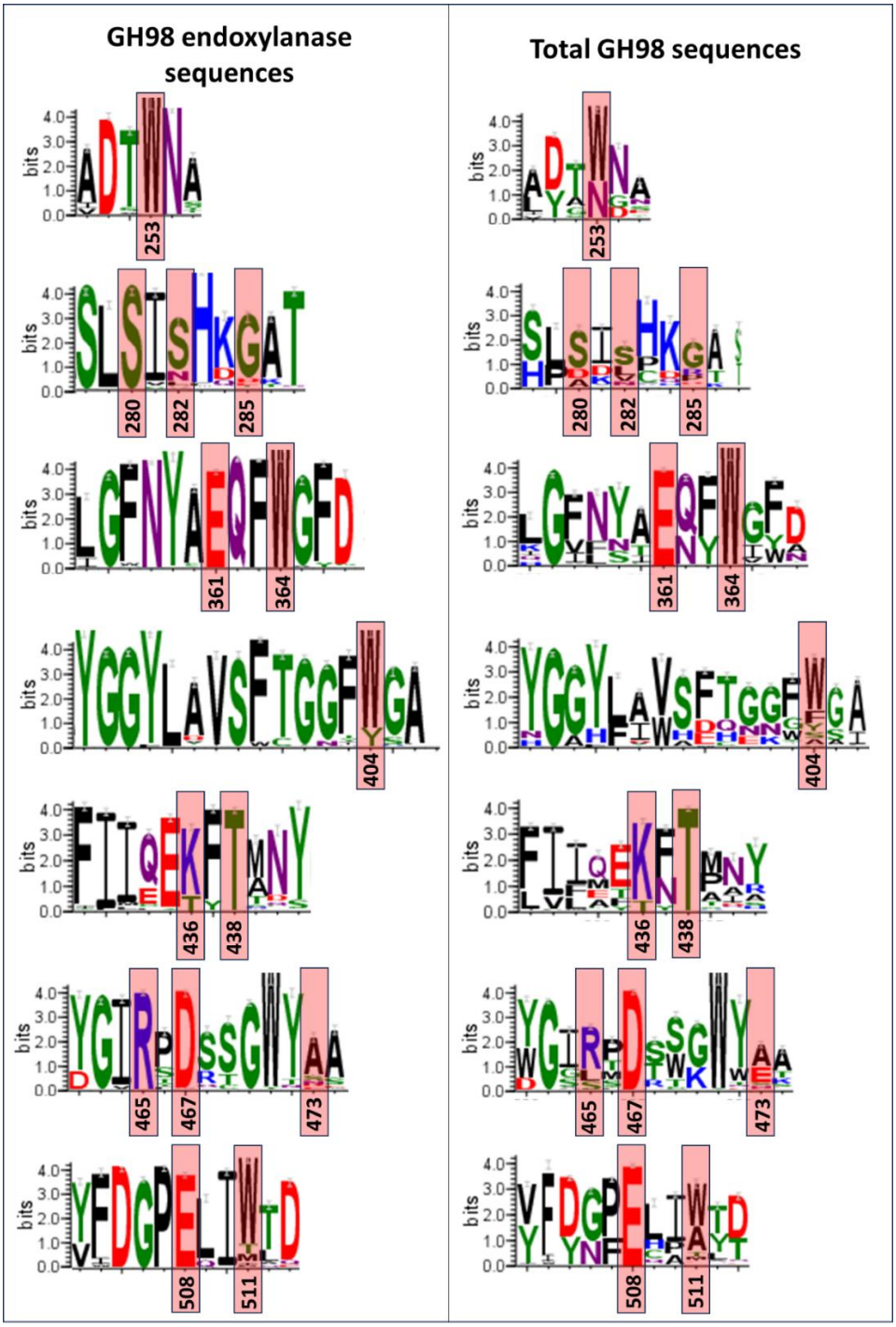


Figure 4-22: WebLogo analysis of active site of enzymes within the GH98 family. Regions containing predicted active site residues from all GH98 sequences within the CAZy database (right) and just predicted endoxylanases (based on phylogeny) (left) were submitted to WebLogo (Crooks *et al.*, 2004). Predicted catalytic and substrate binding residues are highlighted in red boxes, numbered according to Bo98 sequence.

Residues Ser280 and Arg465 from Bo98 were conserved in endoxylanase enzymes, but not in galactosidases, suggesting that these residues may play a role in recognition of the complex xylan substrate. Position and orientation of these residues suggests that they may be involved in preference for arabinose in the pocket of the -1' site of Bo98 (**Figure 4-15**). However, alanine mutation of either only partially impacts Bo98 activity (**Figure 4-19**) indicating that neither Ser280 nor Arg465 are critical for catalytic activity.

Interestingly, when examining conservation of active site residues amongst GH98 endoxylanase enzymes Trp253 has high conservation, whereas Trp511 is not as conserved (**Figure 4-22**), in agreement with mutant activity data where W253A impacts Bo98 enzymatic activity whereas W511A does not (**Figure 4-19**). This suggests that potential pi stacking of W253 alone at the -2 subsite may be sufficient for substrate recognition. Unsurprisingly, neither tryptophan is conserved in β-galactosidase GH98 enzymes, due to the small size of blood-group sugar substrates and exo-like activity requiring recognition of the non-reducing end of these sugars negating the requirement for a -2 enzyme subsite (**Figure 4-18B**).

The glutamate at position 285 of Bo98 seemed to be essential for enzyme activity and we hypothesise that this could be involved in binding with side chain or capping sugars which are shown to be present in GAX (**Figure 1-6**). However, this Glu285 residue is not conserved amongst endoxylanase enzymes, and a glycine residue is most prevalent in that position (**Figure 4-22**). Similarly, Glu473 is not conserved, often with an alanine residue in the equivalent position. Based on structural studies we predicted that E473 may form important interactions with the +2' arabinose decoration of the X2 sugar, however conservation and activity data suggests that this is not critical.

Trp404 in Bo98 appears to form a stacking interaction with the +3 xylose residue, X4, reflecting the ability of the enzyme to recognise and bind to long xylan chains (**Figure 4-15**). Within GH98 endoxylanase enzymes this position is occupied by tryptophan, or occasionally tyrosine, which may be able to form a similar stacking interaction. Less conservation is seen in GH98 galactosidases, again reflecting differences in substrate recognition.

Of interest is the lysine in position 436 of Bo98, which is conserved in Sp4GH98 and said to be important in stabilisation of the active site. However, the role of this residue in GH98 endoxylanases is poorly understood, as alanine mutation does not impact activity, and some GH98 enzymes possess a threonine in this position (**Figure 4-19**). It may be that this residue is critical for recognition of blood group sugars by β-galactosidases, but not for endoxylanase activity.

4.2.9 Biochemical studies into Bo98 substrate specificity

Structural and activity data suggest that the critical specificity determinant of Bo98 is more than simple arabinose side chains, and that double xylose-arabinose decoration at the +2 site may be important in recognition. To improve understanding of enzyme specificity CX was pretreated, individually and in combination, with four different glycosidases from *B. ovatus*, to remove side chains and assess importance of particular decorations for Bo98 activity. Clones of enzymes from Rogowski *et al.* (Rogowski *et al.*, 2015) (**Table 4-1**) were recombinantly expressed in an *E. coli* Tuner expression system as described previously. Protein purification was visualised by SDS PAGE (**Figure 4-22**).

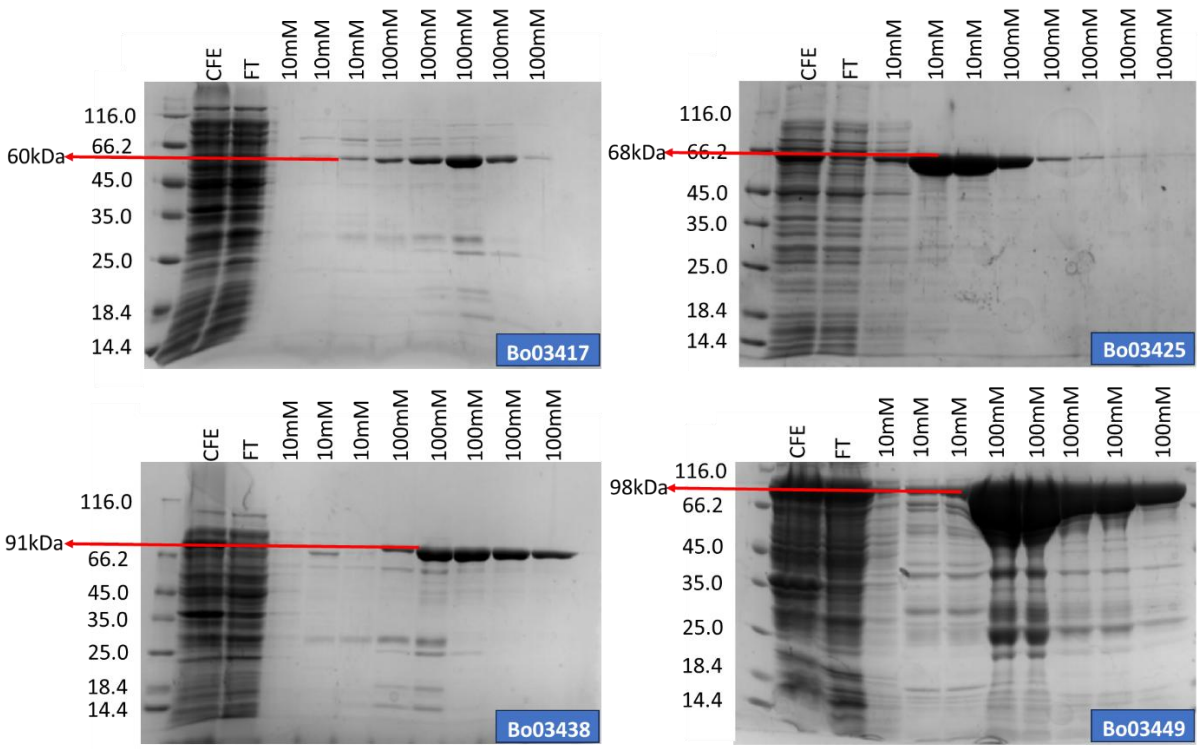


Figure 4-23: Expression and purification of xylan side-chain removing glycosidases. Four xylan active glycosidases from the *B. ovatus* large xylan PUL, previously characterised by Rogowski *et al.* (2015) were recombinantly expressed in *E. coli*. His-tagged protein was subsequently purified by metal affinity chromatography and analysed by SDS PAGE with a 14.4 to 116kDa molecular weight ladder (left hand lane of each gel). Cell free extract (CFE), flow through (FT) and 8 imidazole-eluted fractions were collected: 3 with 10mM imidazole, followed by 5 with 100mM imidazole. Bo03417 GH43 double-specific arabinofuranosidase, Bo03425 GH43 single-specific arabinofuranosidase, Bo03438 GH95 α-L-galactosidase and Bo03449 GH115 α-glucuronidase recombinant enzymes possessed molecular weights of 60kDa, 68kDa, 91kDa and 98kDa respectively.

Enzyme activity was as described previously (Rogowski *et al.*, 2015); Bo03417 GH43 α-L-arabinofuranosidase targets O3 linked xylose or arabinose decorations only where the backbone xylose residue is double decorated at O2 and O3 with L-arabinose, or with D-xylose at O3 and L-arabinose at O2. Bo03425 GH43 arabinofuranosidase targets single O2 or O3 linked substitutions on the xylan backbone, releasing arabinose and a small amount of xylose from CX. Bo03438 GH95 possesses α-L-galactosidase activity, and Bo03449 GH115 α-glucuronidase removes glucuronic acid

(GlcA) or methyl-glucuronic acid (MeGlcA), with a preference for internally substituted xyloses rather than chain ends (**Figure 4-24A+B**).

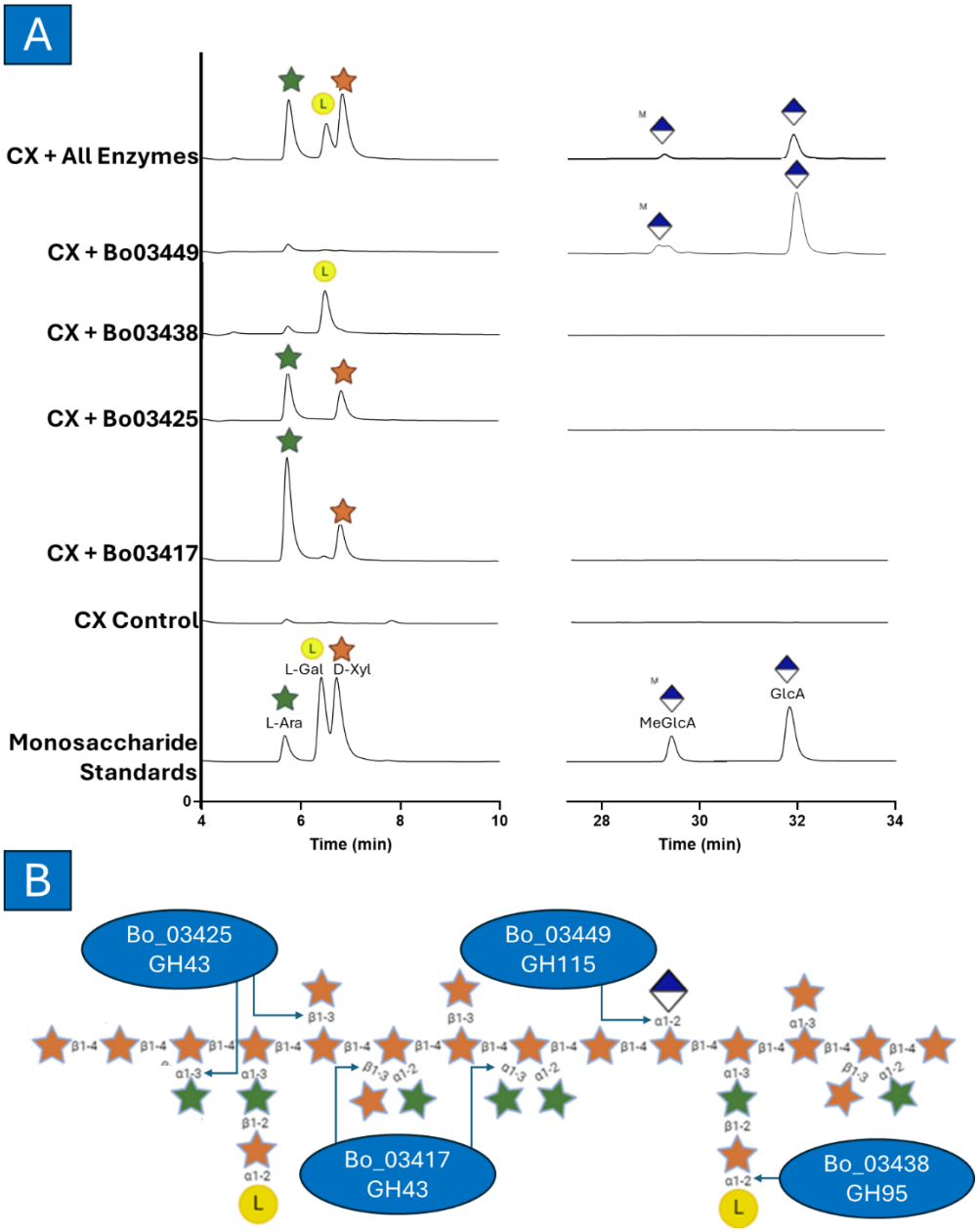
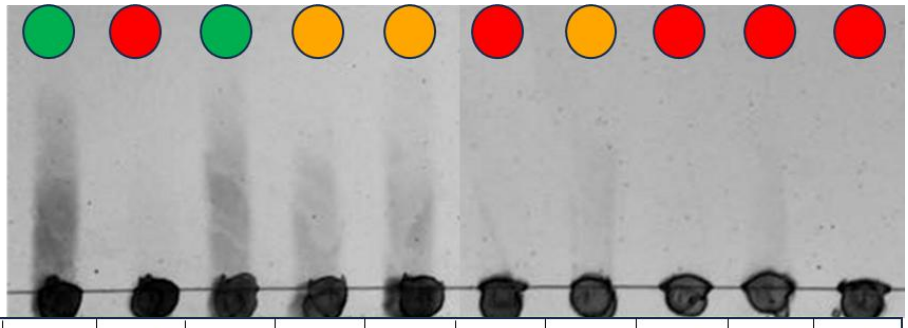


Figure 4-24: Activity of xylan active debranching enzymes used to pre-treat corn xylan. (A) High performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) analysis of four xylan-active enzymes from B. ovatus against corn xylan. Bo03449, a GH115 enzyme, displayed α-glucuronidase activity. Bo03438, a GH95, showed α-L-galactosidase activity, with a small amount of arabinose released. GH43 enzymes Bo03417 and Bo03425 both displayed arabinofuranosidase/xylosidase activity, releasing both arabinose and xylose from corn xylan. (B) Schematic demonstrating the target linkages of the *B. ovatus* large xylan PUL debranching enzymes for a model complex GAX. Although Bo03425 and Bo03417 GH43s both possess arabinofuranosidase/xylosidase activity, releasing arabinose and xylose side chains, Bo03425 targets single sugar decorations, whereas Bo03417 targets double substitutions.

CX was pretreated with each glycosidase individually, and in combination, to remove sidechains, before incubation with Bo98. Interestingly, when GlcA and MeGlcA acid were removed from CX with Bo03449, Bo98 did not show activity against the resulting substrate, suggesting that GlcA or MeGlcA decorations may be important for substrate recognition by Bo98 (**Figure 4-25**).



Bo03449 GH115	X	✓	X	X	X	✓	X	✓	✓	✓
Bo03438 GH95	X	X	✓	X	X	✓	✓	X	✓	✓
Bo03425 GH43	X	X	X	✓	X	✓	✓	✓	X	✓
Bo03417 GH43	X	X	X	X	✓	✓	✓	✓	✓	X

Figure 4-25*:* Impact of pretreatment of corn xylan with xylan active debranching enzymes on Bo98 endoxylanase activity. Corn xylan (CX) was pretreated with each of the four xylan debranching enzymes, individually and in combination. 1 µM final concentration of Bo98 was assayed against the resulting treated substrates for 1 hour at 37 ºC. Green circles represent Bo98 activity equivalent to that of untreated CX. Orange circles represent pretreatment resulting in some loss of Bo98 activity and red circles where pretreatment completely abolishes Bo98 activity.

Despite GlcA and MeGlcA substitutions apparently critical for Bo98 activity, no electron density could be modelled as GlcA in the x-ray crystallography model (**Figure 4-14**). Although GlcA or MeGlcA is clearly important for Bo98 activity we do not know at which subsite it is recognised, or with which residues it may interact. Biochemical data (**Figure 4-25**) suggests that removal of arabinose or xylose double decorations with Bo03417 does not impact on Bo98 activity, seemingly in contrast with previous data (Rogowski *et al.* 2015). However, it is possible that due to spatial constraints from the high branching density of CX, Bo03417 may not be capable of complete removal

of double decoration from CX without prior removal of other side chains, such that when CX is treated with Bo03417 alone, or with the limited repertoire of enzymes utilised in this study, it cannot remove the double decorations recognised by Bo98. Furthermore, structural studies suggest that an arabinose decoration of the X4 sugar in the pocket-like -1' site is tightly recognised (**Figure 4-15**), and therefore likely important for enzyme activity. Again, if Bo03425 failed to remove these arabinose substitutions from CX, these methods may lead to underestimation of the importance of these residues for Bo98 activity.

To further investigate the importance of glucuronic and methyl-glucuronic acid decorations, or double xylose-arabinose decorations, on Bo98 function, we assessed the binding of the D467A Bo98 protein to CX pre-treated with Bo3417 and Bo3449 via ITC (**Figure 4-26**). The D467A mutant was chosen due to its stability during ITC experiments, whereas E361A was unstable during ITC, falling out of solution during the titration process. The data demonstrated that, whilst CX treated with Bo3417 or Bo3449 individually did not greatly interfere with Bo98-CX binding, treatment with both enzymes led to lower heats of binding and potentially lower affinity, indicating dual sugar affinity of Bo98 for GlcA and arabinose (**Figure 4-26**). This suggests that GlcA or MeGlcA and double arabinose/ xylose decorations are important recognition factors, such that when all these substitutions are removed, Bo98 shows reduced affinity for its substrate. Interestingly, treatment with either enzyme alone seemed to improve the affinity of Bo98-D467A for the resulting substrate over the untreated CX. Overall, it is not clear why we see these changes in binding to different substrates. Further experimental data, potentially with other Bo98 mutants, would be required to further elucidate this.

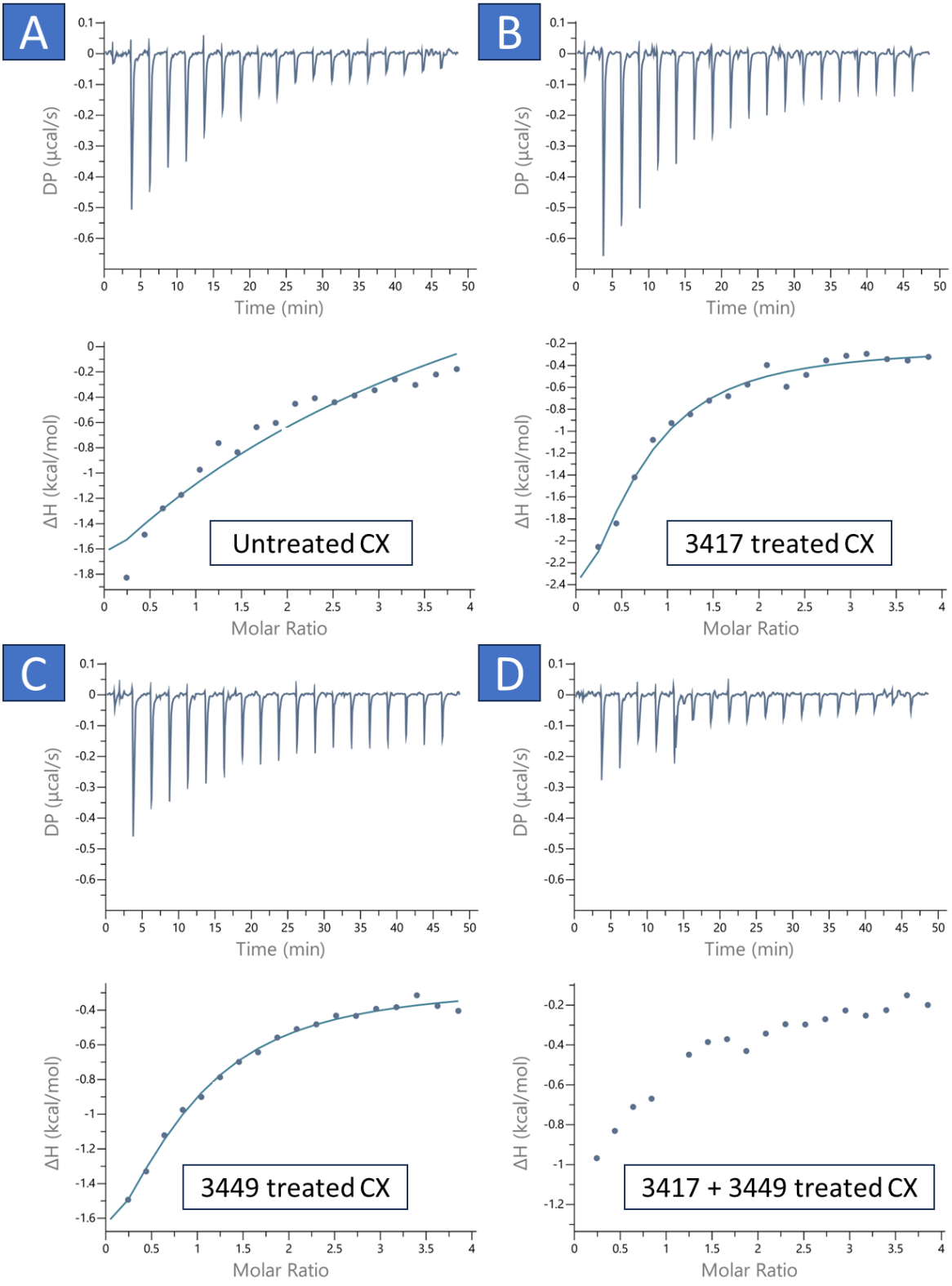


Figure 4-26: Binding of Bo98 D467A mutant to CX and pretreated CX. CX was treated with Bo03417 to remove double arabinose/ xylose decorations, Bo03449 to remove GlcA and MeGlcA, or both enzymes together. Substrates were subsequently dialysed to remove monosaccharide product. 100 µM Bo98D467A was titrated against 20 mg mL^{-1} CX derived substrate. Treatment of CX with both enzymes had a detrimental impact on the ability of Bo98 to bind to the debranched xylan.

Based on structural and biochemical data to date, we suggest that decoration of the +1 xylose with a xylose residue, and the -1 xylose with arabinose is likely critical for Bo98 activity, potentially double xylose-arabinose substitution at the +1 site. GlcA or MeGlcA decorations seem to be essential, albeit at an unknown location. Overall, it is the complexity of these interactions which determines the specificity of GH98 endoxylanase enzymes for complex glucuronoarabinoxylans.

4.3 Discussion

The first characterised members of the GH98 family were shown to possess β1,4-galactosidase activity, with specificity for particular blood group sugars (Anderson *et al.*, 2005; Higgins *et al.*, 2009; Wu *et al.*, 2021). These enzymes to date have been shown to recognise the NR terminal of either blood group A or B sugars or Lewis Y structures and cleave the internally at the Gal-GlcNAc bond, requiring multiple decorations for specificity, depending on the enzyme. These β-galactosidase enzymes are hybrid endo-exo acting (i.e. endo-β-galactosidases), in that they bind and cleave at the end of the sugar chain like exo-enzymes, but they recognise more than one sugar at their negative subsites. Rogowski *et al.* (2015) first identified endo-xylanase activity within the GH98 family; *B. ovatus* ATCC8483 encodes a GH98 enzyme which shows specific endoxylanase activity against highly decorated GAX. Endoxylanase enzymes catalyse the hydrolysis of β-1,4-glycosidic linkages internally in xylans, degrading xylan chains to produce a range of shorter oligosaccharides (Malgas *et al.*, 2019), and most characterised endoxylanase enzymes belong to GH families 10 or 11, however, members of families 5, 8, 30 and 43 also possess this activity (Malgas *et al.*, 2019; Collins *et al.*, 2005). GH8 and GH98 enzymes are the only endoxylanases known to execute catalysis via an inverting mechanism. GH10 endoxylanases generally possess activity against soluble xylans with relatively few branches, with a

catalytic mechanism which involves recognition of the xylan chain but can accept substitutions where they do not cause steric blocks to binding the β1-4 xylose backbone. In comparison, GH11s show greater activity against unsubstituted xylan chains. GH30 endoxylanases specifically target glucuronic acid substituted GX, as they require a GlcA decoration in the -2' site for catalysis (Collins *et al.*, 2005). To date, only families GH5_21, GH5_34 and GH98 have been demonstrated to possess activity against complex GAX without prior debranching (Rogowski *et al.*, 2015; Labourel *et al.*, 2016; Larson *et al.*, 2003), potentially providing a competitive advantage to GH98/ GH5_21 encoding bacteria in certain environmental niches. Recombinant GH5_21 from *B. xylanisolvens* hydrolyses complex GAX, but also simpler AX such as wheat arabinoxylan, whereas the endoxylanase activity of four GH98 enzymes, as presented here, is unique to complex GAX (Rogowski *et al.*, 2015). Little is known regarding the crucial specificity determinants of these GH98 enzymes, nor the biochemical interactions determining such a unique and interesting specificity. Here, we provide significant insights into the structural and functional basis for endoxylanase activity within the GH98 family.

Phylogenetic analysis presented in this study demonstrates that characterised blood group targeting GH98 enzyme sequences sit as part of a distinct cluster, suggesting evolutionary divergence from endo-xylanase members of the same family (**Figure 4-4**). Similarly, SSN analysis shows that predicted endoxylanase GH98s, possessing a CBM35 module, all sit in one cluster, whereas other enzymes, assumed here to possess activity against blood group sugars or Lewis-type antigens, form 2 large clusters, one smaller cluster or lie individually (**Figure 4-6**). Interestingly, sequences seem to cluster according to family of CBM present indicating that the CBM modules may play a role in enzyme specificity. Distinct phylogenetic and SSN clustering

allowed prediction of enzyme activity as endo-β-galactosidase or endoxylanase based on activity of characterised enzymes. Furthermore, sequence domain analysis suggests that GH98 enzymes with blood group targeting β-galactosidase activity lack the CBM35 module, instead possessing multiple CBMs belonging to families 32, 47 or 51 (**Figure 4-6**). It is likely that difference in presence of CBM domains in GH98 enzymes with different activities may function in the recognition or targeting of specific ligands (Boraston *et al.*, 2004). CBM families 32,47 and 51 all show specificity for complex mammalian carbohydrates and show significant structural similarity to each other (Ficko-Blean & Boraston, 2012). CBM32 and CBM51 families contain characterised members with blood group sugar binding activity (Gregg *et al.*, 2008; Ficko-Blean & Boraston, 2009), recognising terminal galactose (Ficko-Blean & Boraston, 2012), whereas CBM47s recognise terminal fucose (Ficko-Blean & Boraston, 2012). The presence of these CBM families in predicted endo-β-galactosidase enzymes supports their known specificity for blood group sugars.

Interestingly, a construct of Bo98 lacking the C-terminus CBM35 domain could not be recombinantly expressed. When examining the structure of Bo98 it is clear that this CBM35 is tightly associated with the catalytic domain in xylanase GH98s and that the structure of catalytic domain would likely be disrupted without the CBM35 (**Figure 4-12A**).

The N-terminal DUF4988 domain seen in Bo98 is not widespread within GH98 endoxylanase enzymes, suggesting that is unlikely to play an important role in catalytic activity. A truncated construct of Bo98 lacking this DUF domain (Bo98ΔDUF) still possessed endoxylanase activity against CX, similar to that of WT, indicating that it is not critical for enzyme function (**Figure 4-9**). However, we would need more detailed time course experiments to determine whether this DUF domain has any impact on

rate of enzymatic reaction. We hypothesise that this domain is involved in spatial orientation of the enzyme on the outer membrane, as part of the utilisome complex with the PUL encoded SusCD and SGBP (White *et al.*, 2023), although there is no direct evidence of this as yet. DUF4988-containing GH98s are encoded solely by members of the Bacteroidota phylum (**Figure 4-5**), enhancing this hypothesis, and are most prevalent in *Bacteroides*, but also present in members of commensal gut genera *Phocaeicola* and *Alistipes*. DUF4988 is lacking from GH98 enzymes encoded by Bacteroidota *Prevotella* or *Paraprevotella* spp. For *Paraprevotella* spp. GH98 enzymes are also contained in PULs encoding Sus-like systems, indicating that this DUF domain is not critical for function as part of a putative utilisome complex, or that GH98 does not form part of a utilisome system in these species, or that it has a different structure. Although DUF4988 modules are found in many proteins, they are associated with numerous different modules, many of which do not have defined functions, providing further evidence that this DUF4988 is not involved in catalytic activity of Bo98.

All predicted blood group sugar active enzymes from the GH98 family are encoded by host associated species (**Figure 4-4**). A recent study characterised a GH98 enzyme with β-galactosidase activity from mucin-degrading human gut symbiont *R. gnavus* hypothesised that this enzyme may play a role in colonisation of individuals with blood group A, or in *R. gnavus* strain acquisition in infants (Wu *et al.*, 2021). GH98 blood group targeting enzymes are prevalent in pathogenic bacterial strains from the genera *Streptococcus,* in particular strains of *Streptococcus suis,* which causes zoonotic disease commonly associated with swine (Wertheim *et al.*, 2009), and *Streptococcus pneumoniae*, an opportunistic pathogen which can cause a wide range of infections (Weiser *et al.*, 2018). GH98 β-galactosidase binding to host specific antigens may play

a role in colonisation and tissue invasion, enhancing pathogenicity (Martens *et al.*, 2018; Higgins *et al.*, 2009; Wu *et al.*, 2021). GH98 encoding genes are not present in commensal *Streptococcus* spp. such as *S. sanguis, S gordonii* or *S. parasanguis*. Interestingly, both endoxylanase and β-galactosidase GH98 activity is present within the gut environment (Rogowski *et al.*, 2015; Wu *et al.*, 2021). Unsurprisingly, GH98 enzymes encoded by other plant cell wall targeting members of the gut microbiome such as ruminal S*. albensis* (Pa98) and human symbiont *R. champanellensis* (Rc98), also possess endoxylanase activity (**Figure 4-8**), whereas RgGH98 from *R. gnavus*, a gut-associated strain known to degrade host associated glycans, possesses β-galactosidase activity (Higgins *et al.*, 2009). This indicates the possibility of divergent evolution of these enzymes within the intestinal niche, as individual bacterial species evolved the ability to degrade and utilise different types of polysaccharides.

However, GH98 endoxylanase enzymes are not solely host associated but are found in a wide range of environmental niches. In fact, prevalence of GH98 enzymes within the human and animal intestinal tract seems to be relatively low (**Figure 4-4**). Although *B. ovatus* ATCC8483 possesses a GH98 endoxylanase encoding gene, these enzymes are not widely present in *B. ovatus* strains, or indeed in members of the gut microbiome. Although xylan-targeting PULs are present in almost all *B. ovatus* strains, few of these encode GH98 enzymes, suggesting that, although the ability to degrade xylans is highly important, the evolutionary role of this enzyme is not critical, despite its necessity for *B. ovatus* ATCC 3484 growth on GAX as a sole carbon source(Rogowski *et al.*, 2015). This may be due to lack of requirement of degradation of complex xylans in the intestinal environment due to overabundance of AXs or GXs which can be degraded by GH10 or GH11 endoxylanases, leading to lack of selection pressure for GH98 requirement. Alternatively, GH98 endoxylanases possess

redundancy in function with members of the GH5 subfamily 21, which has ability to degrade GAX but also simpler AX (Rogowski *et al.*, 2015; Labourel *et al.*, 2016). As discussed previously, Rogowski *et al.* (2015) showed that within *B. xylanisolvens* and *B. ovatus* strains xylan PULs are highly conserved with the exception of a GH98-GH5_21 swap in some strains, providing further evidence for this evolutionary redundancy (**Figure 1-21**). Furthermore, it may be that GH98 endoxylanases are capable of degradation of GAXs of higher complexity than CX, which almost certainly occur in nature, and further studies testing such substrates may be of interest.

Prevalence of GH98 endoxylanase enzymes within plant pathogens, particularly of the genus *Xanthomonas*, is high (**Figure 4-4**). *Xanthomonas* spp. cause disease in more than 400 plant hosts and have a major impact on the agricultural community, as well as food security in general (Timilsina *et al.*, 2020). Many GH98 enzymes are encoded by *Xanthomonas oryzae* pv. oryzae strains which colonise the vascular tissue of rice (Zhang & Wang, 2013) and *Xanthomonas citri* pv citri, a pathogen of citrus plants (Bansal *et al.*, 2017). Repertoires of cell wall degrading enzymes have been noted as important virulence factors for some *Xanthomonas* pathogens, in particular the role of xylanases in degradation of the host plant cell wall and provision of nutrients (Santos *et al.*, 2014; Solé *et al.*, 2015). Of particular interest, the secretion of xylanase enzymes via outer membrane vesicles has been associated with the virulence of *X. campestris* pv. vesicatoria, which causes bacterial spot disease in tomatoes and peppers (Chow *et al.*, 2015). Furthermore, the production of GH10 endoxylanases for the degradation of plant cell wall material has been demonstrated in *X citri* pv. citri, *X oryzae* pv. oryzae and *Xanthomonas axonopodis* pv. citri (Giuseppe *et al.*, 2023; Rajeshwari *et al.*, 2005). To date the role of GH98 endoxylanases in pathogenicity of *Xanthomonas* spp. has not been established, and the activities of such enzymes have

not been characterised. Attempts to recombinantly express Xo98, a GH98 encoded by *X oryzae* pv. oryzae, in the current study failed, however it is likely that this enzyme plays an important role in the degradation of complex GAX found in rice plants (Chiniquy *et al.*, 2012), which are likely recalcitrant to breakdown GH10 or 11 enzymes. Furthermore, Bo PULXylL was shown to be critical for growth on rice and sorghum GAX, indicating that Bo98 may be needed for rice and sorghum GAX degradation (Rogowski *et al.*, 2015). Due to the relative phylogenetic divergence of some Xanthomonas encoded GH98s, in comparison to those from gut environments (**Figure 4-4**), analysis of activity of these enzymes is an important area of future study. As GH98 enzymes are likely to play a role in pathogenicity, improved understanding of their functional mechanisms and role in host invasion could be important in understanding and preventing plant diseases, and their massive financial and food security burdens.

GH98 enzymes are also encoded by bacteria dwelling in the soil environment, such as *Paraflavitalea* spp. or *Cellvibrio japonicus,* where they may play a critical role in nutrient recycling (Gardner, 2016). Rarely, GH98 endoxylanases are encoded my marine microorganisms, namely *Vallitalea guaymasensis* and *Planctomycetes* bacterium K2D. This is of interest as a lot of marine xylans are known to consist of β1,3 linked xylose backbones (Sun *et al.*, 2021), and as such would not be substrates for GH98 endoxylanase enzymes if the activity demonstrated in this study is conserved. Perhaps, this suggests why prevalence of GH98 enzymes is low in marine bacteria.

This work provides the first experimental structure of an endo-xylanase member of the GH98 family. Unsurprisingly for members of the same family, Bo98 appears to share overall conservation of catalytic domain architecture with those of the previously

characterised Sp4GH98 beta-galactosidase enzyme (Higgins *et al.*, 2009) (**Figure 4-16A**). However, major differences are seen in active site morphology between the two enzymes. Sp4GH98 active site consists of a deep pocket, whereas Bo98 possess an open, elongated binding cleft to allow for recognition of heavily decorated substrate (**Figure 4-16B**), as noted by Labourel *et al.* (2016) for a GH5_34 endoxylanase. Furthermore, the catalytic machinery of Bo98 is conserved with that of blood group sugar targeting GH98 enzymes, as demonstrated by conservation in position and orientation of catalytic residues identified in Sp4GH98 by Higgins *et al.* (2009) (**Figure 4-17**) and as such data support conservation in position of cleavage and enzyme subsites.

Multiple other residues were shown by site directed mutagenesis to be important in xylan chain recognition were largely conserved in endo-xylanase enzymes (**Figure 4-19**), namely Trp404 and Trp253, however lacked conservation within blood group targeting enzymes, suggesting a critical role for these residues in xylan recognition. However, although it is clear how these residues are involved in recognition and positioning of a long xylan chain across the active site, from Trp404 at the +3 site to Trp253 at the putative -2 site, these residues are unlikely to function in specificity for complex GAX over less decorated xylans (**Figure 4-15**). Overall, it is not clear from experimental ligand bound data the exact residues involved in substrate recognition by Bo98. Although density does show a double xylose-arabinose decoration at the +1 site (**Figure 4-14C**), in accordance with predictions by Rogowski *et al.* (2015) based on lack of these decorations in simpler arabinoxylans such as WAX, structural data presented here do not elucidate the protein-ligand interactions which may be critical in determining this specificity. Based on ligand density the arabinose decoration at the -1 subsite sits in a tightly formed pocket (**Figure 4-15**). Spatial constraints and

interactions with Ser280 and Arg465 may confer specificity for arabinose at this site, however elucidation of binding interactions at other subsites is challenging from the data presented here.

Furthermore, it is unclear why we do not see density at the -2 subsite and beyond, although this may be due to weak binding at these sites (**Figure 4-14C, Figure 4-18C**) as if the ligand modelled here were a true substrate for Bo98, you would expect to see exo- activity and release of a disaccharide arabinose-xylose product. This is not supported by activity data presented here or in previous study (Rogowski *et al.*, 2015). When the -1 was modelled in a distorted configuration, the position of the O4 changes in a manner which would allow continuation of the chain to the proposed -2 subsite, such that it is also likely that positioning of the -1 xylose residue is not in the ground state as shown by the ligand density presented in **Figure 4-14C**, as it would not allow for continuation of the β1,4 xylose backbone chain into the predicted -2 subsite.

In order to further elucidate the critical specificity determinant of Bo98 we used four xylan debranching enzymes encoded by *B. ovatus* to selectively remove CX side chains and assess the impact on Bo98 activity (**Figure 4-24, Figure 4-25**). Pretreatment of CX with Bo03449 α-glucuronidase to remove glucuronic acid and methyl-glucuronic acid side chains prevented Bo98 activity against the resulting substrate, indicating the importance of these decorations in recognition of and activity against CX. The presence of GlcA or MeGlcA distinguishes GAX such as CX from AX, providing logic for this being critical for activity against Bo98.

Due to the complexity of the CX and its structural heterogeneity (Xu *et al.*, 2022), it is not known precisely where glucuronic acid substitutions occur, and it is likely that there are batch-related differences in CX structure due to many factors including breed,

climate and season of the corn harvest. Furthermore, differences in xylan isolation techniques could contribute to variation in CX decoration and branching structure. This provides great difficulty in assessing at which enzymatic site these GlcA and/or MeGlcA decorations are required for enzyme activity. Furthermore, no ligand density was seen in the x-ray diffraction data which could be modelled as a glucuronic acid or methyl-glucuronic acid sugar (**Figure 4-14C**) and based on the presented ligand position and orientation there is no obvious position for a uronic-acid decoration which would interact with any active site amino acid residue in a manner that could dictate enzyme specificity. It is formally possible that GlcA or MeGlcA could cap the arabinose decoration at the +1 subsite, as shown in **Figure 4-27** interacting with Glu285, which SDM data showed to be critical in Bo98 activity (**Figure 4-19**). It is possible that the glutamate residue may interact with the hydroxyl groups of the uronic acid decoration. However, lack of conversation of this residue in GH98 enzymes with the same activity (**Figure 4-21, Figure 4-22**) suggests that alternative residues may be able to perform a similar role in specificity for GlcA residues.

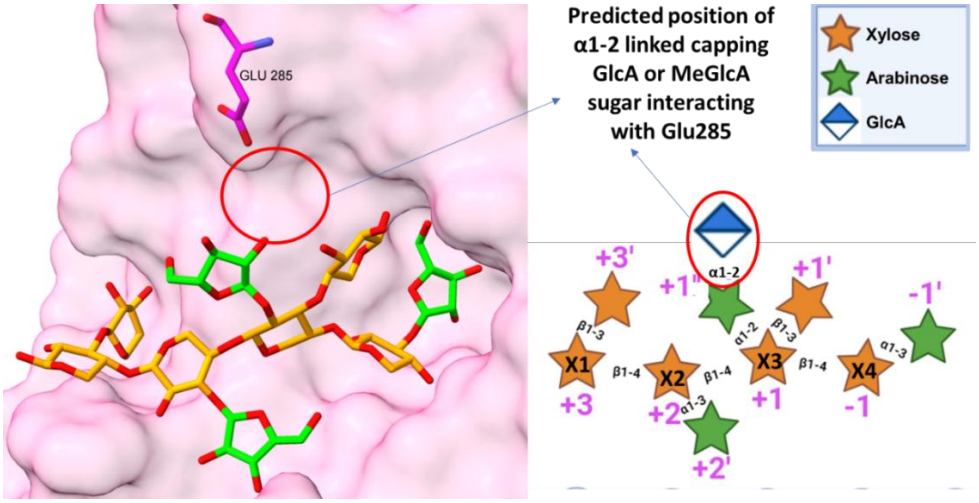


Figure 4-27: Putative position of capping GlcA or MeGlcA residue interacting with Glu285. Biochemical data suggests that GlcA or MeGlcA decorations are critical for Bo98 activity. Based on structure we predict that GlcA may form a cap on the chain decoration, bound to the arabinose decoration of the +1'' site, interacting with Glu285 of Bo98 which is critical for enzyme activity.

Rogowski *et al.* (2015) demonstrated that removal of double-specific arabinofuranosidase/xylosidase Bo3417 from a cocktail of CX-debranching enzymes restored Bo98 activity against CX, however biochemical studies presented here found that treatment of CX with Bo03417 alone did not prevent Bo98 activity (**Figure 4-24**). It is hypothetically possible that without CX debranching by a cocktail of enzymes, Bo03417 fails to remove all double xylose/arabinose decorations due to spatial constraints, and that these remaining decorations are critical for Bo98 substrate recognition. Despite this, substrate specificity determinants are still to be fully elucidated.

The novel data presented here elucidates how complex xylans are broken down in the gut. This is of high importance in human health, as xylans are abundant in the human diet, and their xylooligosaccharide degradation products is known to have positive benefits on digestive health. In fact, the prebiotic potential of xylooligosaccharides has been well documented (Aachary & Prapulla, 2011; Samanta *et al.*, 2015). Furthermore, studies have shown that treatment with xylooligosaccharides may have beneficial effects in patients with diabetes mellitus, cardiovascular disease, obesity, inflammatory bowel disease and others (Jana *et al.*, 2021). These GH98 endoxylanase enzymes may provide a novel mechanism for generation of xylooligosaccharides, which may have structural differences and hence function differently as prebiotics compared to oligosaccharides produced by classical GH10 or GH11 endoxylanases. This means that there is potential for GH98 enzymes to be involved in the *in vitro* generation of novel xylooligosaccharides with prebiotic function.

Furthermore, in the intestinal niche, extracellular degradation of complex xylans by these GH98 endoxylanase enzymes provides opportunity for many other species of the gut microbiota to share in the complex xylooligosaccharide degradation products,

potentially enhancing diversity. Nutrient sharing in the gut has been shown to promote species-specific cooperation (Glowacki & Martens, 2021), for example members of the *Bifidobacterium* genus seem to be well adapted to interspecies carbohydrate cross-feeding (Ventura *et al.*, 2012). Of particular interest, *Bifidobacteria* spp., many species of which are incapable of xylan degradation, can utilise xylooligosaccharide breakdown products generated during co-culture with *Bacteroides* spp. (Zeybek *et al.*, 2020). Although ability of *Bifidobacteria* spp. to utilise GAX degradation products has not been demonstrated as yet, there are likely species with these capabilities. This may mean that, although not generally hugely abundant in the human gut, individuals with a GH98 enriched microbiome may have increased potential for nutrient sharing, and hence an improvement in microbiome diversity and function.

There are also potential industrial applications of these GH98 endoxylanase enzymes. The concept of enzymatic pre-treatment of animal feeds has been studied for many years, removing some of the complex plant-cell wall material to improve accessibility of proteins and more soluble carbohydrates (Stokes & Chen, 1994). Due to the high proportion of CX in the husk of corn grain, pre-treatment with a cocktail of enzymes including a GH98 endoxylanase may prove a novel method of improving feed efficiency.

Although we have many novel insights into the unique specificity of GH98 endoxylanases, we still do not fully understand the exact decorations required for enzymatic activity. Further work in this area is likely to include further crystal soaks to try and improve understanding of where GlcA or mGlcA may be recognised. Further investigation of the role of predicted GH98 endoxylanases in pathogenic *Xanthomonas* spp. may be critical in understanding the causative mechanisms of plant disease and could uncover novel targets for treatment.

Overall, evidence suggests that Bo98 activity involves the recognition of multiple side chains, including the presence of a doubly substituted xylose-arabinose decoration at the +1 site, arabinose decoration at the -1 site and GlcA or MeGlcA decoration at an unknown location, dictating the enzyme's specificity for complex GAX. This research has many potential important applications in the nutritional sciences, in particular development of prebiotics to improve health.

5. Investigating the role of conserved CBM35 domains in GH98 endoxylanase enzymes

5.1 Introduction

Carbohydrate binding modules (CBMs) are non-catalytic proteins with the capacity to bind to soluble and crystalline carbohydrates, mainly carbohydrates. CBMs are generally not discrete proteins, but rather co-occur with catalytic GH or PL modules, and many GH enzymes possess one or more CBM domain (Gilbert *et al.*, 2013). CBMs often function to promote the recognition of plant cell wall substrate and association of the enzymes with the substrate, ultimately enhancing catalytic efficiency (Boraston *et al.*, 2004). CBMs are classified into sequence-based families within the CAZy database (Cantarel *et al.*, 2009), of which there are currently 106 (April 2025), and can also be characterised into three main types according to the shape and degree of polymerisation of their target ligand. Type A CBMs possess a planar binding site formed from aromatic amino acid residues creating a planar surface for binding to crystalline polysaccharides. Type B CBMs are *endo*-type and binding site morphology consists of an elongated groove or cleft which can recognise long polysaccharide chains. Type C, or *exo*-type CBM binding sites consist of small pockets for recognition of the non-reducing ends of polysaccharides, or short oligosaccharides (Boraston *et al.*, 2004).

CBMs possess one of 7 fold types, the most common of which is the β-sandwich fold. This consists of 2 overlapping β-sheets, each with 3 to 6 antiparallel β-strands (**Figure 5-1**) (Boraston *et al.*, 2004). Within CBMs comprising this beta-sandwich fold structure there are 2 distinct glycan binding locations; the variable loop site (VLS) which interconnects β-strands at one end of the sandwich, and the concave face site (CFS)

on one of the β-sheets (Abbott & van Bueren, 2014). Studies show that ligands of multiple different chemistries and structures bind at the VLS, possible due to the variable loops allowing flexibility in binding site morphology. On the other hand, the more restricted morphology of the β-sheet surface means that the CFS is less capable of recognition of highly-decorated or divergent glycans (Abbott & van Bueren, 2014).

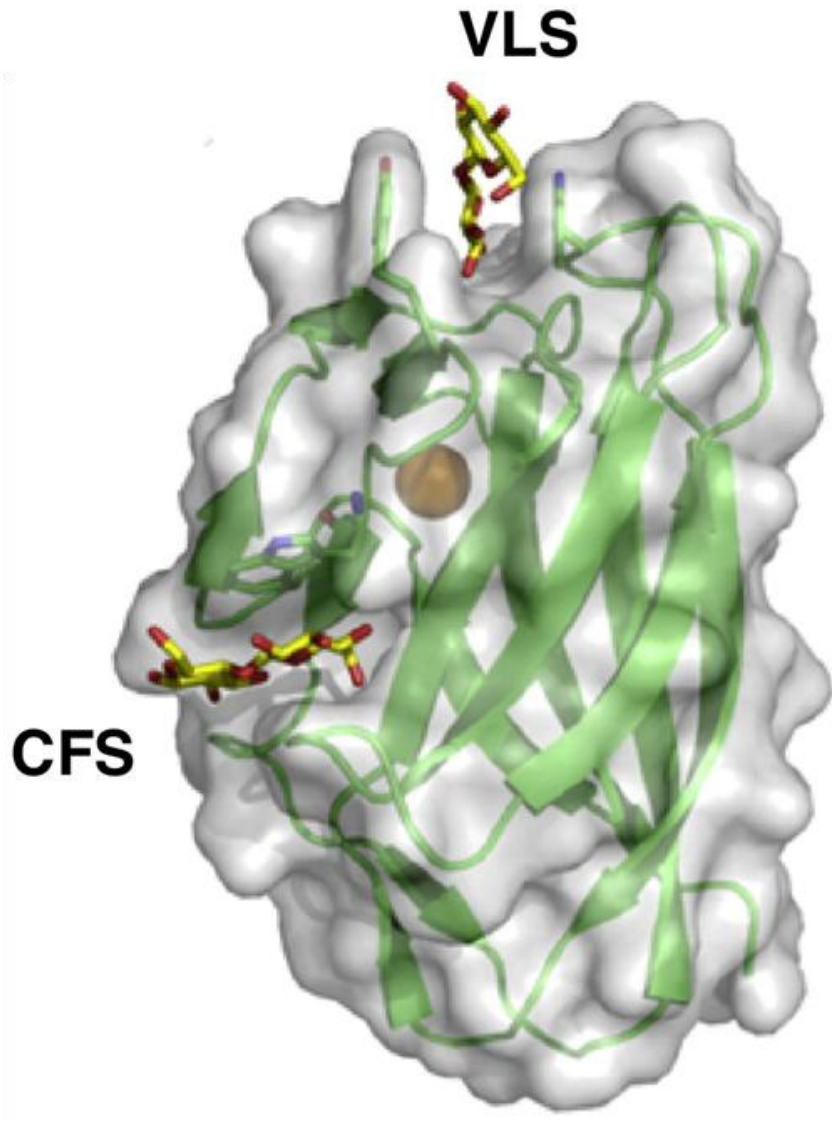


Figure 5-1: Location of binding sites in β-sandwich carbohydrate binding modules in complex with heterogeneous β-linked carbohydrates. Location of the variable loop site (VLS) and concave face site (CFS) on the surface of the CBM6 from *Cellvibrio mixtus* (CmCBM6; 1UYY). The backbone is displayed as a green cartoon with a transparent solvent accessible surface. CmCBM6 is in complex with cellobiose (VLS) and xylobiose (CFS) displayed as yellow sticks, and a structural calcium displayed as an orange sphere. Adapted from Abbott & van Bueren (2014).

As discussed in the previous chapter, GH98 enzymes with endoxylanase activity possess a conserved CBM family 35 domain directly to the C-terminus of the catalytic domains. Structural studies of a GH98 endoxylanase from human gut symbiont *Bacteroides ovatus* demonstrated close association of this CBM with the catalytic domains (**Figure 4-12A**), such that a construct of the enzyme lacking this CBM could not be recombinantly expressed. This evolutionary conservation indicates high functional importance of this CBM35 domain, indicating that it may play an important role in enhancing catalytic efficiency of complex xylan degradation.

The CBM35 family is distantly related to the CBM6, and members possess a similar β-sandwich fold structure (Abbott & van Bueren, 2014). The first characterised members of CBM family 35 were mannan- and xylan- binding proteins from *Cellvibrio japonicus* mannanase Man5C and xylan-targeting arabinofuranosidase Abf62A respectively (Bolam *et al.*, 2004). Xylan-binding by Abf62A-CBM35 was specific to long-chain undecorated xylans, and was shown to be calcium dependent, a phenomenon that had not been previous observed in CBMs (Bolam *et al.*, 2004), and which is not required for all members of the CBM35 family.

The first structure of a CBM35, Man5C-CBM35, was solved by NMR (243). The protein displayed a β-sandwich fold, with substrate recognition at the VLS, as seen in many family 6 CBMs. Furthermore, unlike in most CBM proteins, this CBM35 underwent a significant confirmational change upon ligand binding, indicating flexibility in the loop regions which determine ligand recognition. The binding site of Man5C-CBM35 is a narrow, extended *endo*- style cleft, typical of a type B CBM, with its curved, narrow confirmation likely dictating specificity for mannans (243).

Subsequent studies led to the functional and structural characterisation of CBM35 modules with diverse ligand specificities, including uronic acids from a variety of glycan types (urono- type CBM35) (Montanier *et al.*, 2009; Sainz-Polo *et al.*, 2014), backbone of β1,4-mannans (manno- type) (Tunnicliffe *et al.*, 2005), α-D-galactose sidechains of various oligo- or polysaccharides (galacto- type) (Correia *et al.*, 2010; Matsuyama *et al.*, 2020) or maltoligosaccharides (gluco- type) (Suzuki *et al.*, 2014; Okazawa *et al.*, 2015; Light *et al.*, 2016; Fujimoto, Kishine, *et al.*, 2017; Fujimoto, Suzuki, *et al.*, 2017). Studies show that the predominant binding site of the CBM35 family is the VLS.

Phylogenetic analysis of the CBM35 family by Correia *et al.,* (Correia *et al.*, 2010) (**Figure 5-2**) demonstrated that CBM35 sequences clustered into four main subgroups, which, based on the distribution of CBM35s with known specificities, enabled functional prediction of these subgroups as gluco- manno-, galacto- and urono- binding proteins.

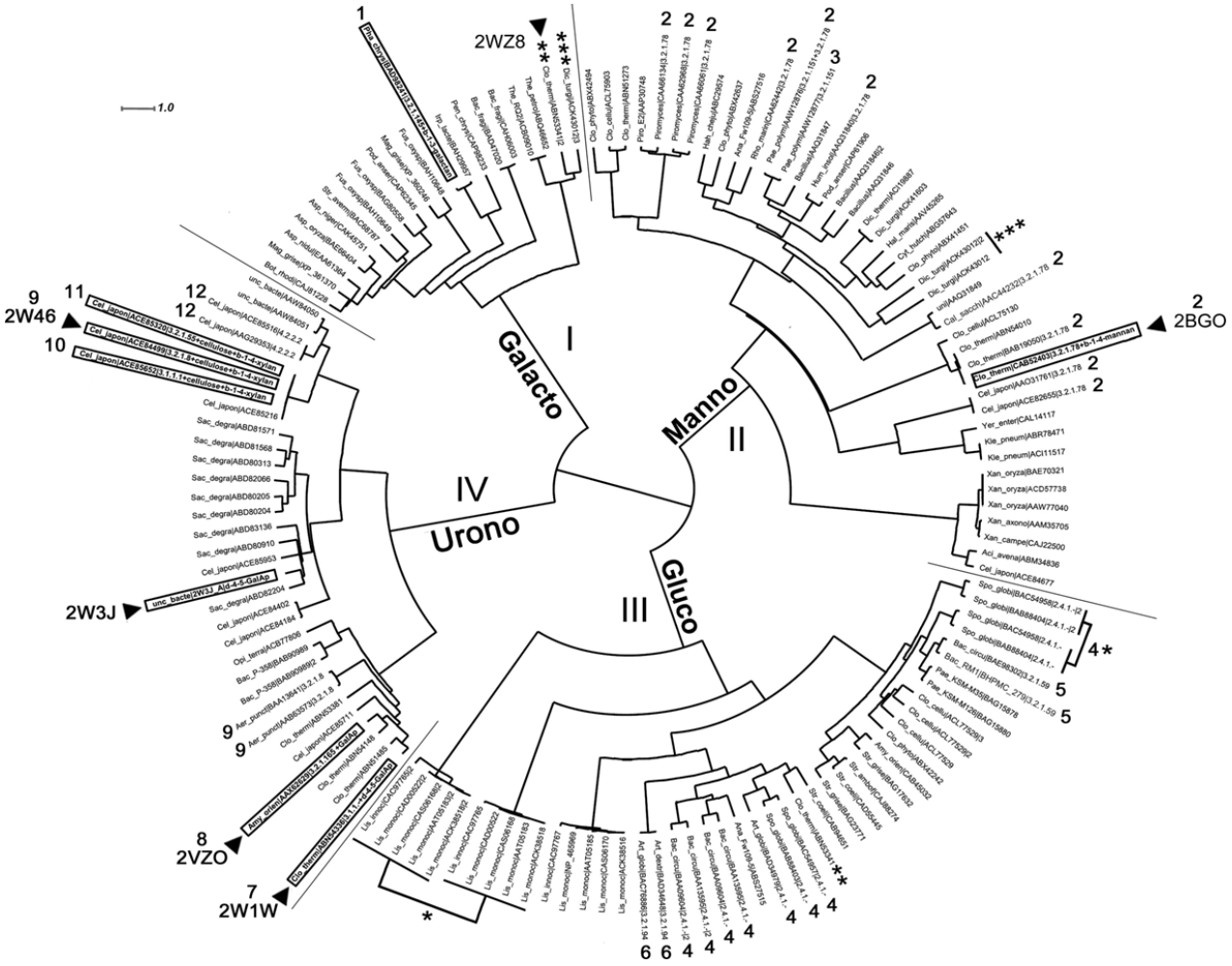


Figure 5-2: Phylogenetic analysis of the CBM35 family. Clustering of 139 sequences demonstrated four main subgroups. Mapping of characterised proteins, shown in black boxes, allowed functional prediction of four main subgroups, each with different binding specificities. From Correia *et al.* (2010).

When reviewing the literature, it is apparent that uronic acid binding members of the CBM35 family are calcium dependent, with the calcium ion playing an important role in ligand recognition (Montanier *et al.*, 2009), however galacto-, manno- and gluco-type CBM35 binding does not seem to involve calcium. Although urono- type CBM35s largely show conservation of key binding site residues (**Figure 5-3A**), they are components of enzymes where the cognate catalytic modules display divergent functions, including acetyl esterases, xylanases, glucosaminidases and pectate lyases. Consequently, different urono- binding CBM35 modules can recognise a

divergent range of uronic acid containing polysaccharides including homogalacturonan, glucuronoxylans and rhamnogalacturonan.

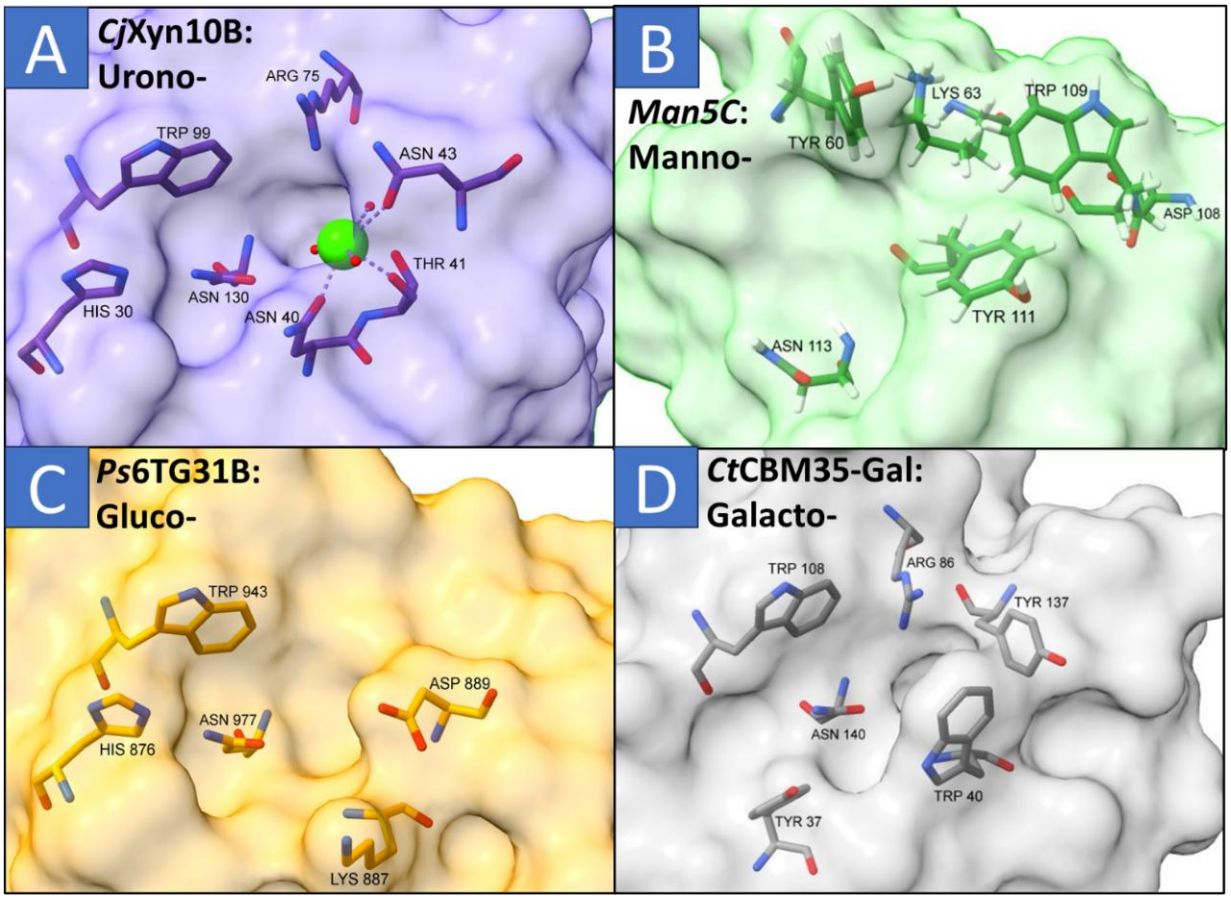


Figure 5-3: Ligand binding sites of characterised CBM35 proteins with different substrate specificities. Four main binding specificities have been identified within the CBM35 family, each associated with characteristic binding residues and binding site morphologies. Structurally characterized examples for each specificity type is shown; A) Uronic acid binding, PDB 2W46 (Montanier *et al.*, 2009), B) Mannan binding, PDB 2BGO (Tunnicliffe *et al.*, 2005), C) Glucan binding, PDB 5X70 (Fujimoto, Suzuki, *et al.*, 2017) and D) Galactose binding, PDB 2WZ8 (Correia *et al.*, 2010).

There is some conservation in key CBM35 binding moieties between subgroups, in particular the tryptophan and asparagine residues in the positions equivalent to Trp99 and Asn130 respectively in *Cj*Xyn10B (Montanier *et al.*, 2009) (**Figure 5-3A, C+D**). However, no conservation is seen in the mannan-binding Man5C-CBM35, in which different binding site residues form an elongated type B CBM binding channel (**Figure**

5-3D), conserved amongst manno- type CBM35s (Ghosh *et al.*, 2014; Couturier *et al.*, 2013).

Interestingly, a fungal CBM35 has been identified with ability to bind galactan chains, which shows significantly different binding site morphology to any of these known subgroups (Matsuyama *et al.*, 2020). Furthermore, the CBM35 from Abf62A, one of the first characterised members of the family (Bolam *et al.*, 2004), binds to unsubstituted xylan chains, so unlikely fits into any of the main subgroups described above. These examples indicate further, as yet undescribed, diversity within the CBM35 family, with respect to substrate specificity and binding site morphology.

Although, as shown, CBM35 proteins normally recognise ligand at the VLS, there is evidence that other sites may function as a secondary, or alternative binding sites in some proteins, as seen in glucuronic acid binding Xyn30D-CBM35 (Sainz-Polo *et al.*, 2014), and in isomaltoligosaccharide binding *Ps*CBM35 (Fujimoto, Kishine, *et al.*, 2017). GlcA-soaked crystal structure of Xyn30D-CBM35 demonstrated two observed GlcA binding sites, at the VLS and at another site (242). Despite this, there are no published examples of CBM35 proteins binding ligand at the CFS.

Overall, members of the CBM35 family display divergent binding activities, with the majority recognising manno-, gluco-, galacto- or urono- type ligands at the VLS. There is evidence of further activities within the family, and potential binding at other sites is largely unexplored.

The conservation of CBM35 domains to the C-terminal of GH98 catalytic modules demonstrates evolutionary importance of these modules, and as such this chapter aims to explore the binding activity and specificity of the CBM35 from *B. ovatus* GH98 endoxylanase, hereafter named Bo98-CBM35. Here, we show that Bo98-CBM35 is a

functional carbohydrate-binding module, with unusual ligand specificity for arabinose-containing polysaccharides. Through structural and biochemical characterisation, we show that Bo98-CBM35 likely binds at the CFS, a novel location within the family, and that this binding site is conserved amongst CBM35s with cognate GH98 catalytic modules.

5.2 Results

5.2.1 Structural analysis of Bo98-CBM35

In order to assess the functional characteristics of Bo98-CBM35 we aimed to solve its structure by X-ray crystallography. Due to instability of the recombinantly expressed His tagged CBM35 module, crystallisation attempts were unsuccessful. Therefore, we utilised the crystal structure of the CBM35 from the full length Bo98 endoxylanase to improve understanding of structure-function relationships of these conserved CBM35 domains within GH98 endoxylanase sequences.

As with other members of the CBM35 family, Bo98 CBM35 possesses a jelly-roll type β-sandwich fold structure made up of 2 anti-parallel β-sheets with 4 and 5 β-strands respectively (Correia *et al.*, 2010; Sainz-Polo *et al.*, 2014) (**Figure 5-4**). Within Bo98, the CBM35 domain appears tightly associated with the catalytic domain. If this is the true association between catalytic and CBM domains, this may indicate coevolution of the two domains. However, the long linker sequence suggests that this may be an artefact of crystallography, despite Alphafold2 predicting the same relative positioning of the 2 modules.

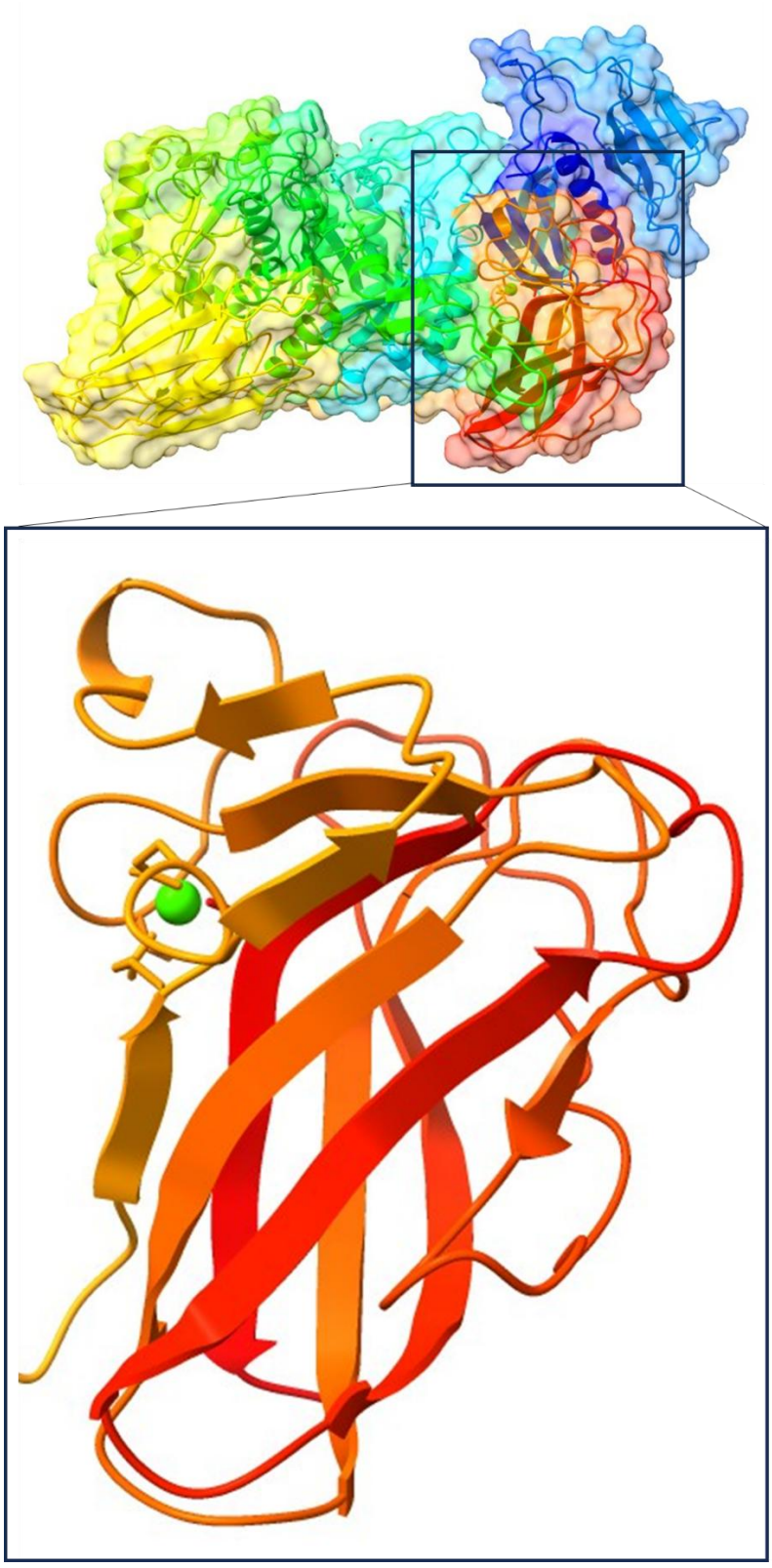


Figure 5-4: Cartoon representation of Bo98-CBM35 structure showing position relative to GH98 catalytic domains. Top panel shows the overall structure of Bo98 colour ramped from blue (N terminus) to red (C terminus). C-terminal CBM belonging to family 35 is shown below in orange/ red possessing the canonical β-sandwich fold structure of the CBM35 family. Structural calcium ion is shown in green,

Bo98-CBM35 contains one calcium ion with a structural role (**Figure 5-4**), common to many CBM35 families, but does not possess a second calcium ion with a role in substrate recognition near the VLS, as is seen in some members of the CBM35 family (**Figure 5-3A**). The loss of this second calcium ion, which is directly involved in uronic acid binding, indicates this CBM35 is unlikely to recognise glucuronic acid side chains of complex GAXs, the substrate for the cognate GH98 catalytic domain.

The positioning and interactions with the structural calcium ion in Bo98-CBM35 shows complete conservation with other members of the CBM35 family, as shown by overlay with the structurally characterised Xyn30D-CBM35 (Sainz-Polo *et al.*, 2014) (**Figure 5-5**) with coordination to Glu 817, Glu 819 and Asp 940 side chains, as well as the carbonyl groups of Thr 842, Gly 845 and Asp940 (Correia *et al.*, 2010).

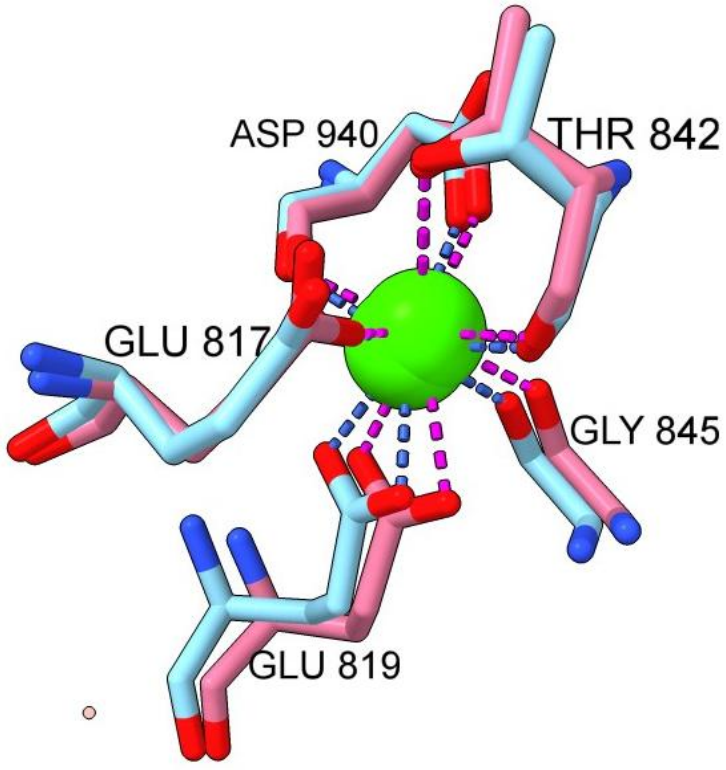


Figure 5-5: Coordination sphere of the structural calcium ion in Bo98-CBM35. Residues interacting with the structural calcium ion of Bo98-CBM35 are shown in pink. Overlay with Xyn30D-CBM35 (cyan) (Sainz-Polo *et al.*, 2014) demonstrates complete conservation in metal binding residues, and in orientation of the coordinate bonds between the metal and amino acids.

Attempts to crystallise His-tagged Bo98-CBM35 were unsuccessful, and due to lack of ligand density surrounding the CBM35 domain when solving the structure of Bo98 following soaks with xylooligosaccharide ligand, structural studies were of limited use to investigate the ligand binding capacity of the GH98-associated CBM35.

5.2.2 Recombinant expression of Bo98-CBM35

To investigate the functional role of this CBM, Bo98-CBM35, the sequence encoding the CBM35 was cloned into a GST-fusion or His-tag vector using primers described in **Supplementary Table 8-1**. Initial studies with the His-tagged CBM revealed the construct was unstable and so all following studies focussed on the GST-CBM35 construct. The binding capabilities of recombinantly expressed GST-tagged Bo98-CBM35 were analysed by isothermal titration calorimetry (ITC). Affinity gel chromatography was also utilised to assess binding to a range of polysaccharides; however, no retardation was observed on the gels and so method was deemed ineffective for studying Bo98-CBM35.

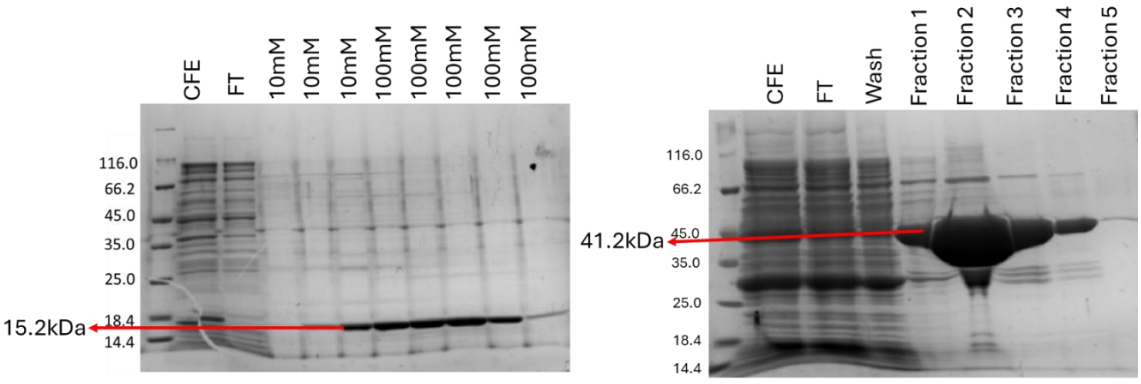


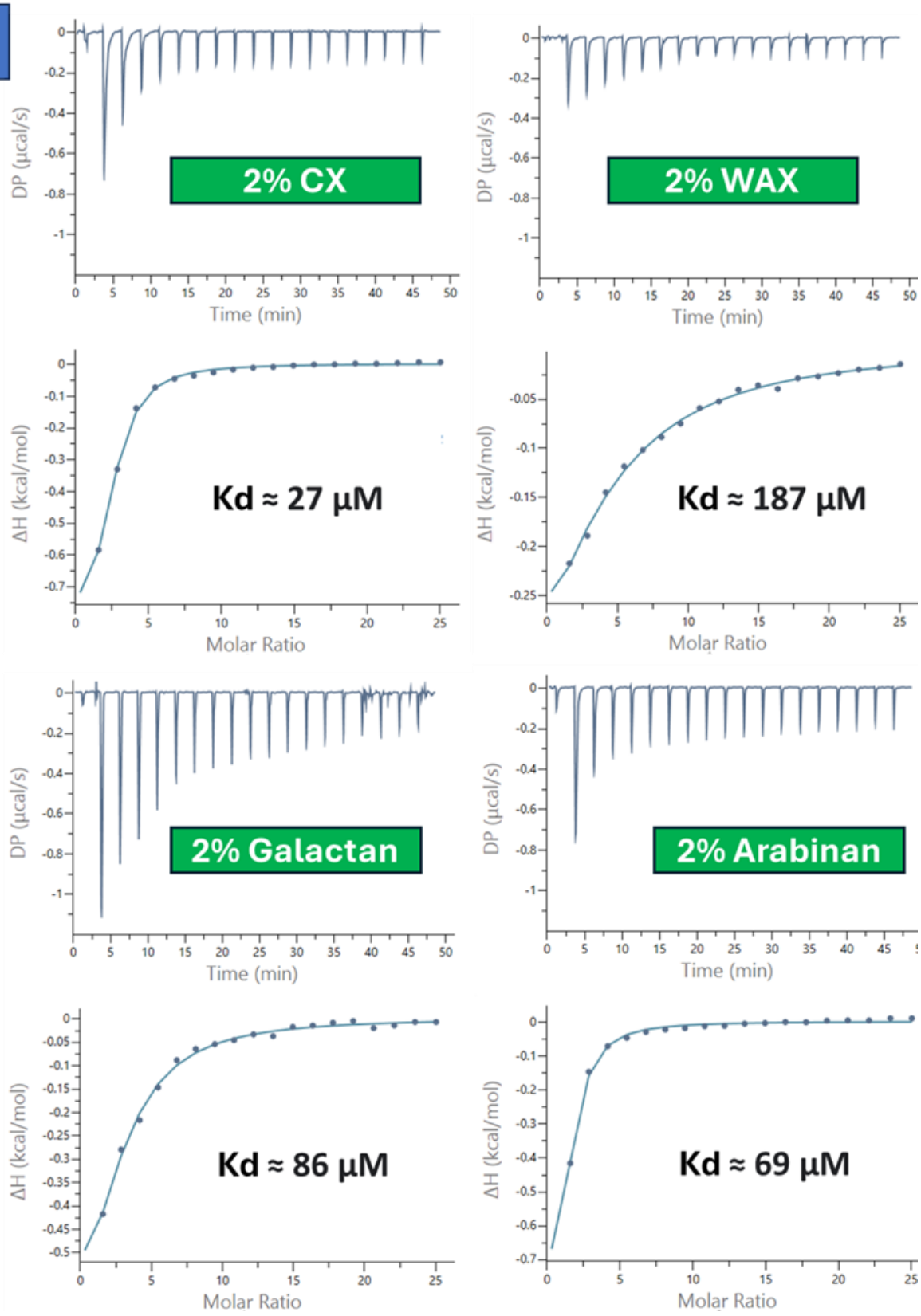
Figure 5-6: Purification of recombinantly expressed His- and GST- tagged Bo98-CBM35. Bo98-CBM35 constructs were cloned and recombinantly expressed in E. coli Tuner cells. Proteins were subsequently purified and analysed by SDS PAGE with a 14.4 to 116kDa molecular weight ladder. For His-tagged Bo98-CBM35 (left) cell free extract (CFE), flow through (FT) and 8 imidazole-eluted fractions were collected: 3 x 10mM, followed by 5 x 100mM. For GST-Bo98-CBM35 (right) five reduced glutathione fractions were eluted. His-tagged and GST-tagged constructs of Bo98-CBM35 had molecular weights of 15.2 and 41.2 kDa respectively.

1.2.3 Ligand specificity of Bo98-CBM35 determined by ITC

The binding capabilities of recombinantly expressed and purified GST-tagged Bo98-CBM35 were analysed by ITC. Bo98-CBM35 was shown to bind to CX with high affinity, but also to WAX (**Figure 5-7**) with a slightly lower affinity. Strong binding was also seen to arabinose containing polysaccharides; arabinan and potato galactan, but not to arabinose or xylose monosaccharides (**Figure 5-7**). Interestingly, no binding was seen to BWX glucuronoxylan, or to OSX with relatively few side chains, indicating that arabinose side chains may be a recognition factor for CBM35 binding. ITC data was fit using ligand is in cell. Due to difficulties with the prediction of molar concentration of polysaccharide substrates, this was assumed to be 13mM based on a concentration of 20 mg mL^{-1} and a rough estimate of 10 available sugars on the chain at the binding site, considering spatial restraints for binding. Although stoichiometry is not accurate, this provided a better estimate of the dissociation constant (Kd). In the present study, due to lack of availability of decorated xylo-oligosaccharides, binding of CBM35 to oligosaccharides was not assessed, however binding affinity of purified Bo98-generated CX degradation products to Bo98-CBM35 would be an interesting area of study in the future.

These data suggest that Bo98-CBM35 is a functional carbohydrate-binding module, recognising a range of polysaccharides whose common feature appears to be the presence of arabinose side chains (**Figure 5-7**). Arabinans and galactans are also found in the plant cell wall, likely in close spatial proximity to CX, such that CBM-mediated binding to these polysaccharides may bring the catalytic domain of Bo98 closer to its substrate, hence improving catalytic efficiency.

A


DP (µcal/s)
Time (min)
2% CX
2% WAX
ΔH (kcal/mol)
Molar Ratio
Kd ≈ 27 µM
Kd ≈ 187 µM
2% Galactan
2% Arabinan
Kd ≈ 86 µM
Kd ≈ 69 µM

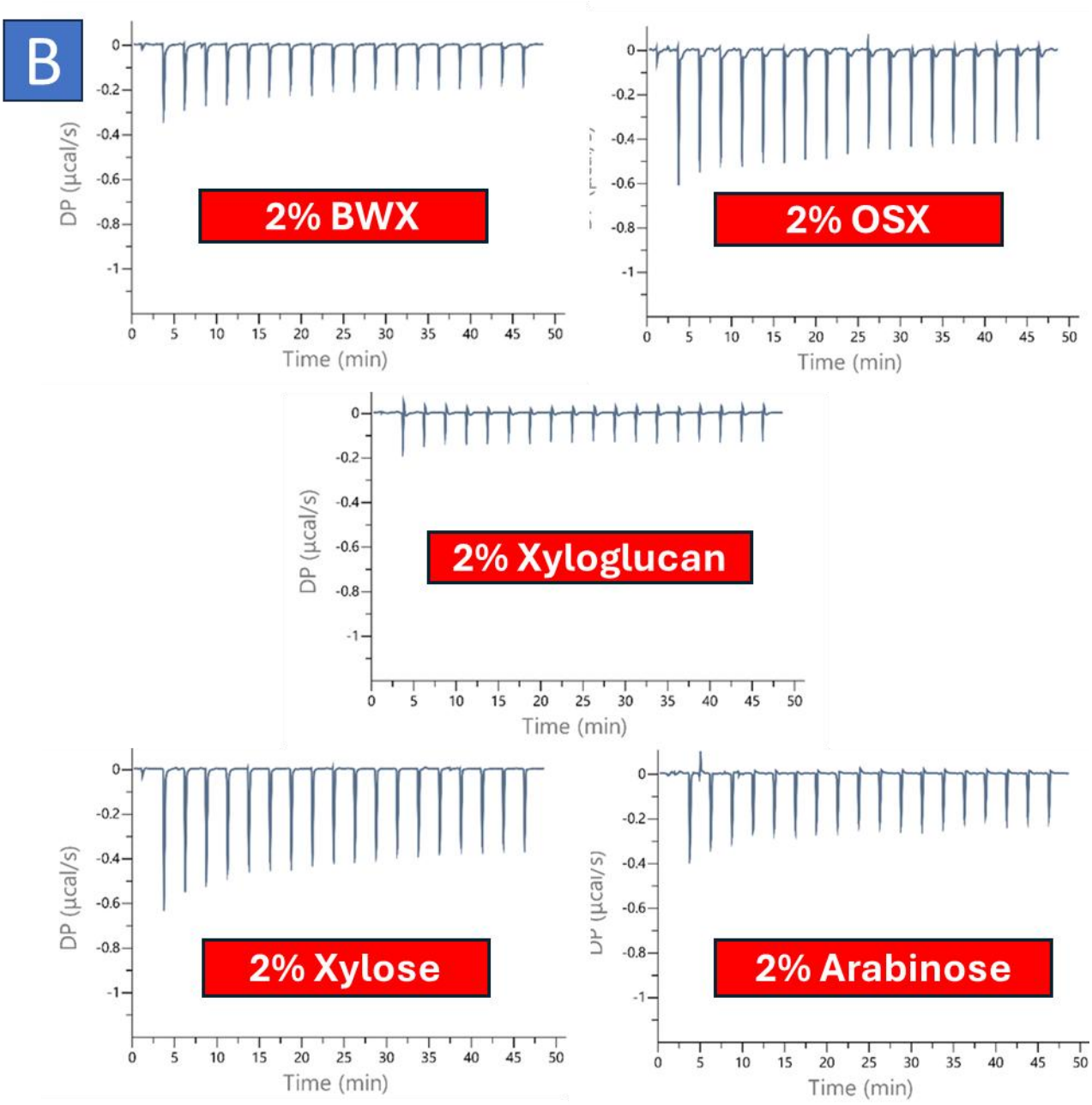


Figure 5-7: Isothermal Titration Calorimetry (ITC) analysis of GST-tagged Bo98-CBM35 against a range of carbohydrates derived from plant cell wall material. Bo98-CBM35 was at a concentration of 100 uM in 50 mM HEPES buffer at pH 7.5. Ligands were at 20 mg mL^{-1} in the same buffer. (A) Binding was seen to corn xylan, wheat arabinoxylan, arabinan and galactan. (B) No binding, or very weak, binding was seen to BWX (glucuronoxylan) or OSX (sparsely decorated xylan chain), xyloglucan or arabinose and xylose monosaccharides. The top frames of each panel show the raw heats, and the bottom show the integrated peak areas fitted to a one site model using MicroCal PEAQ-ITC Analysis Software v1.41. ITC was performed at least in duplicate for each ligand, depending on availability of ligand.

5.2.4 Analysis of the location of the carbohydrate binding site of Bo98-CBM35

As no ligand-bound structure of Bo98-CBM35 could be gained, either from full length Bo98, or the Bo98-CBM35 domain itself, little was understood regarding binding interactions of Bo98-CBM35 and its carbohydrate ligands. This presents a challenge to ascertain the sites at which this functional binding protein recognises ligand, as well as the precise nature of ligand specificity.

As discussed previously, most members of the CBM35 family bind exclusively at the canonical site of the VLS. Notably, when analysing this typical CBM35 binding site in Bo98-CBM35 there is a lack of any exposed aromatic residues, in particular tyrosine or tryptophan, as often required for CBM-ligand recognition (20) (**Figure 5-8A**). Furthermore, structural analysis of the position of the canonical VLS CBM35 binding site in Bo98-CBM35 showed no conservation of the expected binding residues for any known binding activity within the CBM35 family (urono-(6), galacto-(15), manno-(10) or gluco-(5) binding) (**Figure 5-8B,C,D,E**), indicating that Bo98-CBM35 may possess novel binding residues at the VLS site and/or bind ligand at an alternative site.

Furthermore, overlay of the VLS site in Bo98-CBM35 with Alphafold2 models of the CBM35 domains belonging to the other characterised GH98 endoxylanase enzymes, Pa98, Cj98 and Rc98 (**See Chapter 4**), demonstrates limited conservation in site morphology (**Figure 5-9**). As the CBM35 domain is conserved amongst all GH98 enzymes with predicted endoxylanase activity we would expect conservation in function amongst these CBMs, and hence similarities in binding site morphology.

These differences provide further evidence for an alternative binding site in Bo98-CBM35.

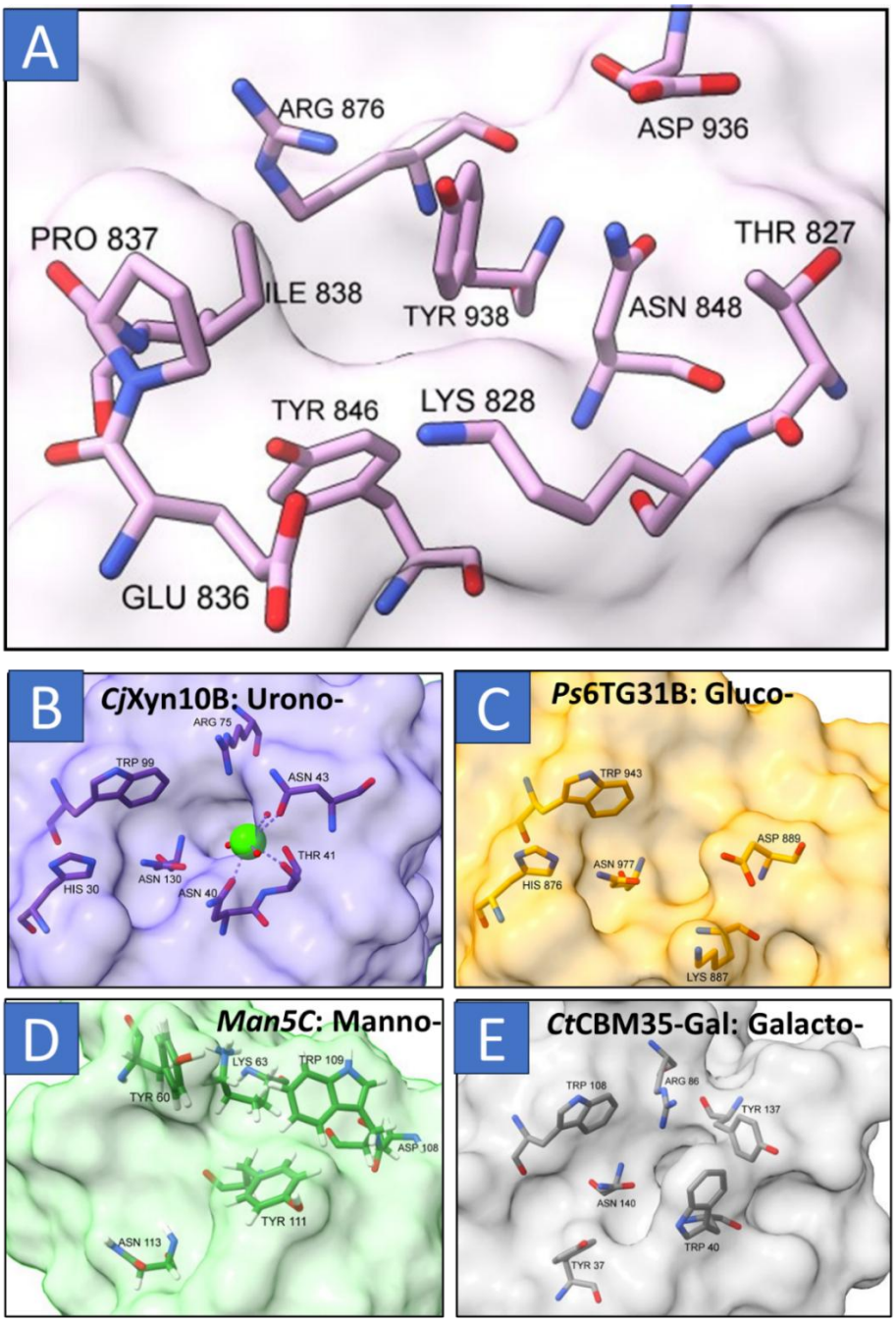


Figure 5-8: Residues forming the canonical VLS binding site position in CBM35 proteins. (A) Position of the VLS in Bo98-CBM35. Analysis of surface residues at that site shows lack of expected binding residues or surface morphology, and tyrosines are not exposed enough to stack with sugar rings. (B, C, D, E) Four known binding specificities have been characterised within the CBM35 family (Montanier *et al.*, 2009; Fujimoto, Suzuki, *et al.*, 2017; Matsuyama *et al.*, 2020; Bolam *et al.*, 2004), each associated with specific binding residues and morphologies. This region of Bo98-CBM35 does not possess any conservation with any of these other characterised binding sites.

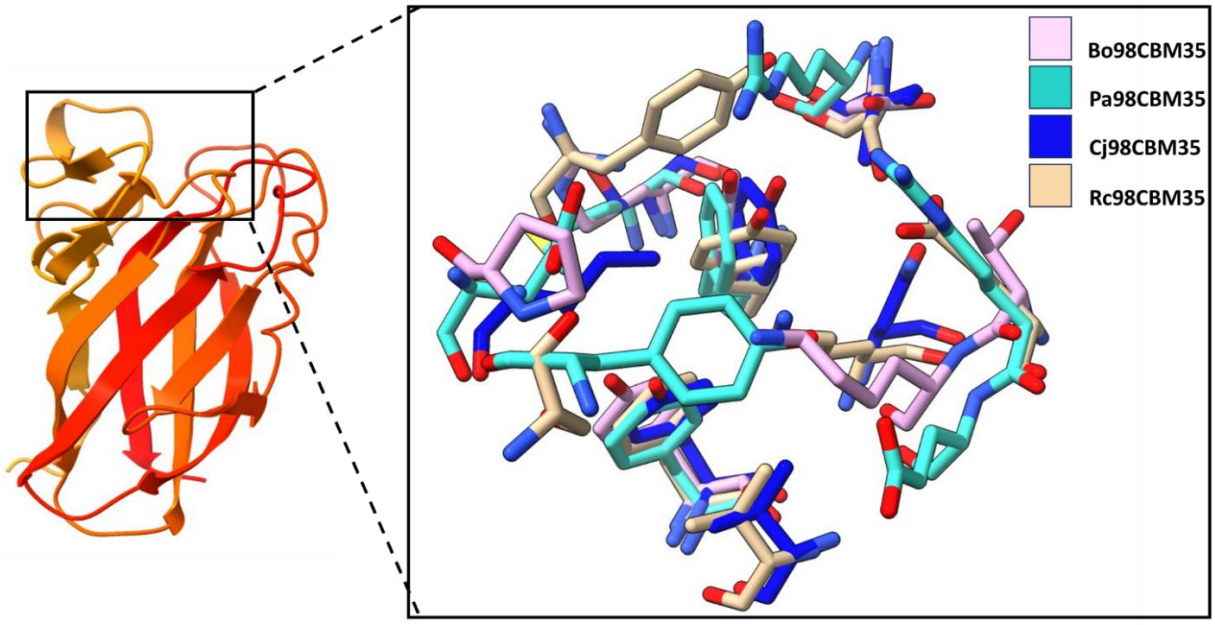


Figure 5-9: Surface residues of CBM35 domains from known GH98 endoxylanases at the VLS. Overlay of surface residues at canonical ligand binding site for CBM35 structures from proven GH98 endoxylanases, Bo98, Rc98, Pa98 and Cj98 shows complete lack of residue conservation suggesting that this site is not important in ligand binding by GH98 coupled CBM35 proteins. Rc98, Pa98 and Cj98 structures were predicted using Alphafold2 (Bryant *et al.*, 2022).

Data thus far suggest that Bo98-CBM35 does not recognise ligand at the VLS, however there is some evidence of alternative, or secondary binding sites, within members of the CBM35 (Sainz-Polo *et al.*, 2014; Fujimoto, Kishine, *et al.*, 2017), or closely related CBM6 families (Pires *et al.*, 2004). Consequently, the ligand recognition site may be elsewhere in the Bo98-CBM35 structure. Based on overlay with other CBM35 and CBM6 structures, 3 potential alternative binding sites were identified (**Figure 5-10**). Site A is the canonical VLS whereas site B is the secondary binding site of *Ps*CBM35, where Fujimoto *et al.* found isomalto-oligosaccharides bound (Fujimoto, Kishine, *et al.*, 2017). Site C is the CFS, and although there is not yet any evidence for the CBM35 family binding at this site, it forms a secondary binding site for *Cm*CBM6, where the protein can bind to cello-oligosaccharides (Henshaw *et al.*, 2004; Pires *et al.*, 2004). Glucuronoxylan binding Xyn30D-CBM35 also possesses a

secondary binding site (Sainz-Polo *et al.*, 2014), and this region has been designated site D (**Figure 5-10**).

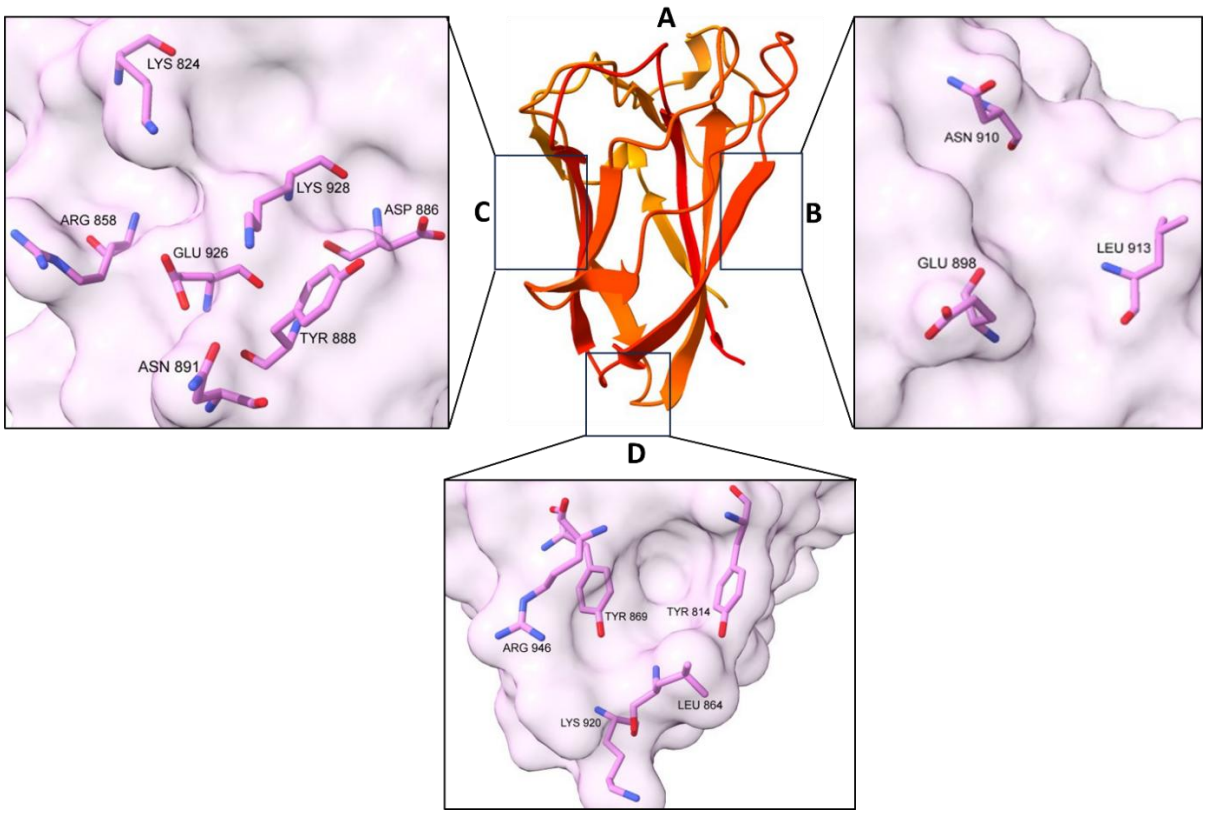


Figure 5-10: Potential alternative binding sites of Bo98-CBM35. Alternative, or secondary, binding sites to the canonical VLS (site A) have been noted within members of the CBM35, or closely related CBM6 families. Overlay of Bo98-CBM35 with four such CBM proteins demonstrates overall conservation in structure and three putative alternative binding sites in Bo98-CBM35, designated here as B to D. PsCBM35 (Fujimoto, Kishine, *et al.*, 2017) binds isomalto-oligosaccharides at a site referred to here as site B. Xyn30D-CBM35 (Sainz-Polo *et al.*, 2014) structural studies showed glucuronic acid in complex with a second binding site, site D. CmCBM6(Pires *et al.*, 2004) possesses a secondary glucan binding site, denoted as site C, also known as the CFS.

Despite evidence for binding at these sites in other CBM35 or CBM6 proteins, none of these sites appear to be classical CBM binding sites, with multiple exposed aromatics creating a planar surface for stacking interactions with sugar ligands (69). Furthermore, Bo98-CBM35 does not show any conservation of known binding residues with the CBM shown to bind at any of these sites (**Figure 5-11**). This indicates

that if the sugar binding site of Bo98-CBM35 is either site B, C or D, ligand recognising likely occurs via a different interaction to those already described in the literature.

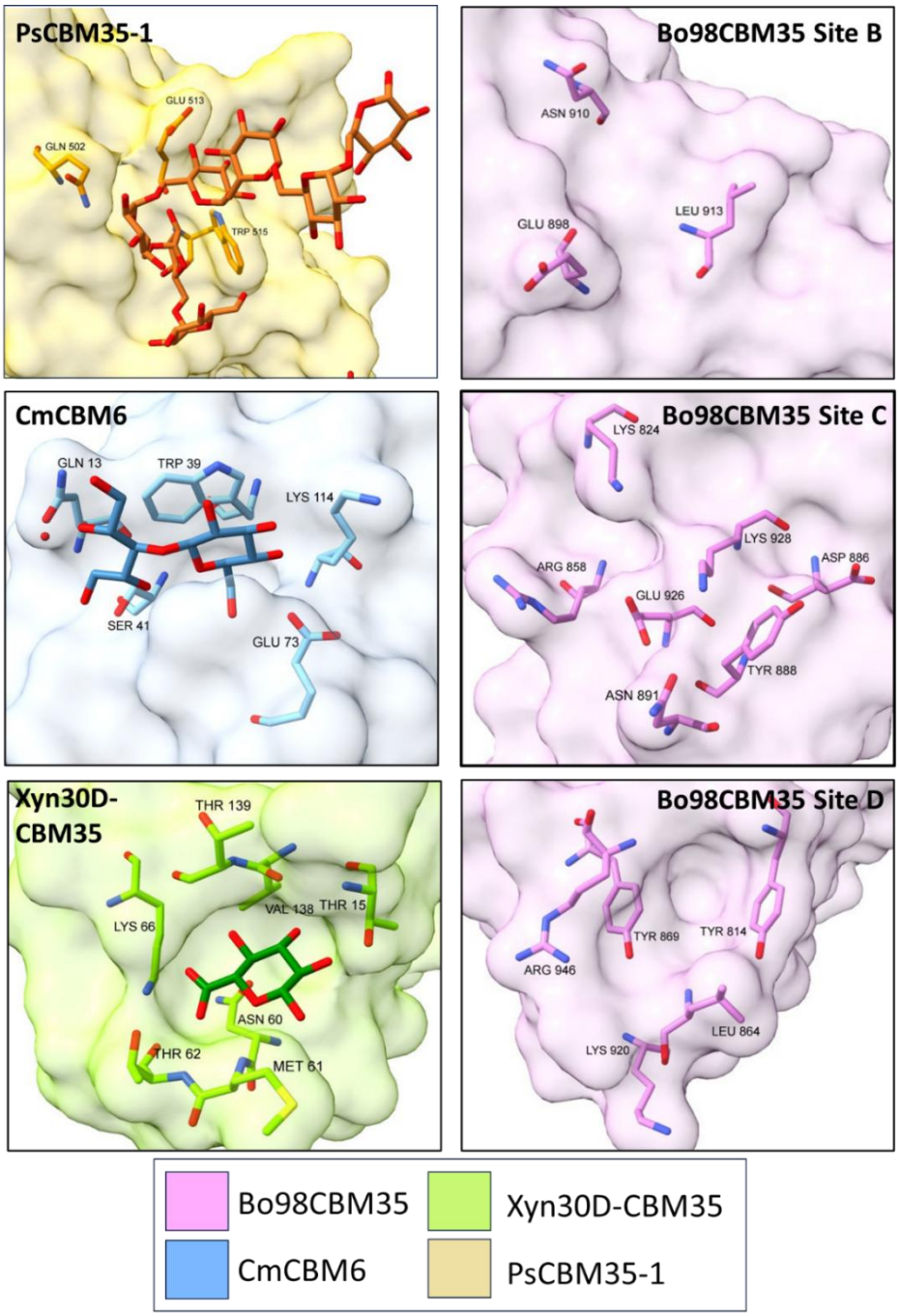


Figure 5-11: Structural comparison of potential Bo98-CBM35 binding sites B, C and D with CBM35 or CBM6 proteins shown to bind ligand at that site. No conservation in site morphology or binding residues was found between Bo98-CBM35 (pink) and characterised binding protein at any site. PsCBM35-1 from *Paenibacillus* sp. 598K in complex with isomaltooligosaccharides is shown in yellow (Fujimoto, Kishine, *et al.*, 2017), Xyn30D-CBM35 from *Paenibacillus barcinonensis* in complex with glucuronic acid (Sainz-Polo *et al.*, 2014) is shown in green, and CmCBM6 from *Cellvibrio mixtus* with cellobiose (Pires *et al.*, 2004) in blue.

When analysing the structure of each site in Bo98-CBM35 for potential binding activity, it is not entirely clear which interactions could lead to recognition of a branched polysaccharide ligand. Site B is sparse in terms of surface residues and the surface morphology suggests that this site is highly unlikely to possess ability to bind to a glycan ligand (**Figure 5-11**). Site C possesses a surface-exposed tyrosine residue (Tyr888) which could be involved in ligand interactions, and surface facing residues Asn891, Arg858 and Lys824 could potentially be involved in polar interactions with the sugars, however from the surface morphology it is not clear how this site could dictate specificity for a branched oligosaccharide chain. In site D there is a small, shallow pocket, formed by two tyrosine residues 869 and 814, which could potentially be involved in exo-type C CBM binding (**Figure 5-11**). Although we could not identify Bo98-CBM35 binding to xylose or arabinose via ITC, it may be that this binding is very weak (**Figure 5-7**).

To investigate the location of the ligand binding site on Bo98-CBM35, we performed alanine site directed mutagenesis of potential binding residues in Bo98-CBM35 sites A to D. 100-200 µM mutant Bo98-CBM35 was titrated against 2% WAX and CX, and impact of single point mutants at each site on binding was assessed (**Figure 5-12**). Five mutants from site A were recombinantly expressed and used in ITC: K828A, T827A, E836A, D936A, N848A; 1 mutant from site B: E898A; 4 mutants from site C: D886A, N891A, K928A and Y888A and 1 from site D; Y814A. Some mutants of Bo98-CBM35 expressed poorly and/ or were relatively unstable, meaning that protein could not easily be concentrated to 200 µM. Therefore, for these mutants, protein was concentrated to as high a level as possible, and ITC was performed with wild-type Bo98-CBM35 at the same concentration for pairwise comparison. ITC data for a

representative mutant from each site are displayed in **Figure 5-12** and remaining data in **Appendix Figure 1**.

When looking at ITC data, it is clear that alanine mutation of residues forming site C, the CFS, has the greatest impact on affinity for WAX and CX. All four site C alanine mutants tested, N891A, D886A, K928A and Y888A caused loss of binding of the mutant protein to WAX (**Figure 5-12C, Supplementary Figure 8-1**), indicating that this site is critical for recognition of AXs. Interestingly, impact of these mutations on binding to CX is not as severe. Heats of binding are considerably lower for these mutants compared to WT and affinity was lower. These data suggest that site C is the primary binding site for Bo98-CBM35.

Alanine mutations at sites B and D had very little impact on binding activity to either substrate (**Figure 5-12B+D**), indicating that Bo98-CBM35 does not recognise ligand at these locations. When analysing ITC data from site A mutants (**Figure 12 A**), the canonical VLS site, some mutants, namely K282A and T287A, had no impact on binding to WAX or CX. However, for mutants E836A, D936A and N848A some differences are seen in the binding profile in comparison to wild-type protein. For these mutants, the binding curve is shallower, indicating a lower affinity for the ligand. For E836A, heats of binding are roughly twice as high for mutant protein compared to WT. These data demonstrate that site A may play some role in ligand recognition but is not critical for binding. Site A may potentially function as a secondary binding site, with limited ability to recognise ligand. Preference for CX binding at site A over WAX may explain why mutation of residues in the primary binding site, site C, prevents binding to WAX but not to CX, however further studies would be required to elucidate this.

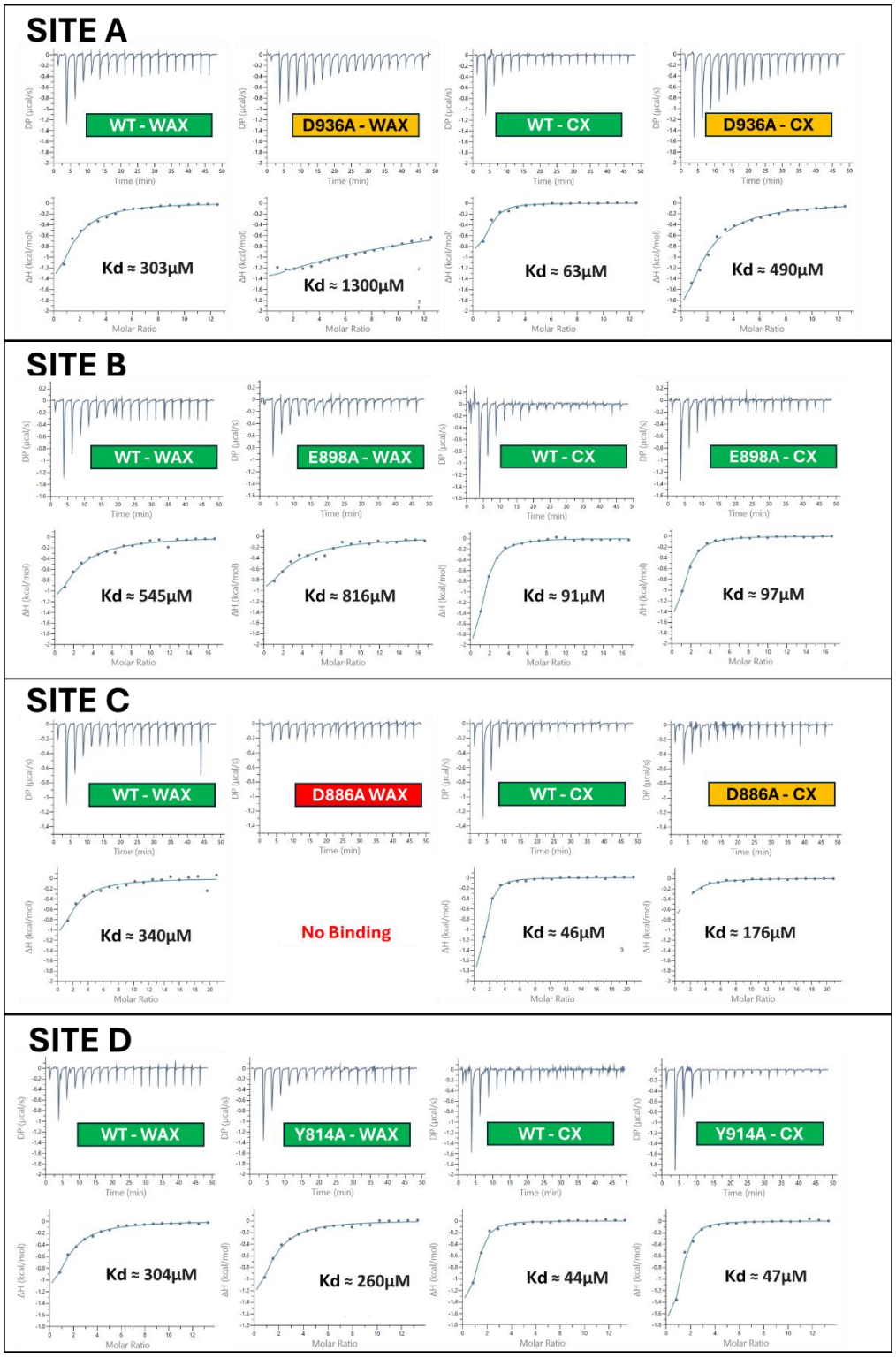


Figure 5-12: Binding of Bo98-CBM35 single point mutants to wheat arabinoxylan (WAX) and corn xylan (CX). Impact of 12 single point mutants from 4 potential binding sites, A to D, on binding activity was studied, examples from each site are displayed here. A) Site A mutant D936A, B) site B mutant T827A, C) site C mutant D886A and D) site D mutant Y914A. The top frames show the raw heats, and the bottom show the integrated peak areas fitted to a one site model using MicroCal PEAQ-ITC Analysis Software v1.41.

When the impact of alanine mutants on WAX and CX binding is mapped onto the structure of Bo98-CBM35 (**Figure 5-13**), it is clear that any disruption to site C has a major impact on binding ability, further demonstrating the importance of this site in ligand recognition.

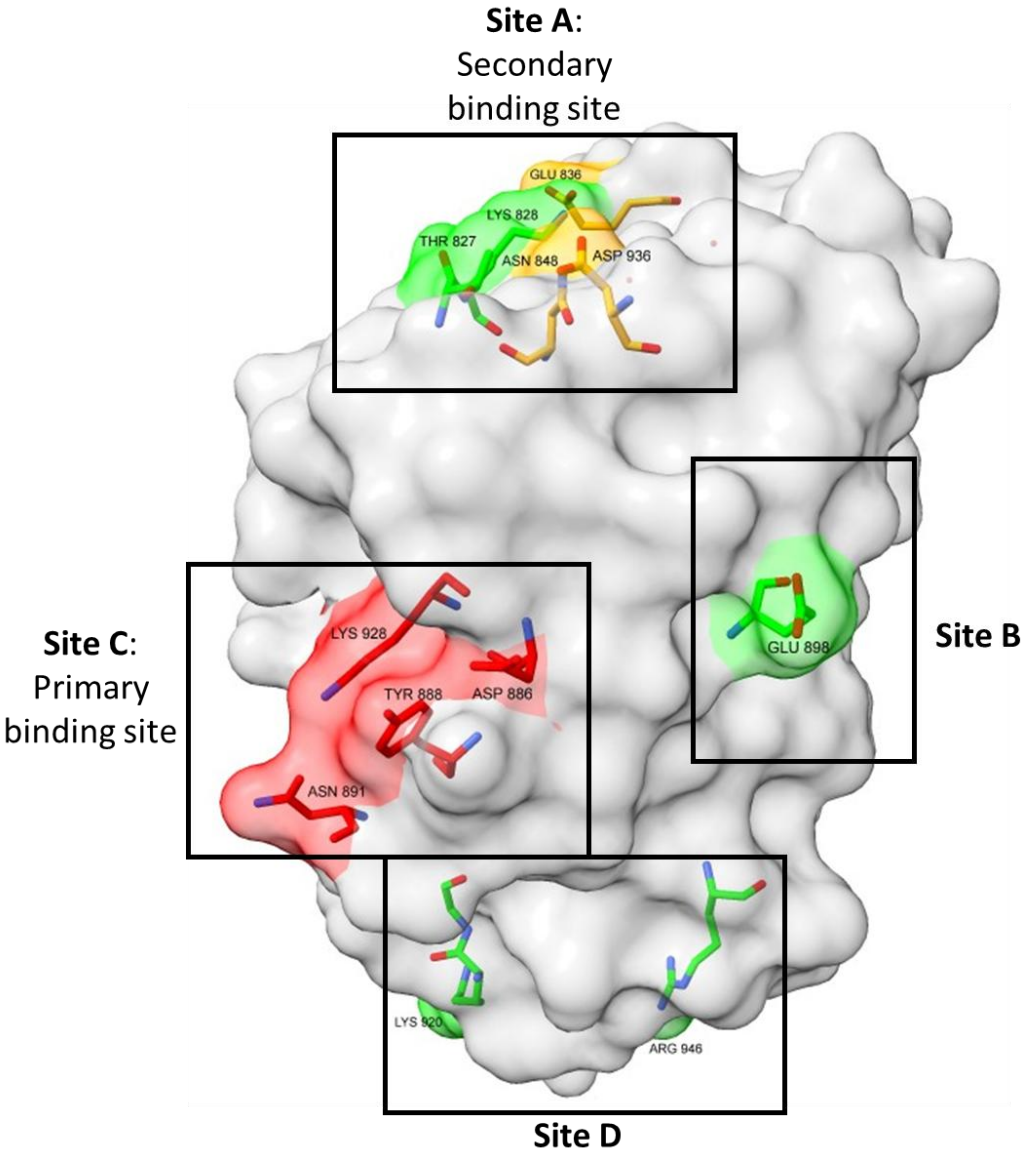


Figure 5-13: Structure of Bo98-CBM35 showing the primary ligand binding site. The 12 residues shown in potential binding sites A-D were individually mutated to alanine and the binding activity of single point mutant proteins was assessed by ITC. Alanine mutation of residues shown in green had no impact on binding to WAX or CX, those in orange had a marginal impact, and those in red had a severe impact on binding affinity.

Based on the hypothesis that site C is the primary binding site of Bo98-CBM35, we looked further into the structure of this site, in particular its relationship with the catalytic domain of Bo98. Site C is located in close proximity to the Bo98 catalytic domain, albeit remote from the active site of the enzyme (**Figure 5-14A**).

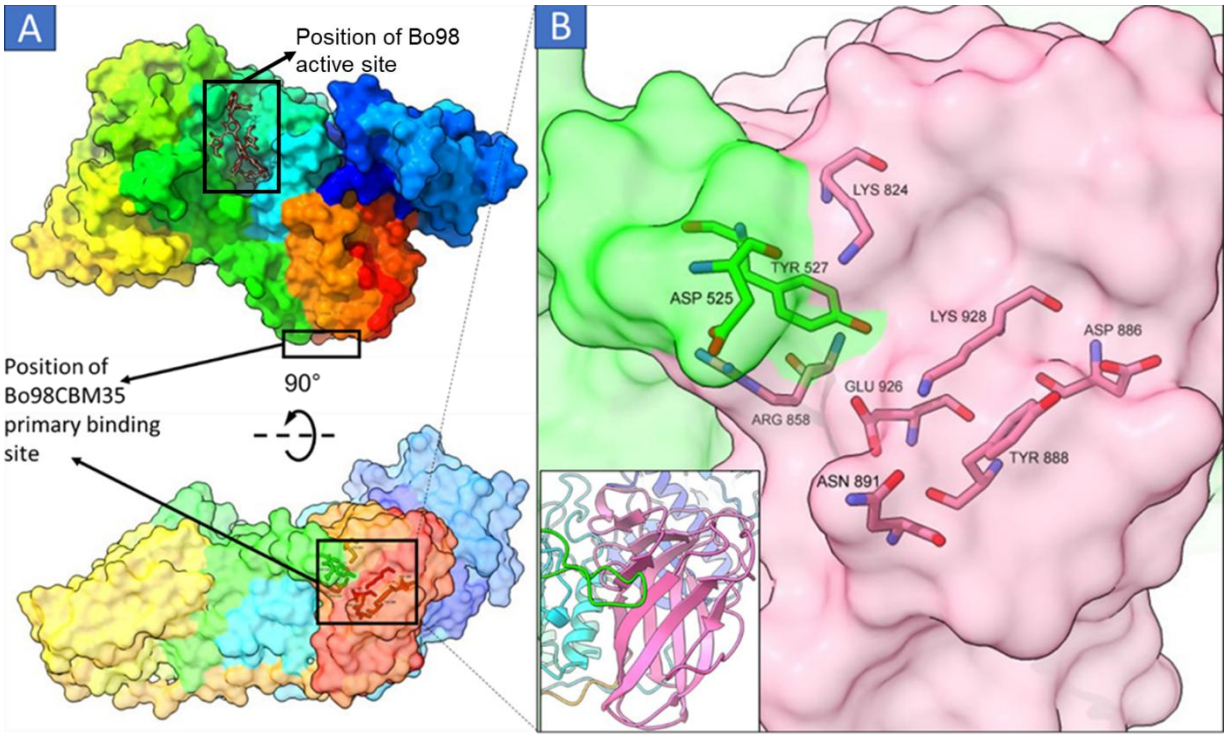


Figure 5-14: Positioning of the primary binding site of Bo98-CBM35 relative to the Bo98 active site. A) Structure of Bo98 showing the positioning of the CBM binding site, which is spatially distant to the active site. Xylooligosaccharide is shown bound to the Bo98 active site. Structure is colour ramped from blue at the N terminal to red at the C terminal. Bo98-CBM35 is shown in orange/red. B) The binding site of Bo98-CBM35 (pink) is closely associated with a loop from the Bo98 catalytic domain (green), such that catalytic domain residues Asp525 and Try527 may contribute to ligand binding.

The catalytic domain of Bo98 contains a loop which folds over the CBM domain and interacts closely with the CBM near to site C (**Figure 5-14B insert**). This loop possesses two surface exposed residues, Asp525 and Try527, which are in close proximity to this CBM binding site such that they may play a role in ligand recognition and binding (**Figure 14B**). Interestingly, Asp525 from the GH98 catalytic domain and

Asn891 from the CBM35 domain, appear to act together to form a channel which could be involved a role in the recognition of long polysaccharide chains. Despite this close association, as shown via ITC, Bo98-CBM35 alone functions as a binding protein, without its cognate GH98 catalytic domain. Although tyrosine and aspartate residues in this GH98 domain loop may support ligand binding, they are clearly not essential. However, it is hypothetically possible that these residues may play a role in defining substrate specificity, such that Bo98-CBM35 may possess slight differences in binding specificities as part of the entire enzyme, compared to those seem for the CBM35 alone.

The spatial divide between the Bo98 active site and the position of the primary binding site of Bo98-CBM35 (**Figure 5-14A**) suggests that the full-length enzyme is capable of concurrent interaction with multiple polysaccharide chains.

To further assess the importance of this primary ligand binding site in Bo98-CBM35, we examined this same site in Alphafold2 models of CBM35 domains from GH98 enzymes Pa98, Cj98 and Rc98. Unlike the VLS, this CFS site is largely conserved amongst GH98-coupled CBM35 domains, indicating functional importance. Interestingly, this site C is highly conserved in Rc98 and Cj98 (**Figure 5-15A**), indicating similarities in function, whereas Pa98 is more divergent in this location, lacking the tyrosine residue corresponding to Y888 in Bo98-CBM35. Interestingly, the asparagine residue at position 891, whose mutation to alanine appeared to prevent WAX and CX binding (**Figure 5-12**), is not conserved in these other proteins, with a glycine in the equivalent position. It is possible that mutation of this residue may have caused structural disruption of the predicted binding site. Furthermore, Tyr527 is completely conserved in each GH98-associated CBM, indicating that this may be important for binding, even if the CBM35 displays binding activity without it. Overall,

based on conservation at this site it is likely that this is also the primary binding site for Cj98, Rc98 and Pa98, and that they will recognise branched polysaccharides via a similar mechanism. However, further studies would be necessary to ascertain their binding activities.

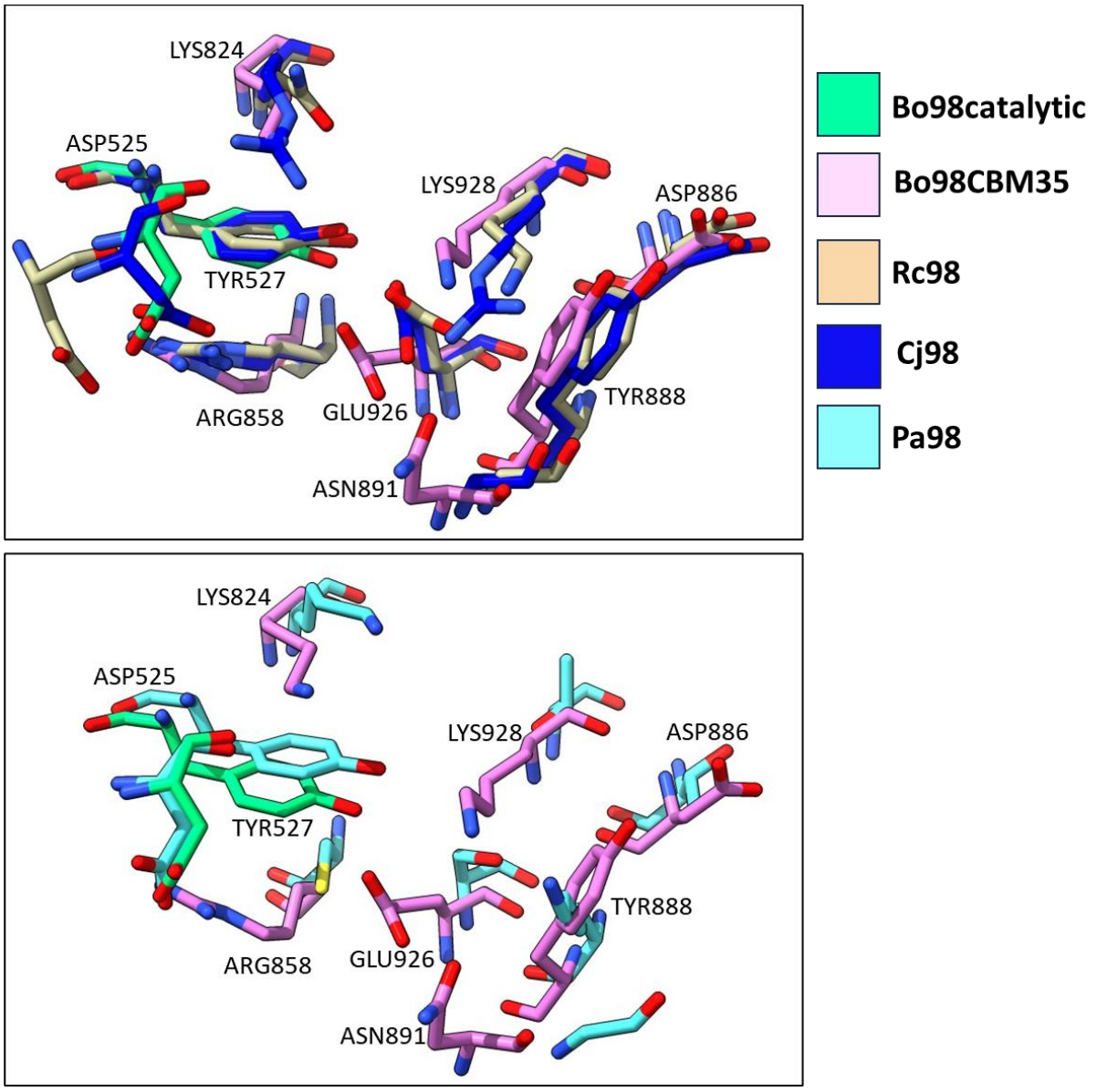


Figure 5-15: Conservation of residues at the CFS in CBM35 domains from other known GH98 endoxylanases. Overlay of key binding residues from Bo98-CBM35 primary binding site (residue numbers labelled) with CBM35s from other characterised GH98 endoxylanases shows overall conservation suggesting that this site is important in ligand binding in these other CBM35s. GH98 structures were predicted using Alphafold2 (Bryant *et al.*, 2022). Top panel shows a high level of conservation of Bo98-CBM35 with CBM35s from Rc98 and Cj98 at this site, whereas the bottom panel shows that Pa98-CBM35 is slightly more divergent.

Based on the data presented in this chapter we suggest that Bo98-CBM35 binds to arabinose-decorated polysaccharides at a novel site for the CBM35 family, corresponding to the CBM CFS. Binding at this location may also involve interactions between ligand and residues in a loop of the Bo98 catalytic domain, which closely interacts with the CBM. Although this is not essential for activity, it may influence ligand specificity. Structural comparisons with other CBM35 domains from GH98 proteins suggests that this binding site and mechanism is likely conserved amongst GH98-associated CBM35 modules.

5.3 Discussion:

Here, we characterise the binding site location and ligand specificity of a CBM35 associated with the GH98 endoxylanase from *B. ovatus*. The CBM35 is a polyspecific family of binding proteins, which display diverse binding profiles, albeit with a conserved β-sandwich fold structure. The presence of a CBM35 domain appears to be a defining feature of GH98 endoxylanases, indicating their functional importance, and as such this study provides insight into the role of these conserved domains.

CBM35 domains linked to non-GH98 xylan-active enzymes have been structurally and/ or biochemically characterised previously, namely Xyn30D-CBM35 which is attached to a GH30 glucuronoxylanase (Sainz-Polo *et al.*, 2014), CjXyn10B-CBM35 GH10 endoxylanase (Montanier *et al.*, 2009) and Abf62A-CBM35-xylan-targeting arabinofuranosidase (Bolam *et al.*, 2004). Studies to date indicate that Xyn30D-CBM35 and CjXyn10B-CBM35 domains bind to uronic acids, allowing recognition of glucuronoxylan target substrates, whereas Abf62A-CBM35 targets xylan chains with low levels of substitution such as OSX (Bolam *et al.*, 2004; McCartney *et al.*, 2006).

Bo98-CBM35 does not possess similar specificities to any of these characterised CBM35 proteins, instead binding to arabinose-containing polysaccharides including WAX, CX, galactan and arabinan (**Figure 5-7**). This presents a novel activity within the CBM35 family.

As Bo98-CBM35 is a component of a plant cell wall degrading enzyme, it likely plays a role in targeting the cognate GH98 domain to regions of the plant cell wall substrate that are accessible to biological degradation. Arabinose-decorated xylans, galactans and arabinans within the plant cell wall are all relatively accessible to members of the HGM, meaning that Bo98-CBM35 likely targets its cognate hydrolase to areas of the plant cell well which can be easily degraded. Conservation of these CBM35 domains within all GH98 endoxylanase enzymes indicated evolutionary importance, suggesting that the binding activity of this CBM35 domain may play a role in enhancing catalytic efficiency of the GH98 against highly decorated xylans. As described in the previous chapter, recombinant construct of Bo98 lacking the CBM35 domain could not be expressed, likely due to the tight interactions between the CBM and cognate catalytic domains meaning that protein without the CBM may have great structural instability. Again, this provides evidence for the evolutionary importance of the CBM35 domain for GH98 endoxylanase enzymes.

Multiple studies have shown that CBMs from different families, including 4, 6, 9, 22 and 35 possess structural calcium ions at a remote location to the binding site, the position of which is conserved in Bo98-CBM35 compared to other members of the CBM35 family (**Figure 5-4 + Figure 5-5**). This structural metal ion likely plays a role in protein stability, and studies suggest that removal of this ion may increase susceptibility to proteolytic degradation (Czjzek *et al.*, 2001). Whilst some members of the CBM35 family, in particular uronic acid targeting CBMs, possess a calcium

cofactor essential for ligand recognition (Montanier *et al.*, 2009), this is not apparent from the Bo98-CBM35 structure, indicating that Bo98-CBM35 binds via a calcium independent mechanism, divergent from characterised uronic acid binding CBM35s associated with xylan-targeting enzymatic domains (Montanier *et al.*, 2009). Future studies will involve ITC experiments with the addition of EDTA to ascertain whether calcium ions are important for ligand binding.

Here we show that the CFS, site C, of Bo98-CBM35 is likely the primary binding site, based on mutant ITC data and site morphology and conservation. This is a novel binding location for the CBM35 family, potentially indication that Bo98CBM35 should be classified within a separate subfamily to previously characterised CBM35. This site was the only one where single point alanine-mutation of residues had a dramatic impact on the ability of the CBM35 to bind WAX and CX, completely preventing any WAX-binding activity (**Figure 5-12**). Despite SDM ITC data suggesting that this site is critical for xylan binding activity, the surface architecture is not typical of a Type B CBM with an endo- style elongated cleft forming the binding site (**Figure 1-15B**). In a typical Type B CBM this cleft would normally function in the recognition of soluble polysaccharide chains, such as those of xylan, arabinan or galactan (Boraston *et al.*, 2004). Without a ligand-bound Bo98-CBM35 structure, it is difficult to predict the orientation or position of ligand in this site. When considering Bo98-CBM35 with its cognate catalytic domain, the Bo98 loop, which closely interacts with the CBM35, helps to form a short channel, defined by residues Asp525 and Asn891 (**Figure 5-14**), which likely functions as a binding cleft. However, these data show that Bo98-CBM35 alone, lacking Asp525 from the GH98 domain, is a functional CBM, meaning that this channel is not critical for ligand binding. Furthermore, Asn891 is not conserved amongst the other GH98 endoxylanase enzymes structurally examined in this study,

Cj98, Rc98 and Pa98, instead replaced by a glycine residue unlikely to play any role in ligand recognition (**Figure 5-15**). Based on these findings, the role of this Asp-Asn channel in Bo98-CBM35 ligand recognition is unclear, making it challenging to predict protein-ligand interactions. However, due to the close interaction between CBM35 and GH98 domains at this site, we hypothesise that the conserved tyrosine and asparagine residues from the GH98 domain loop may play a role in modulating ligand specificity. This could explain why we see the somewhat unexpected binding to arabinan and galactan in Bo98-CBM35 alone, perhaps when the GH98 and CBM35 domains are considered together we would see more specific Bo98-CBM35 binding to AXs and GAXs.

Crystallisation studies using CX or WAX derived oligosaccharides to gain a ligand-bound complex would be necessary to truly understand how the carbohydrate ligand is orientated within this predicted binding site, and how the residues interact with the polysaccharide to determine substrate specificity.

This apparent role of the CFS as the primary binding site of Bo98-CBM35 is in contrast with all other structurally characterised CBM35s who primarily bind at the VLS, with secondary binding sites elsewhere (Sainz-Polo *et al.*, 2014; Fujimoto, Kishine, *et al.*, 2017). Although evidence suggests that this is the true binding site of Bo98-CBM35, it is possible that mutants at this site lead to unstable or poorly folded protein, leading to a non-active phenotype of recombinant protein. Circular dichroism could be used to determine structural stability and to determine whether the lack of binding is due to knocking out the binding residues.

The CFS site in CBM proteins is on the surface of the β-sheet structure (**Figure 5-1**), with side chains of amino acids within the sheet structure forming the binding site. This

provides limited scope for variability in binding site architecture, partly due to limited flexibility of functional groups such as aromatic rings of tyrosine and tryptophan residues. As such, there is limited diversity in the types of ligands recognised at the CFS (Abbott & van Bueren, 2014). Examples of xylan chain recognition at the CFS of CBMs are limited, however this is seen in Xyn10B-CBM15, capable of accommodating various side chains present on different xylan classes, although binding was mediated via two tryptophan residues in the binding groove, absent in Bo98-CBM35, indicating key differences in ligand binding mechanisms (Szabó *et al.*, 2001).

Based on the mild impact of mutation of some site A (VLS) residues on CX and WAX binding (**Figure 5-12**), we hypothesise that this may act as a secondary binding site in Bo98-CBM35. Preference for CX over WAX at this site may explain why disruption of the CFS primary binding site completely abolishes binding to WAX, but some CX binding is still seen.

There are some examples where both VLS and CFS sites are functional binding sites in the same CBM protein, for example CmCBM6-2 (Pires *et al.*, 2004), in which the 2 sites can accommodate cello-oligosaccharides, despite differences in ligand specificity between them. Furthermore, binding of CmCBM6-2 to insoluble cellulose requires synergistic binding activity between these 2 sites. Overall, these sites allow variation in ligand recognition. If Bo98-CBM35 also displayed differential binding at the VLS and CFS sites, this could potentially allow for the recognition of a greater variety of hemicellulolytic, or even cellulolytic substrates which may co-exist with complex GAX in the plant cell wall material, hence enhancing the proximity of the GH98 catalytic domain to its substrate. Further experiments involving testing the impact of mutation of residues at both sites by ITC would be required to further investigate this hypothesis.

Despite this, when examining the morphology of site A, there are no residues obviously involved in ligand binding (**Figure 5-8A**), and overall architecture does not fit that expected of a type A, B or C CBM. Based on this it is highly unlikely that this site will be able to bind ligand with a high affinity. Furthermore, there is no conservation at all of this site in Cj98, Pa98 or Rc98 CBM35 predicted structures (**Figure 5-9**), such that if this VLS is involved in binding it may be unique to Bo98-CBM35 and not conserved in other GH98 coupled CBM35 domains.

When considering the role of Bo98-CBM35 in ligand recognition and function of the congate GH98 domain, it is important to consider that this enzyme may not occur in isolation, but rather potentially forms part of a cell-surface utilisome system, similar to that described by White *et al.* (2023). In this context, Bo98-CBM35 may function as a secondary binding site for the recognition of GAX, alongside the SGBP encoded in *B. ovatus* BoPULXyl-L. Despite this, other GH98 endoxylanases, such as Rc98, from non-Bacteroidota species, will not form part of a utilisome, indicating functional importance of the CBM35 domain regardless of situation.

In conclusion, we have shown that Bo98-CBM35 is a functional CBM that displays a novel specificity for the CBM35 family, recognising arabinose-decorated polysaccharides at the concave face site on the β-sheet structure of the CBM. We also provide the first structural analysis of a CBM35 associated with a GH98 endoxylanase – a conserved feature of these enzymes, although a ligand-bound complex will be critical in order to truly understand ligand specificity. Furthermore, close proximity of a loop from the cognate GH98 catalytic domain at this CFS site likely impacts on binding affinities and specificities, such that Bo98-CBM35 may possess divergent binding behaviours in the context of the entire Bo98 complex.

Further studies will include crystallisation studies to gain a CBM-ligand bound complex of Bo98-CBM35, and of Bo98 in its entirety. Double single-point mutants of the VLS and CFS sites of Bo98-CBM35 will help elucidate whether any secondary binding activity occurs at the CFS.

6. Investigating the mechanism of dietary fibre breakdown by the keystone cellulolytic human gut symbiont *Ruminococcus champanellensis*

6.1 Introduction

Microbiome- mediated cellulose breakdown in the intestinal systems of herbivorous mammals, in particular ruminants, is significant, and provides an important source of nutrients and energy for the host organism (Russell *et al.*, 2009). Members of the rumen microbiota display divergent mechanisms for cellulose breakdown, including the production of multi-enzyme complexes, known as cellulosomes, by Bacillota including *Clostridia. Ruminococci* and *Eubacteria* spp. (Artzi *et al.*, 2017) and the extracellular excretion of cellulase enzymes by *Fibrobacteres* (Burnet *et al.*, 2015). There is even very limited evidence of PUL-mediated cellulose degradation by uncultured ruminal Bacteroidota, although this is not widely supported (Naas *et al.*, 2014).

On the contrary, cellulose degradation by the HGM is minimal, with relatively few individuals from the industrialised world harbouring any cellulose-degrading bacterial species within their microbiota (Chassard *et al.*, 2010). In fact, for many years it was believed crystalline cellulose was not degraded at all by the HGM, with the first evidence for microcrystalline cellulose utilisation provided in 2003, based on the enumeration of reducing end production during incubation of faecal microbiome samples with cellulose (Robert & Bernalier-Donadille, 2003). *Ruminococcus champanellensis,* the first cellulolytic bacterium from the HGM to be isolated and characterised, (Chassard *et al.*, 2012), possessed cellulolytic (avicel and filter paper) and xylanolytic activities, but was unable to utilise starch. Upon sequencing of the *R.*

champanellensis genome, bioinformatic analysis demonstrated close relation to the cellulolytic ruminal bacterium *Ruminococcus flavefaciens,* and that *R. champanellensis* encoded a wide variety of cellulosomal elements, including cohesion and dockerin containing modules (Ben David *et al.*, 2015), indicating the ability to form cellulosome complexes involved in the degradation of crystalline cellulose. This ability indicates that *R. champanellensis* may represent a keystone species in the human gut (Morais *et al.,* 2016), where degradation of cellulose as part of a whole plant cell wall complex may increase the accessibility of other cell wall hemicelluloses to the HGM. The definition of keystone species are those native taxa which play an especially important role in the stability of the ecosystem. As such, *Ruminococcus champanellensis* may only be considered a keystone species when part of the gut microbiota of individuals consuming an entirely plant fibre-based diet, such as those from hunter-gatherer populations.

Upon solving the genome of *R. champanellensis,* studies began examining for the presence of GH enzymes, the majority of which contain dockerin modules and hence are predicted to form part of a cellulosomal system (Cann *et al.*, 2016; Moraïs *et al.*, 2016; Ben David *et al.*, 2015). Morais *et al.* (2016) studied the activities of some of these dockerin-containing GH enzymes (136), allowing characterisation of enzyme activity. These studies that *R. champanellensis* encodes GH enzymes with wide-ranging activities including cellulose-active endoglucanases and a cellobiohydrolase, as well as many hemicellulases (**Figure 6-1**).

Although Morais *et al.* (2016) tested a large proportion of GH enzymes from *R. champanellensis,* this was not exhaustive. Furthermore, studies did not characterise the products of enzymatic degradation, only quantified the production of reducing ends. As such, there is still a lot unknown about the activities of these enzymes, for

example their product profiles (Moraïs *et al.*, 2016). Despite this, it is clear that *R. champanellensis* possesses enzymatic apparatus capable of cellulose degradation, as well as degradation of a range of plant cell wall hemicellulases, indicating that it may play a key role in the degradation of whole plant cell wall material in the intestinal environment.

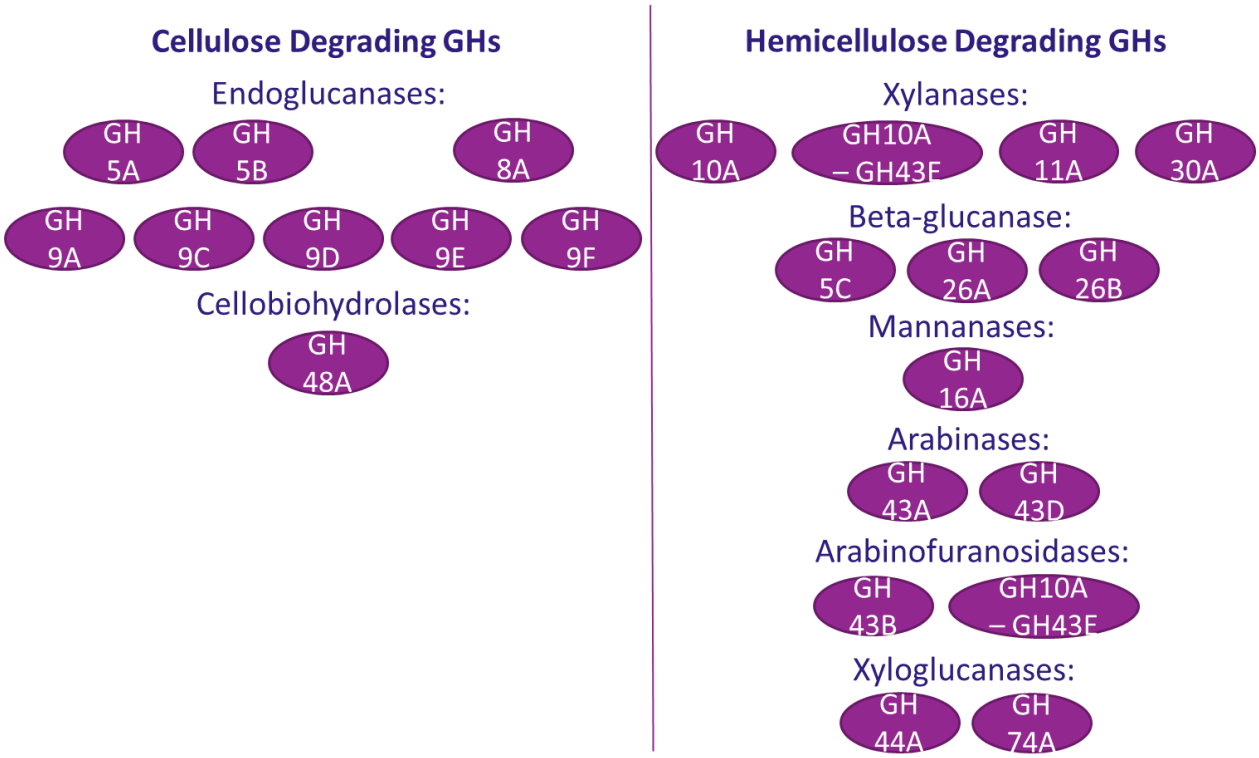


Figure 6-1: Dockerin-containing Glycoside Hydrolase (GH) enzymes encoded by *R. champanellensis* and their characterised activities. A selection of dockerin-containing GHs were cloned and recombinantly expressed, Recombinant protein was assayed for activity against a range of glycans. Activity was confirmed via quanitication of reducing end production. Adapted from Morais *et al.* (2016).

Subsequent studies began to characterise the architecture of cellulosomes produced by *R. champanellensis*. Bioinformatic analysis demonstrated 64 dockerin and 20 cohesin modules within the *R. champanellensis* genome, the parent proteins of all of which except one possess an N-terminal signal peptide, suggesting that they are extracellularly secreted, as expected for all celluosomal enzymes. This includes 25

dockerin-containing GH enzymes. The 20 cohesin domains are found as part of 11 different scaffoldin proteins, named *scaA-scaK* (**Figure 6-2**). Scaffoldin, or *Sca,* proteins are structural proteins, containing cohesin, and often dockerin modules, responsible for dictating the structure of the cellulosome system. Cohesin modules within scaffoldin proteins bind specifically to dockerin domains from other scaffoldins, or those associated with CAZyme modules, allowing the formation of large multi-enzyme complexes. One scaffoldin from *R. champanellensis, ScaE,* possesses a C-terminal putative sortase signal motif, which may be involved in attachment to the cell wall, such that *ScaE* is hypothesised to act as the anchoring scaffolding in *R. champanellensis. ScaI* possesses no dockerin domain and is hypothesised to allow formation of a simple cell-free cellulosome like system, which is released extracellularly to allow access to more spatially distant carbohydrates.

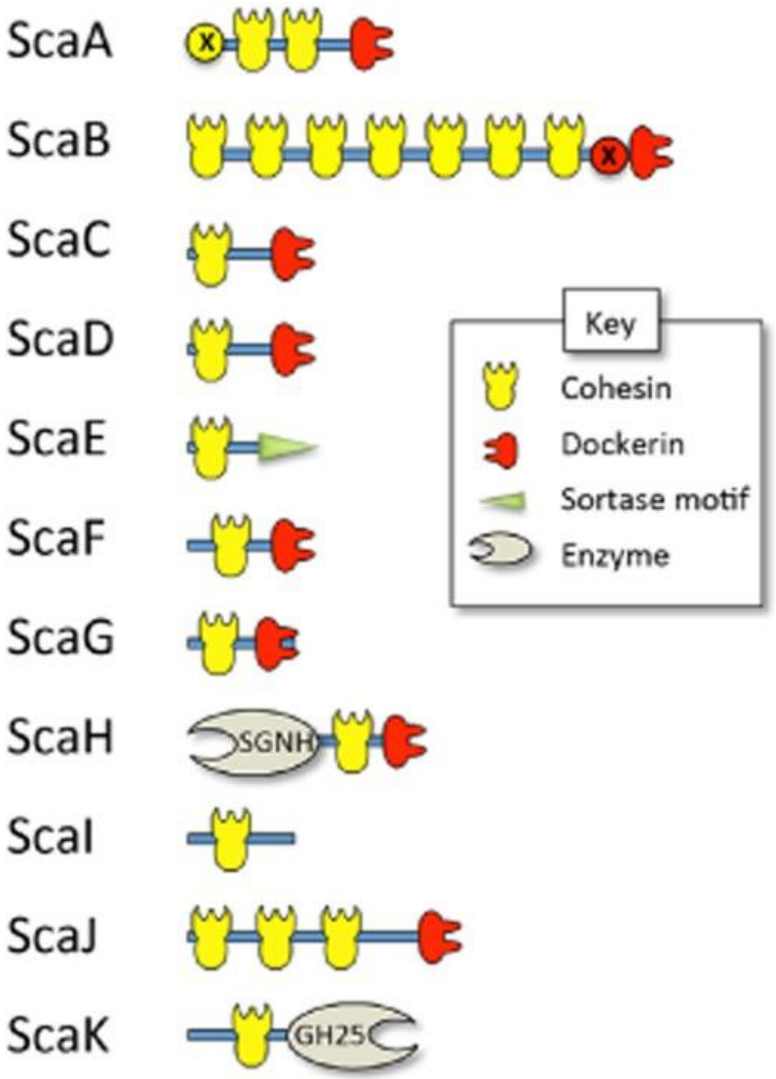


Figure 6-2: Schematic representation of 11 scaffoldin proteins encoded by *R. champanellensis*. Scaffoldin proteins are defined as structural proteins possessing cohesin domains, important in the formation of cellulosome systems. ScaE is hypothesised to act as an anchoring scaffoldin, attaching to the cell wall via the sortase motif. SGNH = a hydrolase type esterase domain. From Ben David *et al.* (2015).

Specific cohesin-dockerin interactions dictate how these scaffoldins are brought together to form the cellulosome system. Dockerin- containing GHs, CBMs or other CAZymes, interact with cohesin domains within the scaffoldin proteins, bringing these components together to form the multi-enzyme complex of the cellulosome. These cohesin-dockerin interactions are specific, such that dockerins with similar sequence similarity usually interact with the same cohesin. Based on sequence conservation, *R. champanellensis* dockerins were clustered into 4 main groups. Via cohesin-dockerin

binding microarrays Ben David *et al.,* studied the interactions between representative dockerin domains from each group, and the 20 cohesin modules. Ultimately, this allowed determination of binding patterns between all cohesin modules belonging to the 11 scaffoldin proteins, and dockerins, and consequently enabled the prediction of *R. champanellensis* cellulosome architecture (**Figure 6-3**).

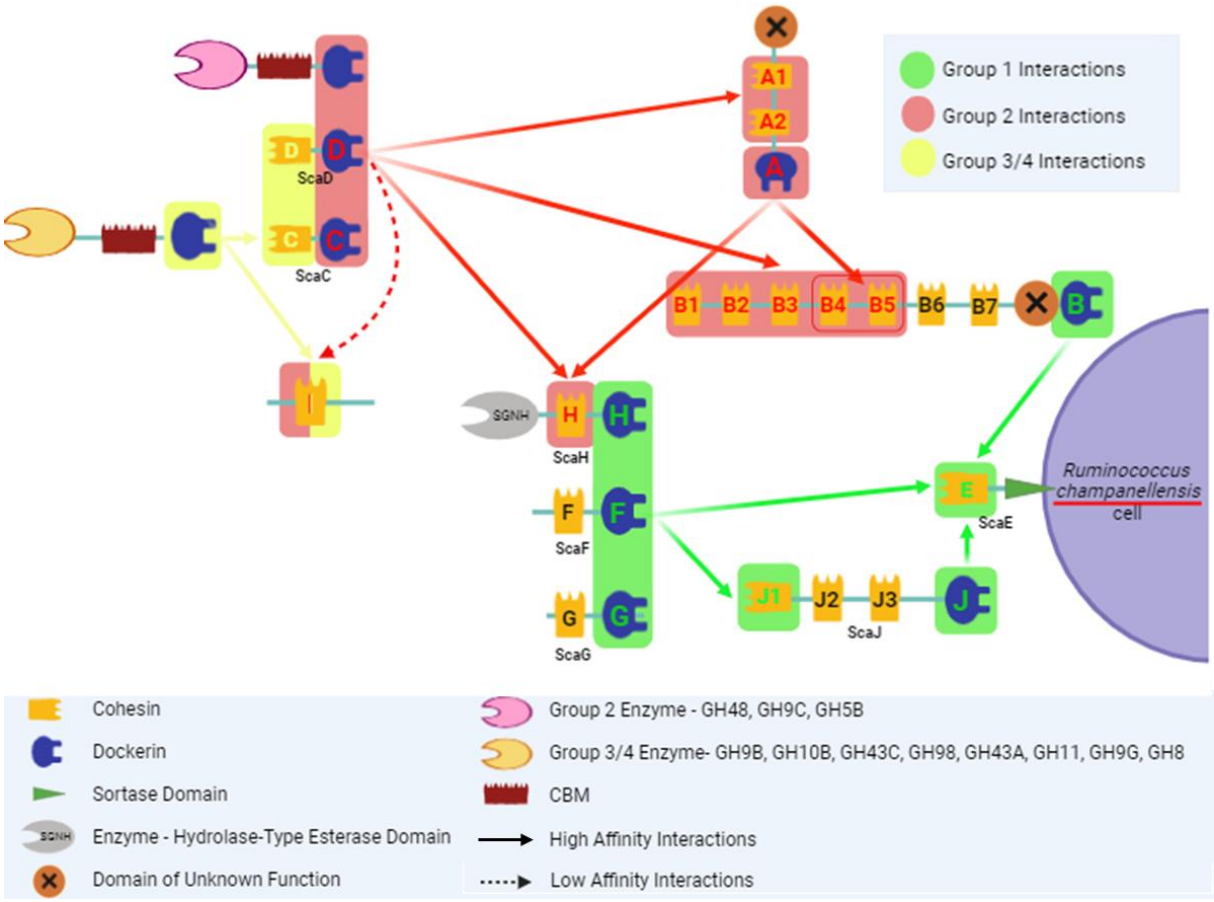


Figure 6-3: Schematic representation of the complex cellulosome system of *R. champanellensis*. *R. champanellensis* is proposed to produce cell free and cell bound cellulosomes, involving up to eleven different scaffoldins, termed ScaA to ScaK. Cell free cellulosomes involve interactions with ScaI, whilst cell bound cellulosomes are attached to the cell surface via anchoring scaffoldin ScaE. Interactions between cohesins and dockerins are grouped into 4 different types, colour coded red, green, and yellow, dictating binding specificity and hence the overall cellulosome structure. Specificities of Sca F and Sca G are as yet undetermined. CAZymes are characterised into group 2 or group 3/4, dependent on the specificity of the integrated dockerin domain. Adapted from Ben David *et al.*, 2015.

Dockerins of different groups, and their consequent interactions with different cohensins, are important in dictating the structure, and hence function of the *R. champanellensis* cellulosome system. Group 1 dockerins were found to interact strongly with the cohesin domain of *ScaE,* termed Group 1 interactions. As *ScaE* is hypothetically involved in membrane anchoring, this suggests that these dockerins are critical for cellulosome assembly, allowing mediation of the attachment of the cellulosome complex to the cell wall through *ScaE.* All cognate proteins of type 1 dockerins seem to have a structural function, rather than enzymatic, indicating that these are adapter scaffoldins, and that *R. champanellensis* is unlikely to produce small, single-scaffoldin cellulosomes. Group 1 dockerins include those from *ScaB, ScaF, ScaG, ScaH* and *ScaJ.*

Group 2 dockerins seems to interact specifically (Group 2 interactions) with cohesins from scaffoldins H and I. Interactions with *ScaI* likely allows the formation of small, free cellulosome complexes, whereas *ScaH* mediated interactions form part of a cell-associated cellulosome. Group 2 dockerins are found in *ScaA, ScaC* and *ScaD*, as well as in GHs belonging to families 48 (cellobiohydrolases), 9 and 5 (cellulose-active endo-glucanases) (Moraïs *et al.*, 2016; Ben David *et al.*, 2015). From this it is clear that *R. champanellensis* can produce cellulosomes of varying sizes and functionalities.

Dockerins belonging to groups 3 and 4 were found to possess the same binding specificities, binding to cohesins from *ScaC, ScaD* and *ScaI.* Parent proteins of group 3 and 4 dockerins include GH enzymes, including those with endoglucanase activity, and hemicellulolytic activity (**Figure 6-3**) (Moraïs *et al.*, 2016). Whilst some dockerins (from GH9, GH10 and GH43 enzymes) interact with all three scaffoldin-contained cohesins, dockerins from GH98 and GH11 only bind *ScaC* and *ScaD.*

This understanding of the varying possibilities for cellulosome assembly, in particular the ability to form cell-bound or free cellulosome systems, is critical to understand how *R. champanellensis* functions in the degradation of complex plant cell wall materials including cellulose and hemicelluloses.

Although *R. champanellensis* is the only isolated cellulolytic species from the HGM to date, there are likely other species which possess similar activities, and produce similar cellulosome structures. In order to understand the prevalence of cellulolytic cellulosome-producing bacteria within the human gut, Moraïs *et al.* (2024) used *R. champanellensis* and *R. flavefaciens* as reference genomes to search for related cellulosome- producing species by identifying key genes involved in cellulose degradation, such as the *scaC* gene, within metagenome-assembled genomes (MAG). This led to the identification of three other candidate cellulolytic species from human-associated MAGs, as well as *R. champanellensis: R. ruminiciens, R. primaciens* and *R. hominiciens.* These species are specific to humans and non-human primates and absent from ruminant samples.

The prevalence and abundance of *R. ruminiciens, R. primaciens* and *R. hominiciens* shows significant variation between populations, at the lowest in industrialised countries with a collective prevalence just 4.6% (**Figure 6-4**). This compares to 20% in geographically diverse rural societies, 21% in hunter gatherer populations and 43% in faecal samples from ancient paleo societies. Furthermore, abundance and heterogeneity in samples from individuals positive for cellulolytic bacteria was significantly lower in industrialised compared to hunter-gatherer and rural populations, as well as paleo samples. These data suggest that these cellulolytic species were previously globally widespread and abundant, however decreases in dietary fibre intake, as well as external factors such as antibiotic consumption means that they are

now found in a limited proportion of the worldwide human population (Moraïs *et al.*, 2024).

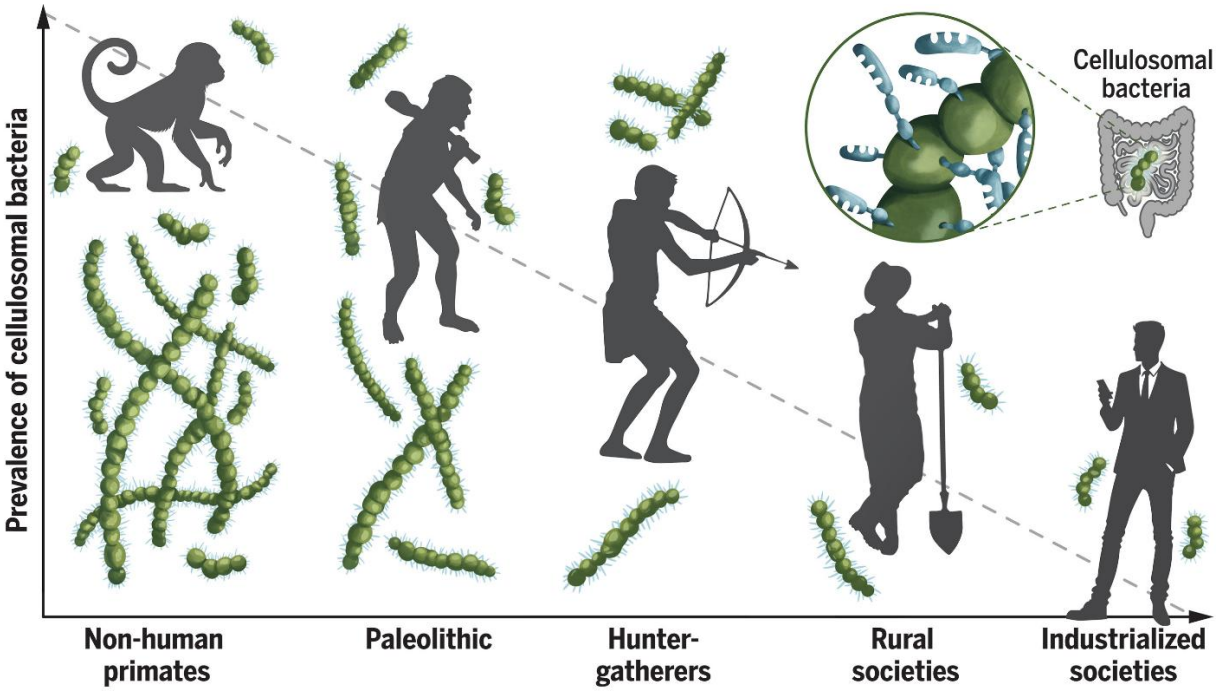


Figure 6-4: Cellulose degrading capacity of hominid gut microbiota over evolutionary time. The prevalence of human gut cellulolytic *Ruminococcus* spp. is relatively high in non-human primates and palaeolithic ancient human populations, as well as in modern day hunter-gatherers and rural societies. Prevalence is lower in urbanised human societies in industrialised countries (Moraïs *et al.*, 2024).

Although the ability to utilise cellulose may not seem important for individuals on a relatively low fibre western diet in terms of nutrient supply, loss of these species may reflect a loss of diversity within the HGM. Furthermore, an increase in proportion of the population following vegan or vegetarian diets, combined with improved understanding of the importance of dietary fibre intake for health and wellbeing, is likely leading to an increase in fibre, and hence cellulose intake. Consequently, the importance of microbiome mediated cellulose degradation may be on the rise, highlighting the need for improved understanding on the mechanisms of cellulose breakdown within the human gut. At present, we do not fully understand how important

these cellulose-utilising species are for maintenance of health, however if they are being lost it is critical to understand this, and potentially to implement studies to enhance their prevalence and abundance within the HGM of western populations.

To date, we understand that *R. champanellensis* can degrade crystalline cellulose via the production of cell bound and free cellulosome systems, and that these systems may be extremely intricate, containing multiple scaffoldin proteins bringing together up to 25 different dockerin-containing GH enzymes via specific cohesin-dockerin interactions. However, we do not currently understand how these cellulosome systems are assembled *in vivo,* whether specific cellulosome architectures correspond to different substrates, or which cellulosomal enzymes are expressed on cellulose compared to intact plant cell wall material.

In this study the aim was to investigate the mechanisms of dietary fibre by *R. champanellensis* in more detail. The data show that, as well as pure cellulose, *R. champanellensis* cinan grow on insoluble corn cell wall derived material and expanded our understanding of the activities of some of the xylanases encoded by this keystone species. In addition, mRNA for transcriptomics was prepped from both cellulose and corn grown cultures, but unfortunately not enough was obtained for RNAseq.

6.2 Results

6.2.1 Ability of *Ruminococcus champanellensis* to utilise a range of plant-cell wall derived materials to support growth

In order to improve understanding of the glycan-degradation capabilities of this unique, cellulosome-producing Bacillota from the human gut microbiota, *R. champanellensis* was screened for growth on a range of hemicelluloses and cellulose, as well as a whole plant cell wall material derived from corn, referred to as corn digest. *R. champanellensis* was grown in YCFA media supplemented with 5% rumen fluid (YCFA-RF), and 5% final concentration of carbohydrate source. Growth on YCFA-RF supplemented with a range of soluble polysaccharides was assessed via measurement of optical density at 600nm in a microplate spectrophotometer (**Figure 6-5**). This demonstrated that although *R. champanellensis* reached exponential phase of growth very rapidly when supplemented with glucose, the lag phase was much longer for all other carbohydrates, with the maximum optical density considerably lower (**Figure 6-5**). As discussed in the introduction to this thesis, *R. champanellensis* encodes 57 GH modules, belonging to 27 families, including 4 xylanases and 2 arabinofuranosidases (Ben David *et al.*, 2015), suggesting that this bacterium should be able to degrade xylans and utilise their components to support growth. However, with the exception of birchwood glucuronoxylan (BWX), growth on xylan-supplemented media was poor, indicating that xylans are not a high-priority substrate for *R. champanellensis.* Furthermore, previous studies have demonstrated that *R. champanellensis* is capable of insoluble cellulose degradation. Here, we demonstrate that this bacterium is incapable of utilising two soluble cellulose-derivatives, hydroxy-ethyl-cellulose (HEC) and carboxymethyl cellulose (CMC) (**Figure 6-5**), perhaps unsurprisingly as these are modified forms of cellulose and so are unlikely to be a

target for cellulosome systems. Of note, error bars for these growths are large, indicating large variability in growth kinetics between replicates, indicating that data presented here may not truly represent the polysaccharide utilisation capabilities of *R. champanellensis,* and other factors may be influencing cell viability.

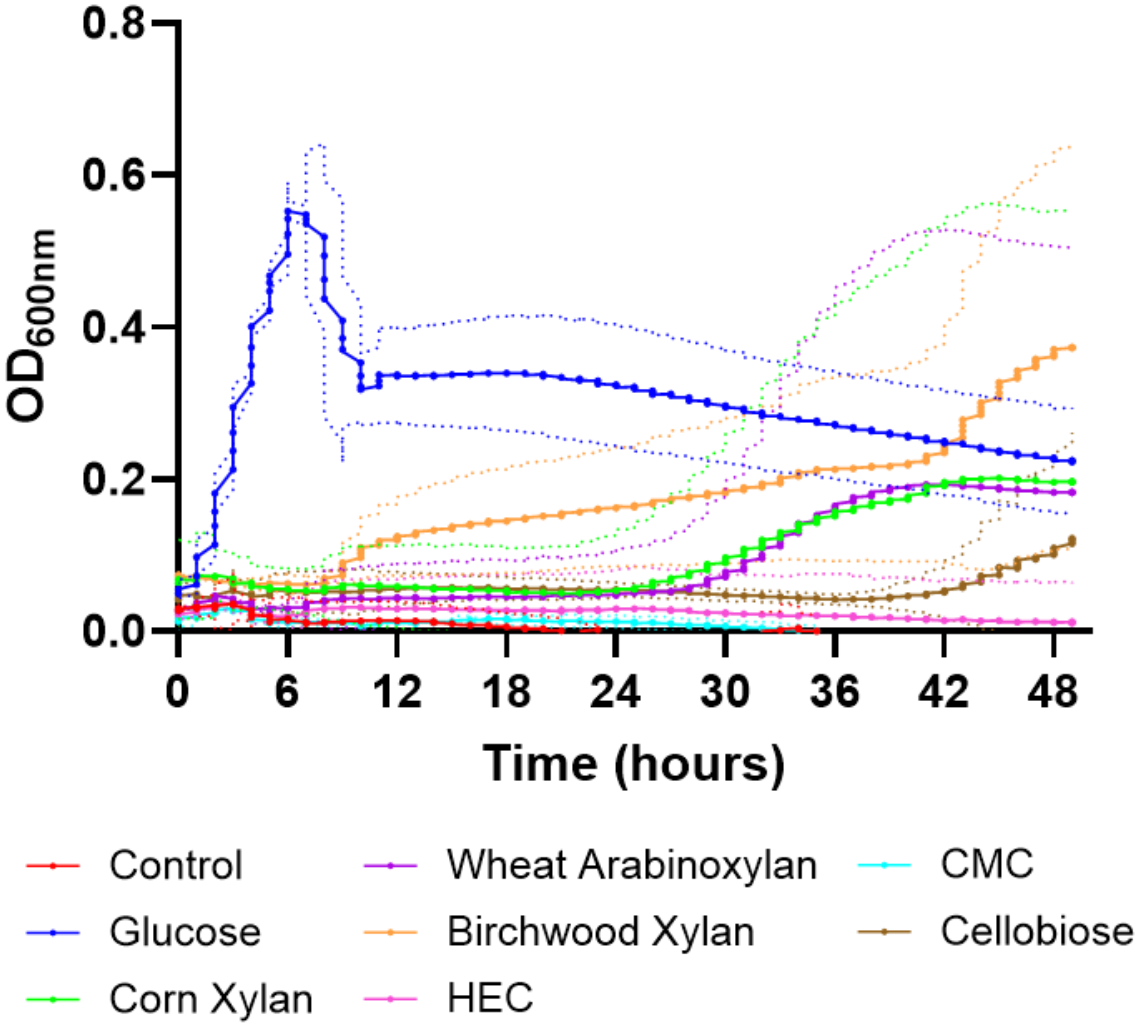


Figure 6-5: Growth of *R. champanellensis* on YCFA media supplemented with rumen fluid and various polysaccharide substrates. *R. champanellensis* cultures were run in triplicate and optical density at 600 nm was recorded every 10 minutes.

Bacillota cellulosomal systems are known to be more efficient at the degradation of insoluble or crystalline glycans, compared to other systems such as those encoded by Bacteroidota PULs. Therefore, we examined the growth kinetics of *R.*

champanellensis on YCFA-RF supplemented with 1% Avicel (insoluble pure crystalline cellulose) (**Figure 6-6**). In agreement with previous studies (Chassard *et al.*, 2012), we show that *R. champanellensis* is capable of utilising Avicel, as shown by an increase in CFU per µL of culture media (**Figure 6-6**) over that seen on YCFA-RF media alone. This is unsurprising given the multitude of GH5 and GH9 putative cellulases encoded within its genome. *R. champanellensis* reached a peak CFU count quickly when YCFA-RF was supplemented with glucose, and also reached a high CFU count on cellobiose. This correlates with the presence of a putative GH48 cellobiohydrolase encoded within the *R. champanellensis* genome.

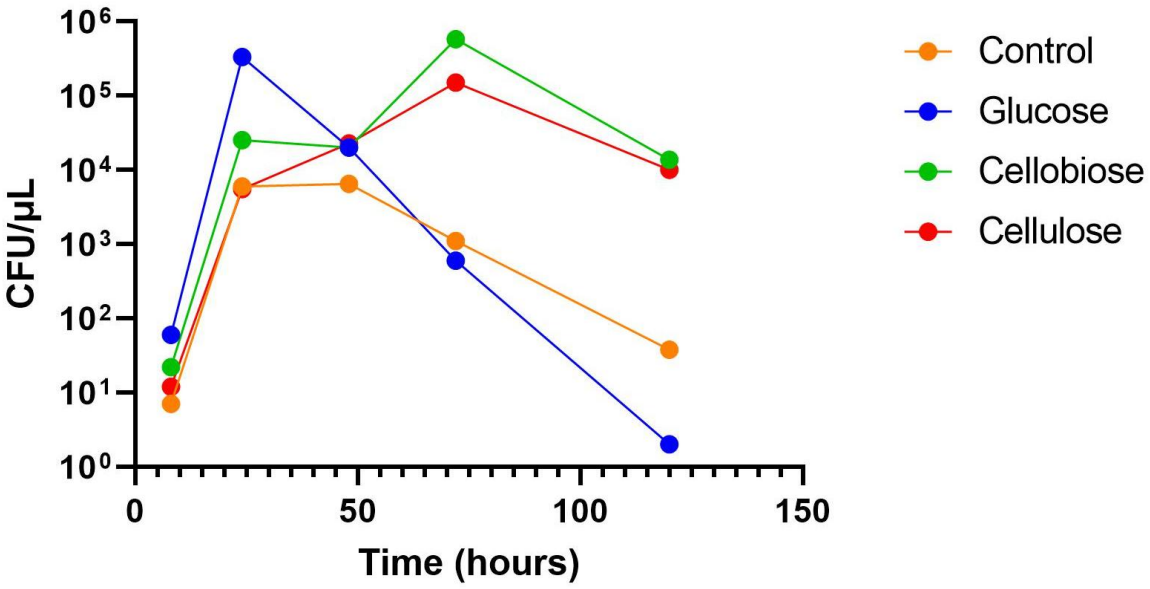


Figure 6-6: Growth of *R. champanellensis* on YCFA media supplemented with rumen fluid and various glucose-containing carbohydrates. *R. champanellensis* growth was quantified by harvesting of cells, performing a dilution series and plating on Brain Heart Infusion agar plates. Carbohydrate sources were supplemented at a final concentration of 10 mg mL^{-1}.

In the intestinal environment, members of the human gut microbiota are not exposed to isolated carbohydrates, but rather complex mixtures of polysaccharides. In order to simulate the "natural" substrate that *R. champanellensis* may target *in vivo,* we

assessed growth on corn digest; a corn bran derived substrate, treated with pepsin and pancreatin, containing amylases, peptidases and lipases, to simulate degradation of starches, lipids, proteins and peptides which occurs in the human saliva, stomach and small intestine, prior to food reaching the colonic microbiota. Here we show that *R. champanellensis* can utilise this complex plant fibre, referred to as corn digest, for growth (**Figure 6-7**), demonstrated by increase in CFU count over time. Despite this, we do not yet know which components of this complex corn digest mixture are degraded preferentially. Of note, *R. champanellensis* appears to reach stationary phase of growth more rapidly when grown with corn digest, compared to Avicel, likely due to the presence of some more easily degradable polysaccharides within the corn digest. This stationary phase may be due to depletion of other nutrients within the media, rather than availability of usable polysaccharides. Likewise, YCFA-RF corn digest seems to support a greater number of viable cells at maximum CFU count when supplemented with corn digest over Avicel (**Figures 6-6 + 6-7**).

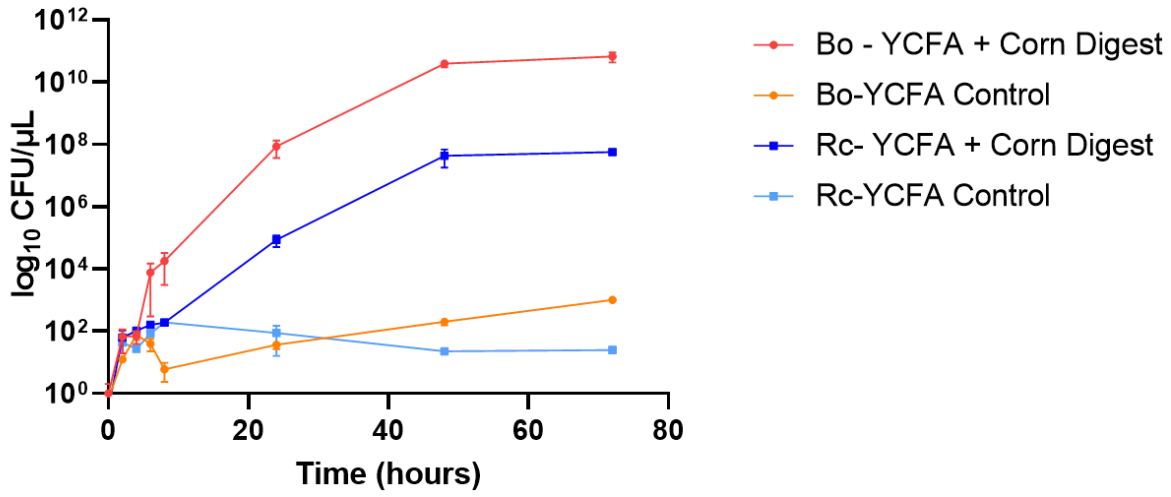


Figure 6-7: Growth of *R. champanellensis* (Rc) and *B. ovatus* (Bo) on YCFA media supplemented with rumen fluid and a corn cell wall derived substrate. Bacterial growth was quantified by harvesting of cells, performing a dilution series and plating on Brain Heart Infusion agar plates. Carbohydrate sources were supplemented at a final concentration of 10 mg mL^{-1}. The corn cell wall derived substrate, referred to as Corn Digest, was produced via pre-digestion of corn bran with pepsin and pancreatin to simulate the removal of starches, lipids and proteins that would occur in the stomach and small intestine.

6.2.2 Investigating the accessibility of corn digest components to other members of the human gut microbiota

In order to improve understanding of the accessibility of hemicelluloses within this material to degradation by members of the HGM, we also assessed growth of *B. ovatus* on corn digest supplemented YCFA-RF. This was of interest as it is well established that *B. ovatus* cannot degrade cellulose, and as such is likely to only use components of corn digest which are not cellulose-associated, especially hemicelluloses. Data show that *B. ovatus* was capable of degrading components of this complex substrate, reaching a maximum CFU count in the region of 10^{10} CFU mL^{-1}. However, as with *R. champanellensis,* we do not yet understand which components of corn digest are being degraded.

As it is well established that Bacteroidota PULs are upregulated in response to presence of particular polysaccharides, we utilised this concept to hypothesise which components are targeted by *B. ovatus* (Martens *et al.*, 2011). After 24 hours of growth on corn digest, or glucose, *B. ovatus* cells were harvested, washed in PBS, and incubated at 37 ºC with various polysaccharides for 16 hours. Supernatants were also saved and assayed against hemicelluloses to test for extracellular enzyme secretion. Assays were subsequently analysed by TLC (**Figure 6-8**). Where a hemicellulose substrate was degraded by corn digest grown *B. ovatus* cells, but not glucose grown cells, we predict that *B. ovatus* accessed that hemicellulose as a component of corn digest and hence upregulated the relevant PUL-encoded degradation apparatus (**Figure 6-8**). These whole cell assays demonstrated degradation of pectic galactan, galactomannan, linear β1-4 mannan, and β-glucan, but not xylans, or xyloglucan. Supernatant degradation profiles were similar, except that no activity was seen against β-glucan or galactomannan. This provides partial information on the composition of

corn digest, as it shows that *B. ovatus* can likely target a range of hemicelluloses without cellulose molecules preventing availability. However, these results are surprising given that cereal grains are thought to have relatively low levels of mannan, and higher abundance of other hemicelluloses such as xylans. Interestingly, whole cell assays do not suggest upregulation of *B. ovatus* xylan -targeting PULs. This is somewhat unexpected, as it is known that xylans are an abundant component of the cell wall of corn grain. Furthermore, *B. ovatus* xylan and xyloglucan capacities have been well established (Larsbrink *et al.*, 2014; Rogowski *et al.*, 2015)

Although *R. champanellensis* can also degrade cellulose and consequently may have access to a myriad of other glycans within corn digest, which are inaccessible to *B. ovatus,* this provides some preliminary data on the types of glycans which may be degraded by members of the human gut microbiota within corn digest.

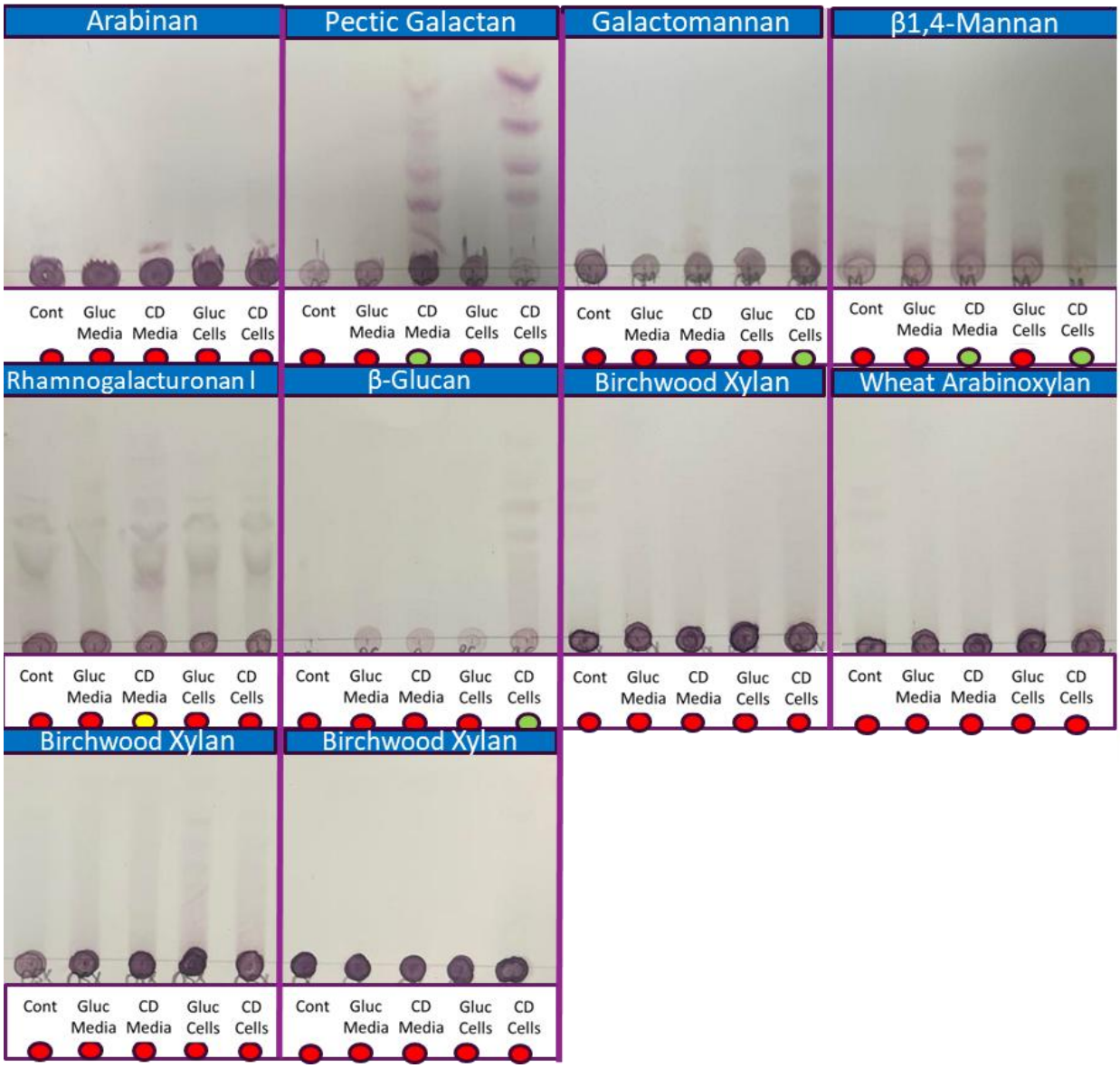


Figure 6-8: Activity of *Bacteroides ovatus* cells and spent media on a range of plant polysaccharides following 24 hours of growth on glucose or corn digest. Cells or media, following 24 hours growth on glucose or corn digest, were harvested and incubated with 5 mg mL^{-1} final concentration of polysaccharide for 16 hours at 37 ^{o}C. Red circles denote no activity, yellow circles slight activity, and green shows significant activity against the polysaccharide tested. CD – corn digest derived from whole corn bran, Gluc – glucose.

6.2.3 Investigating the activity of dockerin-containing xylanase enzymes from *Ruminococcus champanellensis*

In order to improve understanding of the polysaccharide-degrading capabilities of *R. champanellensis* cellulosomes, we aimed to recombinantly express and characterise

the activity of dockerin-containing GH enzymes encoded within the *R. champanellensis* genome. Due to the focus of this project on xylans, and their abundance within the cell wall of cereal grains, the focus was on the activity of putative xylan-active enzymes. This work was performed by undergraduate student Natalia Łoś, working within the research group. As shown in chapter 4, Rc98 is a functional GH98 endoxylanase, with specificity for complex glucuronoarabinoxylans, such as those from corn xylan. Furthermore, *R. champanellensis* encodes enzymes from GH families 10 and 11, which are predicted to possess endoxylanase activity. These enzymes were recombinantly expressed with a C-terminal His-tag (**Figure 6-9 + Table 6-1**).

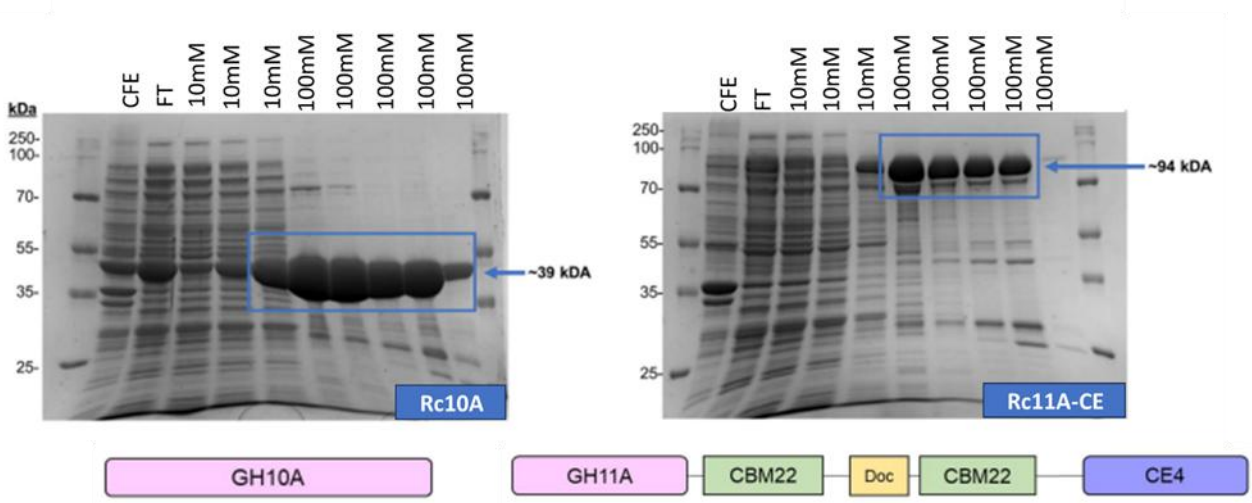


Figure 6-9: Expression and purification of recombinantly expressed putative xylanases from *R. champanellensis*. Enzymes were cloned and recombinantly expressed in E. coli Tuner cells His-tagged protein was subsequently purified and analysed by SDS PAGE with a 25 to 250 kDa molecular weight ladder (labelled on the left). Cell free extract (CFE), flow through (FT) and 8 imidazole-eluted fractions were collected: 3 x 10 mM, followed by 5 x 100 mM. Rc10A GH10 and Rc11A-CE GH11 had calculated molecular weights of 39 and 94 kDa respectively. Domain composition of the two enzymes tested here are shown below. Work carried out by Natalia Łoś.

Table 6-1; Cloning strategy and details of gene fragments amplified and analysed in this study.

Code name for recombinant protein	Locus Tag	Region cloned from DNA	Cloning vector	Cloning sites in vector	Theoretical molecular weight (kDa)	Extinction coefficient ($M^{-1}\,cm^{-1}$)
Rc10A	RUM_03400	575-1685	pET21a	NheI/XhoI	39	189385
Rc11A-CE	RUM_23130	93-2754	pET21a	NheI/XhoI	94	70945

Theoretical molecular weights (kDa) and extinction coefficients ($M^{-1}cm^{-1}$) were calculated using the Expasy Protparam tool at web.expasy.org/cgi-bin/protparam/protparam.

Studies showed that both Rc10A and Rc11A-CE are active xylan-degrading enzymes, with activity against GX (BWX) (**Figure 6-10A**) and AX (WAX) (**Figure 6-10B**), but were not active against CX, as expected for a complex GAX (**Figure 6-10C**).

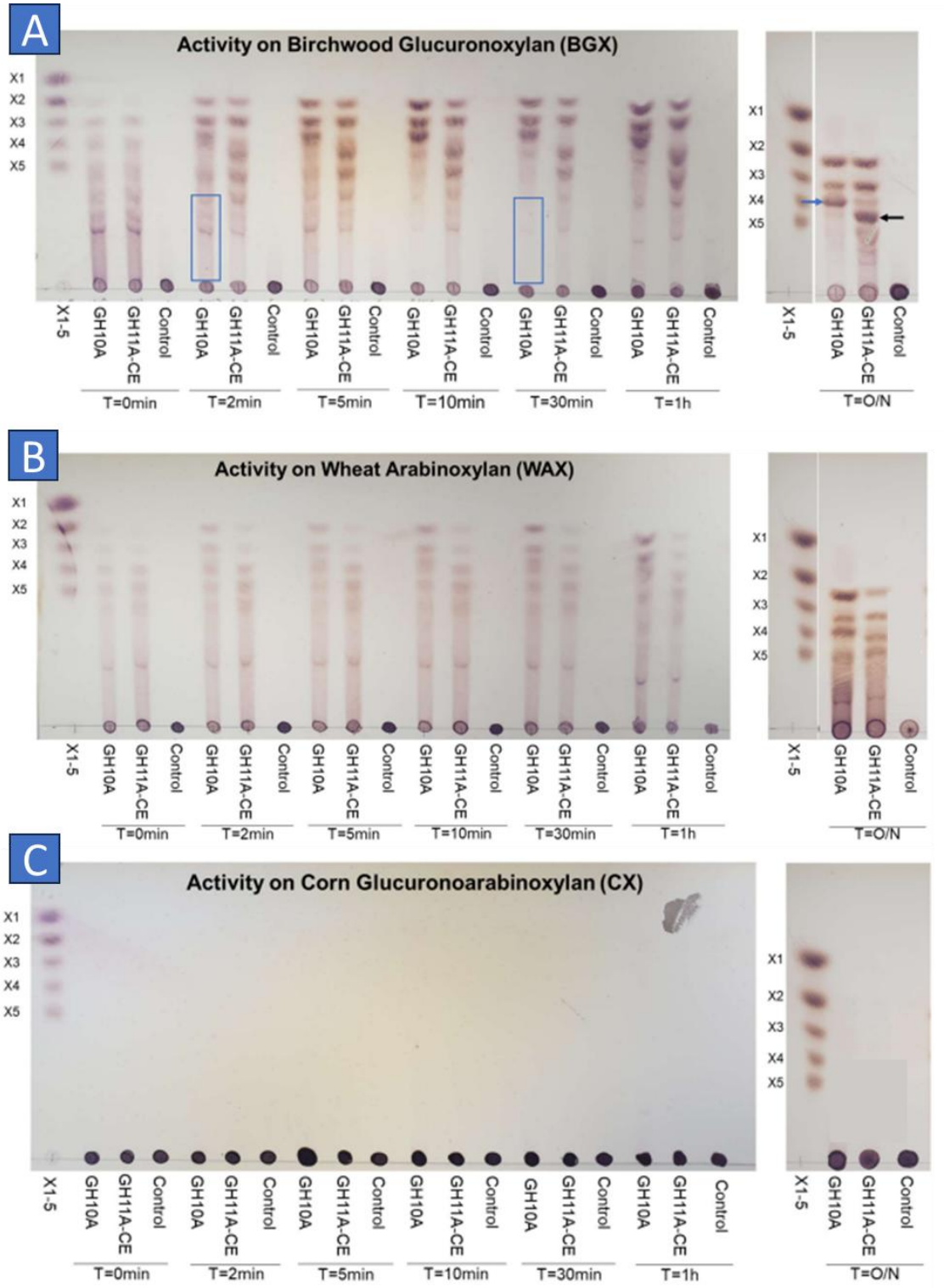


Figure 6-10: Activity of Rc10A and Rc11A-CE enzymes against the three major classes of xylans. Breakdown products of GH10 and GH11 enzymes against BWX, WAX and CX are shown via TLC. 1 µM final enzyme concentration was incubated with 5 mg mL^{-1} final concentration of xylan for 16 hours in 20 mM HEPES buffer, pH7.0. Control assays consisted of HEPES buffer with 5 mg mL^{-1} xylan. For both enzymes, activity was seen at T=0, immediately after addition of enzyme. X1-5 represents a linear β1-4 xylo-oligosaccharide standard mix from xylose to xylopentaose. Work performed by Natalia Łoś.

Both enzymes were rapidly acting against WAX and BWX (**Figure 6-10A+B**), with activity seen immediately upon addition of the enzyme to the substrate, at time T=0. Furthermore, each enzyme displayed different xylan degradation patterns. On BGW, Rc10A initially produced longer xylo-oligosaccharides (**Figure 6-10A**), as shown by the blue boxes, however by 10 minutes these had been further cleaved into shorter fragments. Rc10A was also more efficient at producing shorter xylo-oligosaccharides, migrating closer to X2 and X3 on the TLC, whereas Rc11A-CE seemed to produce longer oligosaccharides, even after 16 hours of incubation (**Figure 6-10A**). In fact, Rc11A-CE degradation products were identical at 2 minutes, compared to 16 hours, indicating that the reaction had reached completion after just 2 minutes, or that enzyme lost activity rapidly.

On WAX, for both enzymes the production of small oligosaccharide products was less for the first hour, and rather a range of larger oligosaccharides were produced, as shown by the smear on the TLC and fainter bands (**Figure 6-10B**). Again, degradation patterns between the two enzymes differed, with Rc11A-CE producing fewer very small oligosaccharides in comparison to Rc10A.

Overall, these assays, and the work on Rc98 from chapter 3, reveal that *R. champanellensis* possesses a range of xylan-active enzymes, which between them have the ability to degrade all major classes of xylans (GAX, AX and GX). This suggests that despite the poor utilisation of xylans as a sole substrate to support growth, this bacterium does possess to the enzymatic apparatus to break them down.

In this chapter we aim to understand more about the cellulolytic capabilities of *R. champanellensis.* To build on work performed in previous studies (Moraïs *et al.*, 2016, Ben David *et al.*, 2015), we aimed to clone putative GH5, GH8, GH9 and GH48

cellulases, shown in **Figure 6-1,** and examine their activity against a range of soluble and insoluble cellulose sources, and characterise degradation products by TLC and IC. Attempts to clone and recombinantly express putative cellulases from *R. champanellensis* were performed by undergraduate students Alexandra Musk and James Allen, however technical difficulties with both amplification and expression meant that none of the putative cellulases were characterised.

6.2.4 Transcriptomic studies of *Ruminococcus champanellensis* gene upregulation during growth on insoluble cellulose and corn cell-wall derived material

In order to elucidate the mechanisms of cellulosome assembly by *R. champanellensis,* as well as to understand more about the expression of cellulosomal components in response to different substrates, we aimed to perform RNAseq transcriptomics to examine the upregulation of cellulosome associated genes, as well as transporters, regulators, carbohydrate binding proteins, and non-cellulosomal CAZymes, during *R. champanellensis* growth on cellulose and corn digest, in comparison to growth on glucose. The insoluble growth substrate, as well as relatively low cell yield from growth of *R. champanellensis,* posed great challenges in the extraction of RNA. With the exception of glucose-grown cells, insufficient RNA could be extracted for RNAseq, and this was of inadequate quality (**Figure 6-11**) to perform sequencing. Even if sufficient RNA could be extracted from *R. champanellensis* to gain sequencing data, samples from glucose and cellulose/ corn digest grown cells would need to have similar RNA integrity scores to allow comparison. Unfortunately, this meant that no transcriptomic datasets have yet been gained for presentation in this thesis. As can be seen in **Figure 1-11,** RNA from glucose grown cells was of high quality, as indicated by the high RNA Integrity Number (RIN^e), where 10 represents the highest quality RNA sample. RNA

from Avicel and Corn Digest grown cells was of poor integrity, as demonstrated by the low RIN^{e} number. Future studies will focus on the optimisation of RNA extraction from this system to enable transcriptomic sequencing.

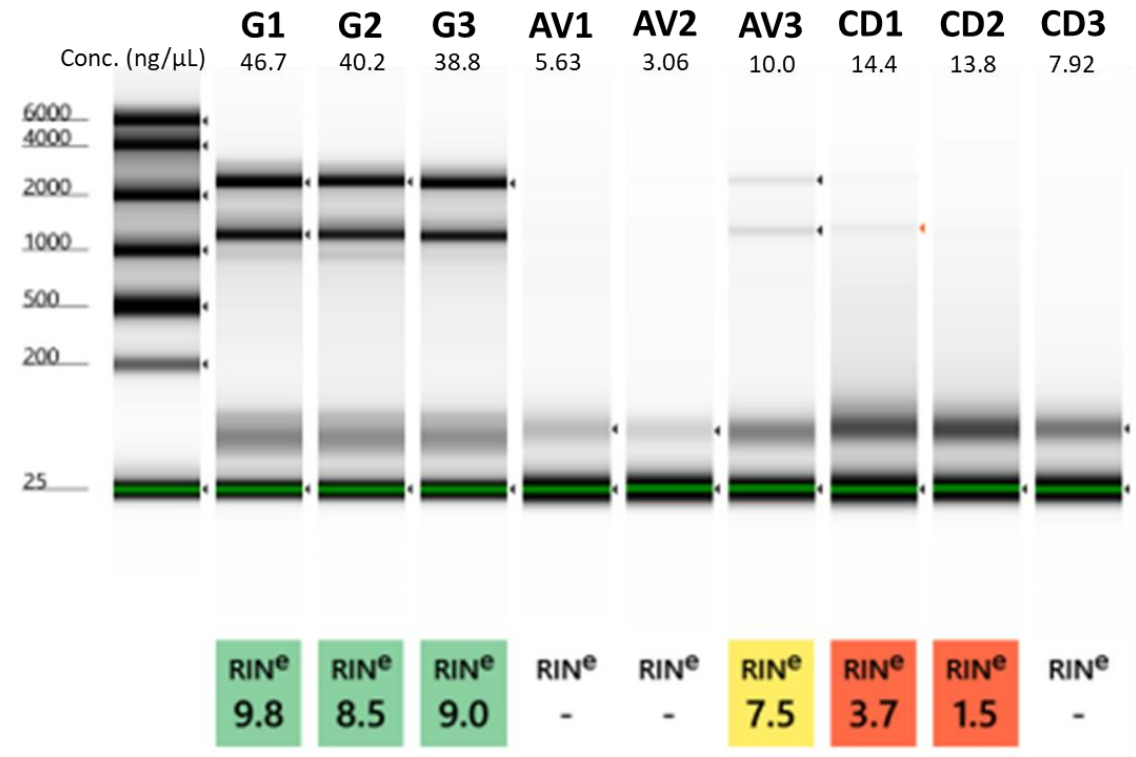


Figure 6-11: Quality control of RNA extracted from *R. champanellensis* for RNAseq transcriptomic analysis. *R. champanellensis* was grown on 3 substrates, glucose (G), avicel (AV) and corn digest (CD) in triplicate. RNA Integrity Number (RINe) demonstrated high RNA quality from glucose-grown cells, but insufficient integrity from AV or CD grown cells for RNAseq transcriptomics.

6.3 Discussion and Future Work

6.3.1 Growth of *Ruminococcus champanellensis* using xylans as a carbon source

As demonstrated here (**Figures 6-1 + 6-10**) and in previous studies (Moraïs *et al.*, 2016), *R. champanellensis* possesses multiple xylan-active enzymes, capable of degrading complex GAX (Rc98) as well as simpler AX and GX (Rc10A and Rc11A-CE) into smaller oligosaccharide products. It is interesting that *R. champanellensis* encodes both GH10 and GH11 enzymes, as gut *Bacteroides* spp. seem to only possess GH10. Although it is unclear from the data presented here, this may allow degradation of a wider range of xylans. Based on this capability, we hypothesised that *R. champanellensis* would be capable of import of these xylo-oligosaccharide products, and further metabolism supporting cell growth and proliferation. Despite this, growth on xylans as a sole carbon source was poor, with a long lag phase, and relatively low increase in optical density in comparison to growth on glucose, for all xylans except BWX. One reason for this may be that the three xylanases examined in this study all contain dockerin modules, and hence likely form part of cellulosome systems when natively expressed by *R. champanellensis.* As discussed, cellulosomal systems are adapted for the degradation of cellulose and/or insoluble or complex plant cell wall material (Bayer *et al.*, 1998), and as such incorporation of xylanases into such a complex enzyme system may reduce accessibility, and hence activity against pure, soluble xylans. In fact, catalytic activity of these enzymes may be improved against xylans within a complex matrix of cellulose and hemicelluloses due to the binding of the many CBM modules within the cellulosome to different components of plant cell wall material. Further studies into the activity of cellulosome systems against complex

plant cell wall material, such as the transcriptomic studies that we aimed to perform here, would provide a wealth of information on the xylanolytic activity of this bacterium against a substrate that it is likely to encounter within the intestinal environment. Examining the binding activity of CBMs within the cellulosome system via ITC would also be of great interest as this will help improve understanding of how a cellulosome system is targeted towards its substrate.

The poor growth of *R. champanellensis* on pure xylan may also be due to lack of ability to import xylo-oligosaccharide degradation products into the cell cytoplasm. *Ruminococcus* species are known to import glycan degradation products across the cytoplasmic membrane, largely via ABC transporters, as demonstrated by the cellulosome mediated degradation of xyloglucan by *Ruminoclostridium cellulolyticum* (Ravachol *et al.*, 2016), where degradation products are imported by highly-specific ABC transporters. Studies on the import systems of *R. champanellensis* have not yet been performed, however it is possible that this species does not possess xylo-oligosaccharide specific transporters and instead releases these degradation products for “sharing” with other members of the HGM. Removal of xylans from plant cell wall material may allow access to other components of the cell wall matrix. Further studies involving characterisation of *R. champanellensis* import systems are of great interest if we wish to fully elucidate the role of this bacterium as part of the HGM.

6.3.2 Growth of *Ruminococcus champanellensis* on cellulose and corn digest

Upon the initial isolation of *R. champanellensis,* Chassard *et al.* (2012) demonstrated its ability to utilise microcrystalline cellulose via growth in cellulose-supplemented media. Here, we present further evidence of the ability of *R. champanellensis* to utilise cellulose as a carbon source to support growth, a rare activity within the HGM.

However, growth on cellulose was slow, with a maximum CFU/ mL only reached after approximately 80 hours of growth (**Figure 6-6**). Given that the human digestive transit time averages 24 to 48 hours, albeit with much variation, this timeframe for cellulose degradation in the human gut may not be physiologically relevant, indicating that cellulose degradation in the human gut may be minimal, even where cellulose utilising strains, such as *R. champanellensis,* are present. This slow cellulose utilisation may also indicate why prevalence and abundance of cellulose utilising strains is low in many human populations (Moraïs *et al.*, 2024), perhaps with individuals or diets favouring a longer transit time enhancing survival of these bacteria within the intestinal tract. Furthermore, low dietary fibre, and hence cellulose, content in the western diet is known to be low, leading to low selection pressure for survival of these cellulose-targeting strains. These western diets are also associated with dysbiosis of the HGM, and a myriad of health conditions including neurological, cardiovascular, respiratory and metabolic disease (Illiano *et al.*, 2020). This indicates a potential role for these cellulose-utilising strains in prevention of dysbiosis and maintaining diversity within the HGM. This shows that future studies of cellulosome-producing species from the human gut are important to improve understanding of the link between the microbiome and human health. Indeed, multiple studies have shown beneficial health effects of supplementation of the diet with cellulose, particularly microbiome-mediated improvements in immune homeostasis (Fischer *et al.*, 2020; Xie *et al.*, 2022; Wen *et al.*, 2022) and further metagenomic studies into the impact on prevalence and abundance of particular taxa would be of great interest.

Here, we also demonstrate the capability of *R. champanellensis* to grow on corn digest supplemented media (**Figure 6-7**), reaching a maximum CFU count after around 45 hours of growth. This is unsurprising given that many components within the corn

digest are likely hemicelluloses and may be subject to more rapid enzymatic degradation in comparison to crystalline cellulose, in which cellulose microfibrils are tightly packed (Rongpipi *et al.*, 2019), potentially placing spatial restraints on cellulolytic enzymes. This demonstrates that *R. champanellensis* is capable of utilisation of whole plant cell wall material, similar to that which it may access within the intestinal niche, with a timeframe relevant to the human digestive transit time. Despite this, it is not clear from the limited results presented here which components of corn digest are being utilised by *R. champanellensis.* As discussed previously, it may be that this strain possesses hemicellulose-targeting GHs (**Figure 6-1**) simply to remove these from a preferred substrate, potentially cellulose, within the complex material, despite poor ability to import and utilise their degradation products, as seen in the *Ruminiclostridium cellulolyticum* xylan-degradation model (Ravachol *et al.*, 2016). Further studies are necessary to elucidate glycan-utilisation preferences of *R. champanellensis,* and this information could be critical in understanding dietary strategies for enhancing the prevalence and abundance of cellulosome-producing strains within the HGM.

6.3.3 Extraction of RNA from *Ruminococcus champanellensis* cells grown on an insoluble growth substrate

Here we discuss the wealth of information that could be gained from transcriptomic studies of *R. champanellensis* glycan utilisation when grown on cellulose, or on a complex cell wall material. Despite this, RNA extracted in this study was insufficient for sequencing and no transcriptomic data could be gained. There are several potential reasons for this, which can be taken as learning for future studies aiming to perform transcriptomic experiments on complex, insoluble growth materials. Firstly, the cell count for *R. champanellensis* is relatively low, as indicated by 1000 times lower CFU

count at stationary phase in comparison to *B. ovatus.* There is potential to overcome this by growing *R. champanellensis* in larger batches to increase cell count. Transcriptomic studies of Avicel utilisation by *Clostridium thermocellum* involved growth of the bacteria in 3 L fermentation tanks with 5g/L cellulose (Raman *et al.,* 2011). Another potential issue was the huge amount of insoluble growth substrate required for sufficient increases in *R. champanellensis* CFUs. This may have posed an issue with cell lysis, reducing the efficiency of lysozyme and lysis buffer mediated cell lysis. Future experiments could aim to separate the cells from corn digest/ cellulose substrate, potentially based on density, allowing harvesting of cells.

Proteomic studies provide a potential alternative to transcriptomics to understand the glycan-degradation strategies of *R. champanellensis.* In fact, Moraïs *et al.* (2016) performed small-scale proteomic studies, looking at the comparative proteome of *R. champanellensis* during growth on cellobiose and filter paper cellulose. The major finding was increased expression of GH9 endoglucanase and GH48 cellobiohydrolase during growth on cellulose. There is great potential to build on these proteomic experiments by examining the complete proteome of various growth fractions including the intracellular proteome, the extracellular (free) proteome, the cell-surface proteome and potentially even the substrate-bound proteome. This could provide a wealth of information on the expression of cellulosomal components, and whether these are more abundant in their cell-bound or free state. Furthermore, there are proteomic studies investigating cellulosome systems in other *Ruminococcus* spp. including *R. flavefaciens* (MaÅ¡a Vodovnik *et al.*, 2013; Rakotoarivonina *et al.*, 2009). In the present study, proteomic studies were not performed due to potential issues with tight CBM and cohesin/ dockerin mediated interactions between the insoluble substrate, cellulosome complexes, and cells leading to inability to separate fractions, and

potential problems in understanding how/where cellulosome components are functioning. However, proteomic experiments should definitely be considered when planning future studies in this area.

6.3.4 Information to be gained from "omic" studies of *Ruminococcus champanellensis* glycan utilisation

Data gained from transcriptomic or proteomic experiments could allow improved understanding of cellulosome function within the human gut, as well as the importance of cellulolytic strains in plant cell wall degradation. At present, there are many questions regarding the expression of cellulosome systems by *R. champanellensis.*

Firstly, whilst it is unlikely that cellulosome systems are constitutively expressed, due to the high energy requirements, it is not yet known whether *R. champanellensis* cellulosomes are upregulated in response to the presence of plant cell wall material. Studies have shown that cellulosomes systems from other Bacillota spp. are regulated by catabolite repression, with upregulation of many components when grown on cellulose or hemicelluloses compared to cellobiose (Artzi *et al.*, 2015; Han *et al.*, 2004). Whilst it is likely that *R. champanellensis* cellulosomes are regulated in a similar manner, the proteomic or transcriptomic studies described here would be required to elucidate this.

Furthermore, it is not yet established whether *R. champanellensis* cellulosome constituents are always the same, regardless of growth substrate, or whether cellulosome structure and CAZyme content differs according to the glycans present. Artzi *et al.* (2015) demonstrated differences in the expression level of *Clostridium clariflavum* cellulosomal proteins varied as a function of the carbon source present, however cellulosomes from microcrystalline cellulosome grown cells demonstrated

hydrolytic activity against cellulose, xylans and switchgrass, indicating a broad repertoire of CAZymes despite lack of presence of xylans in the growth substrate. Again, the future studies described here will elucidate cellulosome architecture on different growth substrates.

When *R. champanellensis* accesses plant cell wall material within the intestinal niche, the carbohydrate sources present are not constant, and rather there is likely to be fluctuation. Time course "omic" studies, taking samples at different stages of the 80 hours of growth on corn-digest, will demonstrate whether *R. champanellensis* cellulosome production is dynamic with respect to availability of different glycans within the plan cell wall structure.

6.3.5 Comparison of plant cell wall material breakdown between *Ruminococcus champanellensis* and other members of the human gut microbiota

As demonstrated in the present study, *B. ovatus,* has the ability to grow utilising corn digest, despite not encoding any cellulases, and hence is likely targeting accessible hemicelluloses within the complex plan cell wall mixture. Whole cell assays indicate that *B. ovatus* is likely utilising mannans, pectic galactans, and β-glucans within corn digest, but not xylans, despite being specialised for xylan utilisation, as described in the previous chapters of this thesis. One potential explanation for these glycan utilisation strategies is accessibility of these hemicelluloses within the corn digest material. It may be that whilst other substrates are "free" and not bound tightly to cellulose or lignins, xylans are all tightly associated with cellulose, such that the chains are inaccessible to *B. ovatus.*

Within the intestinal niche, different members of the HGM will target different glycans and may work in synergy by targeting different components of plant cell wall materials,

even sharing degradation products between phyla (Fernandez-Julia *et al.,* 2022) If a lot of xylan material within the plant cell wall is held tightly by cellulose, this presents and interesting role for cellulolytic strains within the HGM, in realising hemicelluloses, potentially xylans, but not degrading them. This would provide a nutrient source and potentially enhance survival of hemicellulose-targeting species and increase diversity within the gut, with its many associated health outcomes (Jovel *et al.,* 2018).

Transcriptomic studies, looking at gene upregulation of *B. ovatus* and *R. champanellensis* individually during growth on corn digest would provide some indication of the differences in glycan utilisation preferences of these two strains, as well as a more informative picture of which hemicelluloses found within corn digest might be accessible to *B. ovatus.* Subsequently, transcriptomic analysis of the 2 strains in co-culture would highlight potential nutrient sharing or synergistic cell-wall degradation activities between these two species. This would highlight the potential importance of cellulosome producing strains as a keystone member of the HGM.

6.3.6 Conclusions and future work

Here we demonstrate that *R. champanellensis* seems to be specialised for cellulose degradation, despite possessing enzymatic capabilities for degradation of hemicelluloses including xylans. These specialisms, combined with lack of cellulose in the western diet, may explain the low prevalence and abundance of cellulose-utilising *Ruminococci* in the human gut microbiomes of industrialised populations.

Overall, transcriptomic studies focussing on isolated *Ruminococcus* spp. are limited, likely to difficulties such as those experienced here. Proteomic studies may offer a valuable alternative through which we can study the assembly of cellulosome systems

during growth on divergent substrates, providing a wealth of information on the degradation of complex plant glycans by these unique members of the HGM.

Furthermore, comparisons of plant cell wall degradation strategies with other species, such as well-characterized hemicellulose-targeting *Bacteroides* spp. via transcriptomic or proteomic studies, could provide critical information on the interactions between species in the intestinal niche.

7. Final Discussion

Glycan breakdown by *Bacteroides* spp. is well established, including the PUL paradigm. Whilst PULDB provides a carefully curated database of predicted PULs from many Bacteroidota, here we show the ability to bioinformatically identify xylan-targeting PULs, as well as the necessity of such a PUL to degrade glycan in a varied subset of *Bacteroides* and *Prevotellaceae* species. Recently, structural studies have demonstrated that PUL-encoded proteins come together to form polymeric utilisome complexes, consisting of a SusCD complexes, which form dimers in the levan and dextran utilisomes of *B. thetataiotaomicron* as proven by cryoEM, SGBPs as well as one or more GH enzyme (White *et al.*, 2023). As shown in the present study, and many others, *Bacteroides* PULs are diverse in their composition, and it is not yet known whether the described utilisome system (White *et al.*, 2023) is conserved amongst species, amongst target substrates and even amongst PUL compositions. In fact, there could be huge divergence in the architecture of utilisomes.

To date, no structure of a *Bacteroides* xylan-targeting utilisome complex has been experimentally solved. This means that there are many unknowns as to whether PUL-encoded xylan-utilisation systems form complexes comprising SusC/D pairs, SGBPs and xylan-active enzymes. Due to the diversity in xylan-targeting PULs demonstrated in this study, there are vast possibilities for the potential structure of these complexes. For example, *B. ovatus* PULXyl-L possessed 2 SusC/D-like pairs, 4 extracellular xylan-active GHs from families 3, 30, 98 and 43, and a SGBP (Rogowski *et al.*, 2015), all of which could potentially form an incredibly complex utilisomes. Likewise, the two SusC/D pairs may each form their own utilisome, with distinct enzyme content and structure. Proximity of different enzymes, binding proteins and transport systems will

undoubtedly impact on function with respect to xylan degradation and may explain differences in xylan utilisation preferences which cannot be explained by simply assessing GH families encoded by a given species. Further structural studies are required to improve understanding of this concept. Here we show that *Bacteroides* xylan PULs possess distinct groupings of SusC/D pairs, based on their sequence similarity, and further studies into the relevance of this for utilisome formation would be fascinating and provide huge advances in our understanding of xylan degradation capabilities in the human gut.

Of particular interest in this study is the possibility that Bo98, a cell-surface endoxylanase enzyme, forms part of a utilisome. Here we hypothesise that the DUF-4988, which is unique to GH98 endoxylanases from the *Bacteroides* genus, may function in interaction with SusC/D-like complexes, enabling utilisome formation. Despite this, the potential impact of such interaction on enzyme function are unknown. It is likely that the close association with SusC/D-like pairs would enhance spatial proximity, improving efficiency of import of large xylo-oligosaccharide products of Bo98.

Whilst Bo98 may form part of a utilisome system, other Bacteroidota GH98 enzymes, such as those from *Prevotellacaea* spp., lack the DUF hypothesised to be involved in utilisome incorporation. Furthermore, many *Prevotellacaea* GH98 endoxylanases are not encoded within PULs, but rather isolated from other CAZyme encoding genes, indicating that these enzymes are unlikely to form part of a utilisome. On the other hand, Rc98 contains a dockerin module, indicating that is likely to form part of a cellulosome system. Despite differences in situation with respect to the plant cell wall, all these GH98 enzymes possess the same endoxylanase activity, indicating that the surrounding glycan-targeting apparatus does not impact on enzyme function.

Another important consideration when considering whether Bo98 forms part of a utilisome is the potential complexity of these complexes, and whether all extracellular enzymes encoded within *B. ovatus* PULXylL sit within it. This would include a GH3 β-xylosidase, GH43 arabinofuranosidase, GH30 glucuronoxylanase and GH98 endoxylanase as well as a SGBP. If all of these enzymes were to sit within the utilisome system, this would be exceptionally complex, and further studies to understand whether this occurs, and which enzymes interact directly with the SusC/D-like proteins, would be of great interest.

If, as discussed here, Bo98 sits as part of a utilisome complex, the function of Bo98-CBM35 within this system is of interest, given the potential presence of one SGBP, and potentially two different SusD-like binding proteins within the system. Whilst we show in the present study that Bo98-CBM35 binds to arabinose-decorated glycans, both Bo_03427 SusD-like proteins and Bo_03432 SGBP bind to GX and GAX, a slightly different binding specificity (Rogowski *et al.*, 2015). Without structural studies demonstrating the orientation of Bo98-CBM35 within the utilisome, it is difficult to predict the function of this in xylan degradation.

In this study we discuss two distinct types of glycan utilisation systems for targeting plant cell wall material, cellulosome systems and PUL-encoded utilisome systems. The key differences between these two systems seems to be that cellulosomes are larger complexes, consisting mostly of GH enzymes and CBMs, held together by strong cohesin-dockerin interactions, targeting a range of polysaccharides including cellulose and hemicelluloses. On the other hand, utilisomes are smaller and target one distinct glycan type, and contain SusC/D-like binding and transport systems, likely enhancing the efficiency in the import of degradation products. The differences between these two systems indicate the importance of diversity within the intestinal

niche, as it is clear that they allow different species to target different components of dietary fibre.

Whilst a lot is known about dietary glycan degradation by distinct groups of the gut microbiota, in particular the *Bacteroides*, it is important to consider that these species do not exist in isolation, but rather in an extremely complex ecosystem consisting of many, many different genera, each with different abilities and preferences for glycan degradation. Bacillota belonging to the Clostridial Clusters XIVa and IV are hugely prevalent in the HGM, and although many of them are known to possess glycan degrading capabilities, they are largely understudied, suggesting a wealth of unknown information which could have a significant impact on our understanding of the role of the gut microbiome as a whole. Of particular interest are *Ruminococcus* spp., first identified from the rumen, where their importance in glycan utilisation has been well documented.

R. champanellensis remains the only characterised member of the human gut microbiota capable of cellulose degradation, likely due to lack of reliance of humans on utilising such recalcitrant carbohydrate sources as a means to provide energy. Because of this, *R. champanellensis* should not be considered as a keystone member of the HGM, particularly in industrialised populations, despite potentially degrading hemicellulose-adhered cellulose. This cellulose material is normally cross-linked by hemicelluloses such as xylans, potentially rendering hemicelluloses inaccessible to other members of the HGM, including *Bacteroides* spp. *R. champanellensis* mediated removal of this cellulose may improve availability of hemicelluloses to degradation by other members of the HGM, drastically influencing population dynamics with respect to glycan utilisation, due to altered substrate availability.

As shown by Morais *et al.* the prevalence of cellulose utilising strains is greater in rural and hunter-gatherer populations, who eat a diet much higher in dietary fibre than individuals in western populations, increasing reliance on microbial degradation of plant cell wall material for host energy supply. In these human populations it is possible that even a small amount of cellulose degradation could lead to shift in gut microbiome composition towards carbohydrate-utilising species, and as such improve efficiency in nutrient acquisition from high-fibre foods such as whole cereal grains, fruits, vegetables and other grasses. Interestingly, trends towards vegan and vegetarianism in western populations are increasing. However, widespread loss of cellulose-utilising strains from the HGM may impact on the ability of individuals on fibre-based diets to obtain sufficient energy sources to sustain health and wellbeing. Reintroduction of these strains, naturally, or via probiotic supplementation may play an important role in this, leading to beneficial health outcomes, however studies will be necessary to elucidate this.

Human and animal supplementation with prebiotics, probiotics, postbiotics or synbiotics, with the aim of improving digestive health and function, is an ever-expanding market. In order to understand what impact biotic supplementation could have on microbiome function, we need to understand the glycan-degrading capacities of individual species, and potential impact of their upregulation. The more we understand the glycan degradation strategies employed by different members of the HGM, the more we can understand what impact their supplementation could make and start to implement targeted approaches.

These strategies could include feeding particular glycans as prebiotics, with the aim of upregulating particular species within the gut microbiota, which consequently impacts on overall microbiota composition and its ability to degrade dietary fibre, for

example in one study, seaweed polysaccharides were shown to enrich *Bacteroides* spp. within the HGM, whereas galacto-oligosaccharides enriched certain *Bifidobacteria* species (Wang *et al.*, 2020). Likewise, supplementation with certain probiotic strains can alter HGM composition, with the potential for many positive health outcomes (Latif *et al.*, 2023). The more that we understand about the glycan-degradation capacities of different species within the HGM, the better we can make targeted implementations to improve diversity and function of the HGM, particularly. As discussed here, the loss of cellulose utilising strains from the HGM of western populations shown by Moraïs *et al.* (2024), may be associated with loss of diversity and health problems. Whilst direct probiotic supplementation with *R. champanellensis* may pose practical problems, such as maintaining live bacterial cells in a format viable for supplementation, this is without doubt an interesting area for future study. Furthermore, *in vitro* or *in vivo* studies could be used to screen probiotics and prebiotics which may support the survival of cellulosome-producing strains within the HGM and study the impact that this may have on overall diversity.

One other interesting consideration is the potential for supplementation of artificial cellulosomes, such that they could initiate the degradation of cellulose within plant cell wall material in a manner similar to cell-free cellulosomes produced by *R. champanellensis* and ruminal *Ruminococcus* spp. The concept of designer cellulosomes has been applied in biomass degradation for the production of biofuels (Lamote *et al.*, 2023; Kahn *et al.*, 2019; Moraïs *et al.*, 2010), however this has not yet been applied in human or animal nutrition.

These interventions are not only important in human health but also when considering domestic species, in particular food production systems. Whilst the utilisation of antibiotics as growth promoters has long been banned in the European Union, there

is potential for targeted “biotic” supplementation to improve utilisation of glycans within feed sources and hence improve efficiency of agricultural systems. Utilisation of CAZymes to pre-treat feed materials, improving the speed and efficiency of nutrient acquisition by the host organism is also an interesting area of study.

In vitro studies into the mechanisms of dietary fibre degradation by individual members of the gut microbiota, such as those presented here, are a critical first step in understanding the role of individual strains within the gut ecosystem, as well as to make some initial hypothesis on potential nutritional strategies to improve digestive health. However, if we wish to make targeted nutritional approaches to alter the gut microbiome, and its glycan-degrading capacity as a whole, further studies can be performed using *in vitro* intestinal models, the most complex of which can take into consideration the entire microbiome, dietary components and host factors such as the immune system, peristalsis and compartmentalisation. The impact of biotic supplementation in these systems can provide a more realistic representation of any potential beneficial health outcomes prior to *in vivo* studies. Furthermore, shotgun metagenomic sequencing can be utilised to provide functional information on an intestinal ecosystem, including the relative abundances of CAZyme-encoding genes from which we can infer glycan degrading capacities. For example, in this study we show that GH98 enzymes encoded by plant-glycan targeting bacterial strains possess endo-xylanases activity. This information can potentially be inferred in functional annotation of metagenomic datasets to imply the ability of a system to target complex xylans.

Overall, the novel information presented in this study contributes to a rapidly expanding bank of information on glycan degradation by the gut microbiota, improving understanding of the roles of divergent members of the gut microbiota in breakdown

of dietary material. This is critical in order to employ dietary implementations to improve digestive health of humans and animals alike.

8. Bibliography

Aachary, A.A. & Prapulla, S.G. (2011) 'Xylooligosaccharides (XOS) as an Emerging Prebiotic: Microbial Synthesis, Utilization, Structural Characterization, Bioactive Properties, and Applications', *Comprehensive Reviews in Food Science and Food Safety*, 10(1), pp. 2–16.

Aakko, J., Pietilä, S., Toivonen, R., Rokka, A., Mokkala, K., Laitinen, K., Elo, L. & Hänninen, A. (2020) 'A carbohydrate-active enzyme (CAZy) profile links successful metabolic specialization of Prevotella to its abundance in gut microbiota', *Scientific Reports*, 10(1), p. 12411.

Abbott, D.W. & van Bueren, A.L. (2014) 'Using structure to inform carbohydrate binding module function', *Current Opinion in Structural Biology*, 28pp. 32–40.

Accetto, T. & Avguštin, G. (2021) 'Non-oral Prevotella stepping into the spotlight', *Anaerobe*, 68p. 102321.

Accetto, T. & Avguštin, G. (2015) 'Polysaccharide utilization locus and CAZYme genome repertoires reveal diverse ecological adaptation of Prevotella species', *Systematic and Applied Microbiology*, 38(7), pp. 453–461.

Accetto, T. & Avguštin, G. (2019) 'The diverse and extensive plant polysaccharide degradative apparatuses of the rumen and hindgut Prevotella species: A factor in their ubiquity?', *Systematic and Applied Microbiology*, 42(2), pp. 107–116.

Agger, J., Viksø-Nielsen, A. & Meyer, A.S. (2010) 'Enzymatic Xylose Release from Pretreated Corn Bran Arabinoxylan: Differential Effects of Deacetylation and Deferuloylation on Insoluble and Soluble Substrate Fractions', *Journal of Agricultural and Food Chemistry*, 58(10), pp. 6141–6148.

Agirre, J., Atanasova, M., Bagdonas, H., Ballard, C.B., Baslé, A., Beilsten-Edmands, J., Borges, R.J., Brown, D.G., Burgos-Mármol, J.J., Berrisford, J.M., Bond, P.S., Caballero, I., Catapano, L., Chojnowski, G., Cook, A.G., Cowtan, K.D., Croll, T.I., Debreczeni, J.É., Devenish, N.E., *et al.* (2023) 'The *CCP* 4 suite: integrative software for macromolecular crystallography', *Acta Crystallographica Section D Structural Biology*, 79(6), pp. 449–461.

Anderson, K.L. & Salyers, A.A. (1989) 'Biochemical evidence that starch breakdown by Bacteroides thetaiotaomicron involves outer membrane starch-binding sites and periplasmic starch-degrading enzymes', *Journal of Bacteriology*, 171(6), pp. 3192–3198.

Anderson, K.M., Ashida, H., Maskos, K., Dell, A., Li, S.-C. & Li, Y.-T. (2005) 'A Clostridial Endo-β-galactosidase That Cleaves Both Blood Group A and B Glycotopes', *Journal of Biological Chemistry*, 280(9), pp. 7720–7728.

Ardèvol, A. & Rovira, C. (2015) 'Reaction Mechanisms in Carbohydrate-Active Enzymes: Glycoside Hydrolases and Glycosyltransferases. Insights from *ab Initio* Quantum Mechanics/Molecular Mechanics Dynamic Simulations', *Journal of the American Chemical Society*, 137(24), pp. 7528–7547.

Armenta, S., Moreno-Mendieta, S., Sánchez-Cuapio, Z., Sánchez, S. & Rodríguez-Sanoja, R. (2017) 'Advances in molecular engineering of carbohydrate-binding modules', *Proteins: Structure, Function, and Bioinformatics*, 85(9), pp. 1602–1617.

Artzi, L., Bayer, E.A. & Moraïs, S. (2017a) 'Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides', *Nature Reviews Microbiology*, 15(2), pp. 83–95.

Artzi, L., Bayer, E.A. & Moraïs, S. (2017b) 'Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides', *Nature Reviews Microbiology*, 15(2), pp. 83–95.

Artzi, L., Morag, E., Barak, Y., Lamed, R. & Bayer, E.A. (2015) 'Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis', *mBio*, 6(3),

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., *et al.* (2011) 'Enterotypes of the human gut microbiome', *Nature*, 473(7346), pp. 174–180.

Aspeborg, H., Coutinho, P.M., Wang, Y., Brumer, H. & Henrissat, B. (2012) 'Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5)', *BMC Evolutionary Biology*, 12(1), p. 186.

Bansal, K., Midha, S., Kumar, S. & Patil, P.B. (2017) 'Ecological and Evolutionary Insights into Xanthomonas citri Pathovar Diversity', *Applied and Environmental Microbiology*, 83(9).

Bayer, E.A., Kenig, R. & Lamed, R. (1983) 'Adherence of Clostridium thermocellum to cellulose', *Journal of Bacteriology*, 156(2), pp. 818–827.

Bayer, E.A., Lamed, R., White, B.A. & Flint, H.J. (2008) 'From cellulosomes to cellulosomics', *The Chemical Record*, 8(6), pp. 364–377.

Bayer, E.A., Shimon, L.J.W., Shoham, Y. & Lamed, R. (1998) 'Cellulosomes—Structure and Ultrastructure', *Journal of Structural Biology*, 124(2–3), pp. 221–234.

Berg Miller, M.E., Antonopoulos, D.A., Rincon, M.T., Band, M., Bari, A., Akraiko, T., Hernandez, A., Thimmapuram, J., Henrissat, B., Coutinho, P.M., Borovok, I., Jindou, S., Lamed, R., Flint, H.J., Bayer, E.A. & White, B.A. (2009) 'Diversity and Strain Specificity of Plant Cell Wall Degrading Enzymes Revealed by the Draft Genome of Ruminococcus flavefaciens FD-1', *PLoS ONE*, 4(8), p. e6650.

Biely, P. (2012) 'Microbial carbohydrate esterases deacetylating plant polysaccharides', *Biotechnology Advances*, 30(6), pp. 1575–1588.

Biely, P., Singh, S. & Puchart, V. (2016) 'Towards enzymatic breakdown of complex plant xylan structures: State of the art', *Biotechnology Advances*, 34(7), pp. 1260–1274.

Biely, P., Šuchová, K. & Puchart, V. (2023) 'Diversity of microbial endo-β-1,4-xylanases', in *Glycoside Hydrolases*. [Online]. Elsevier. pp. 135–163.

Bindelle, J., Buldgen, A., Boudry, C. & Leterme, P. (2007) 'Effect of inoculum and pepsin–pancreatin hydrolysis on fibre fermentation measured by the gas production technique in pigs', *Animal Feed Science and Technology*, 132(1–2), pp. 111–122.

Boisen, S. & Fernández, J.A. (1997) 'Prediction of the total tract digestibility of energy in feedstuffs and pig diets by in vitro analyses', *Animal Feed Science and Technology*, 68(3–4), pp. 277–286.

Bolam, D.N. & van den Berg, B. (2018) ‘TonB-dependent transport by the gut microbiota: novel aspects of an old problem’, *Current Opinion in Structural Biology*, 51pp. 35–43.

Bolam, D.N. & Koropatkin, N.M. (2012) ‘Glycan recognition by the Bacteroidetes Sus-like systems’, *Current Opinion in Structural Biology*, 22(5), pp. 563–569.

Bolam, D.N., Xie, H., Pell, G., Hogg, D., Galbraith, G., Henrissat, B. & Gilbert, H.J. (2004a) ‘X4 Modules Represent a New Family of Carbohydrate-binding Modules That Display Novel Properties’, *Journal of Biological Chemistry*, 279(22), pp. 22953–22963.

Bolam, D.N., Xie, H., Pell, G., Hogg, D., Galbraith, G., Henrissat, B. & Gilbert, H.J. (2004b) ‘X4 Modules Represent a New Family of Carbohydrate-binding Modules That Display Novel Properties’, *Journal of Biological Chemistry*, 279(22), pp. 22953–22963.

Boraston, A.B., Bolam, D.N., Gilbert, H.J. & Davies, G.J. (2004) ‘Carbohydrate-binding modules: fine-tuning polysaccharide recognition’, *Biochemical Journal*, 382(3), pp. 769–781.

Brody, J.R. & Kern, S.E. (2004) ‘History and principles of conductive media for standard DNA electrophoresis’, *Analytical Biochemistry*, 333(1), pp. 1–13.

Bryant, P., Pozzati, G. & Elofsson, A. (2022) ‘Improved prediction of protein-protein interactions using AlphaFold2’, *Nature Communications*, 13(1), p. 1265.

Burnet, M.C., Dohnalkova, A.C., Neumann, A.P., Lipton, M.S., Smith, R.D., Suen, G. & Callister, S.J. (2015) ‘Evaluating Models of Cellulose Degradation by Fibrobacter succinogenes S85’, *PLOS ONE*, 10(12), p. e0143809.

Cann, I., Bernardi, R.C. & Mackie, R.I. (2016) ‘Cellulose degradation in the human gut: *Ruminococcus champanellensis* expands the cellulosome paradigm’, *Environmental Microbiology*, 18(2), pp. 307–310.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009) ‘The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics’, *Nucleic Acids Research*, 37(Database), pp. D233–D238.

Carding, S., Verbeke, K., Vipond, D.T., Corfe, B.M. & Owen, L.J. (2015) ‘Dysbiosis of the gut microbiota in disease’, *Microbial Ecology in Health & Disease*, 26(0).

Carter, M.M., Olm, M.R., Merrill, B.D., Dahan, D., Tripathi, S., Spencer, S.P., Yu, F.B., Jain, S., Neff, N., Jha, A.R., Sonnenburg, E.D. & Sonnenburg, J.L. (2023) ‘Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes’, *Cell*, 186(14), pp. 3111-3124.e13.

Centanni, M., Hutchison, J.C., Carnachan, S.M., Daines, A.M., Kelly, W.J., Tannock, G.W. & Sims, I.M. (2017) ‘Differential growth of bowel commensal Bacteroides species on plant xylans of differing structural complexity’, *Carbohydrate Polymers*, 157pp. 1374–1382.

Chanliaud, E., Saulnier, L. & Thibault, J.-F. (1995) ‘Alkaline extraction and characterisation of heteroxylans from maize bran’, *Journal of Cereal Science*, 21(2), pp. 195–203.

Chassard, C., Delmas, E., Robert, C., Lawson, P.A. & Bernalier-Donadille, A. (2012) ‘Ruminococcus champanellensis sp. nov., a cellulose-degrading bacterium from human gut microbiota’, *International Journal of Systematic and Evolutionary Microbiology*, 62(1), pp. 138–143.

Chassard, C., Delmas, E., Robert, C. & Bernalier-Donadille, A. (2010) 'The cellulose-degrading microbial community of the human gut varies according to the presence or absence of methanogens', *FEMS Microbiology Ecology*, 74(1), pp. 205–213.

Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S.J., Webb, C., Wu, D., Varghese, N.J., Reddy, T.B.K., Mukherjee, S., Ovchinnikova, G., Nolan, M., Seshadri, R., Roux, S., Visel, A., Woyke, T., *et al.* (2023) 'The IMG/M data management and analysis system v.7: content updates and new features', *Nucleic Acids Research*, 51(D1), pp. D723–D732.

Chen, M., Liu, S., Imam, K.Md.S.U., Sun, L., Wang, Y., Gu, T., Wen, B. & Xin, F. (2020) 'The Effect of Xylooligosaccharide, Xylan, and Whole Wheat Bran on the Human Gut Bacteria', *Frontiers in Microbiology*, 11.

Chen, T., Long, W., Zhang, C., Liu, S., Zhao, L. & Hamaker, B.R. (2017) 'Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota', *Scientific Reports*, 7(1), p. 2594.

Chen, Z., Li, S., Fu, Y., Li, C., Chen, D. & Chen, H. (2019) 'Arabinoxylan structural characteristics, interaction with gut microbiota and potential health functions', *Journal of Functional Foods*, 54pp. 536–551.

Chiniquy, D., Sharma, V., Schultink, A., Baidoo, E.E., Rautengarten, C., Cheng, K., Carroll, A., Ulvskov, P., Harholt, J., Keasling, J.D., Pauly, M., Scheller, H. V. & Ronald, P.C. (2012) 'XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan', *Proceedings of the National Academy of Sciences*, 109(42), pp. 17117–17122.

Cho, K.H. & Salyers, A.A. (2001) 'Biochemical Analysis of Interactions between Outer Membrane Proteins That Contribute to Starch Utilization by *Bacteroides thetaiotaomicron*', *Journal of Bacteriology*, 183(24), pp. 7224–7230.

Chow, V., Shantharaj, D., Guo, Y., Nong, G., Minsavage, G. V., Jones, J.B. & Preston, J.F. (2015) 'Xylan Utilization Regulon in Xanthomonas citri pv. citri Strain 306: Gene Expression and Utilization of Oligoxylosides', *Applied and Environmental Microbiology*, 81(6), pp. 2163–2172.

Cockburn, D.W. & Koropatkin, N.M. (2016) 'Polysaccharide Degradation by the Intestinal Microbiota and Its Influence on Human Health and Disease', *Journal of Molecular Biology*, 428(16), pp. 3230–3252.

Cohen, S.N., Chang, A.C.Y. & Hsu, L. (1972) 'Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of *Escherichia coli* by R-Factor DNA', *Proceedings of the National Academy of Sciences*, 69(8), pp. 2110–2114.

Collins, T., Gerday, C. & Feller, G. (2005) 'Xylanases, xylanase families and extremophilic xylanases', *FEMS Microbiology Reviews*, 29(1), pp. 3–23.

Correia, M.A.S., Abbott, D.W., Gloster, T.M., Fernandes, V.O., Prates, J.A.M., Montanier, C., Dumon, C., Williamson, M.P., Tunnicliffe, R.B., Liu, Z., Flint, J.E., Davies, G.J., Henrissat, B., Coutinho, P.M., Fontes, C.M.G.A. & Gilbert, H.J. (2010) 'Signature Active Site Architectures Illuminate the Molecular Basis for Ligand Specificity in Family 35 Carbohydrate Binding Module', *Biochemistry*, 49(29), pp. 6193–6205.

Cosgrove, D.J. (2005) 'Growth of the plant cell wall', *Nature Reviews Molecular Cell Biology*, 6(11), pp. 850–861.

Cosgrove, D.J. & Jarvis, M.C. (2012) ‘Comparative structure and biomechanics of plant primary and secondary cell walls’, *Frontiers in Plant Science*, 3.

Couturier, M., Roussel, A., Rosengren, A., Leone, P., Stålbrand, H. & Berrin, J.-G. (2013) ‘Structural and Biochemical Analyses of Glycoside Hydrolase Families 5 and 26 β-(1,4)-Mannanases from Podospora anserina Reveal Differences upon Manno-oligosaccharide Catalysis’, *Journal of Biological Chemistry*, 288(20), pp. 14624–14635.

Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. (2004) ‘WebLogo: A Sequence Logo Generator: Figure 1’, *Genome Research*, 14(6), pp. 1188–1190.

Crost, E.H., Tailford, L.E., Le Gall, G., Fons, M., Henrissat, B. & Juge, N. (2013) ‘Utilisation of Mucin Glycans by the Human Gut Symbiont Ruminococcus gnavus Is Strain-Dependent’, *PLoS ONE*, 8(10), p. e76341.

Curry, T.M., Peña, M.J. & Urbanowicz, B.R. (2023) ‘An update on xylan structure, biosynthesis, and potential commercial applications’, *The Cell Surface*, 9p. 100101.

Cuskin, F., Lowe, E.C., Temple, M.J., Zhu, Y., Cameron, E.A., Pudlo, N.A., Porter, N.T., Urs, K., Thompson, A.J., Cartmell, A., Rogowski, A., Hamilton, B.S., Chen, R., Tolbert, T.J., Piens, K., Bracke, D., Vervecken, W., Hakki, Z., Speciale, G., *et al.* (2015) ‘Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism’, *Nature*, 517(7533), pp. 165–169.

Czjzek, M., Bolam, D.N., Mosbah, A., Allouch, J., Fontes, C.M.G.A., Ferreira, L.M.A., Bornet, O., Zamboni, V., Darbon, H., Smith, N.L., Black, G.W., Henrissat, B. & Gilbert, H.J. (2001) ‘The Location of the Ligand-binding Site of Carbohydrate-binding Modules That Have Evolved from a Common Sequence Is Not Conserved’, *Journal of Biological Chemistry*, 276(51), pp. 48580–48587.

Dąbrowska, K. & Witkiewicz, W. (2016) ‘Correlations of Host Genetics and Gut Microbiome Composition’, *Frontiers in Microbiology*, 7.

D’Argenio, V. & Salvatore, F. (2015) ‘The role of the gut microbiome in the healthy adult status’, *Clinica Chimica Acta*, 451pp. 97–102.

Ben David, Y., Dassa, B., Borovok, I., Lamed, R., Koropatkin, N.M., Martens, E.C., White, B.A., Bernalier-Donadille, A., Duncan, S.H., Flint, H.J., Bayer, E.A. & Moraïs, S. (2015a) ‘Ruminococcal cellulosome systems from rumen to human’, *Environmental Microbiology*, 17(9), pp. 3407–3426.

Ben David, Y., Dassa, B., Borovok, I., Lamed, R., Koropatkin, N.M., Martens, E.C., White, B.A., Bernalier-Donadille, A., Duncan, S.H., Flint, H.J., Bayer, E.A. & Moraïs, S. (2015b) ‘Ruminococcal cellulosome systems from rumen to human’, *Environmental Microbiology*, 17(9), pp. 3407–3426.

Davies, G. & Henrissat, B. (1995) ‘Structures and mechanisms of glycosyl hydrolases’, *Structure*, 3pp. 853–859.

Déjean, G., Tamura, K., Cabrera, A., Jain, N., Pudlo, N.A., Pereira, G., Viborg, A.H., Van Petegem, F., Martens, E.C. & Brumer, H. (2020) ‘Synergy between Cell Surface Glycosidases and Glycan-Binding Proteins Dictates the Utilization of Specific Beta(1,3)-Glucans by Human Gut *Bacteroides*’, *mBio*, 11(2), .

D’Elia, J.N. & Salyers, A.A. (1996a) ‘Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron’, *Journal of Bacteriology*, 178(24), pp. 7180–7186.

D'Elia, J.N. & Salyers, A.A. (1996b) 'Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron', *Journal of Bacteriology*, 178(24), pp. 7180–7186.

Despres, J., Forano, E., Lepercq, P., Comtet-Marre, S., Jubelin, G., Chambon, C., Yeoman, C.J., Berg Miller, M.E., Fields, C.J., Martens, E., Terrapon, N., Henrissat, B., White, B.A. & Mosoni, P. (2016) 'Xylan degradation by the human gut Bacteroides xylanisolvens XB1AT involves two distinct gene clusters that are linked at the transcriptional level', *BMC Genomics*, 17(1), p. 326.

Dodd, D., Mackie, R.I. & Cann, I.K.O. (2011) 'Xylan degradation, a metabolic property shared by rumen and human colonic Bacteroidetes', *Molecular Microbiology*, 79(2), pp. 292–304.

Dodd, D., Moon, Y.-H., Swaminathan, K., Mackie, R.I. & Cann, I.K.O. (2010a) 'Transcriptomic Analyses of Xylan Degradation by Prevotella bryantii and Insights into Energy Acquisition by Xylanolytic Bacteroidetes', *Journal of Biological Chemistry*, 285(39), pp. 30261–30273.

Dodd, D., Moon, Y.-H., Swaminathan, K., Mackie, R.I. & Cann, I.K.O. (2010b) 'Transcriptomic Analyses of Xylan Degradation by Prevotella bryantii and Insights into Energy Acquisition by Xylanolytic Bacteroidetes', *Journal of Biological Chemistry*, 285(39), pp. 30261–30273.

Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B. & Terrapon, N. (2022) 'The carbohydrate-active enzyme database: functions and literature', *Nucleic Acids Research*, 50(D1), pp. D571–D577.

Ebringerová, A., Hromádková, Z. & Heinze, T. (n.d.) 'Hemicellulose', in *Polysaccharides I*. [Online]. Berlin/Heidelberg: Springer-Verlag. pp. 1–67.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E. & Relman, D.A. (2005) 'Diversity of the Human Intestinal Microbial Flora', *Science*, 308(5728), pp. 1635–1638.

Emsley, P. & Cowtan, K. (2004) '*Coot*: model-building tools for molecular graphics', *Acta Crystallographica Section D Biological Crystallography*, 60(12), pp. 2126–2132.

Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. (2010) 'Features and development of *Coot*', *Acta Crystallographica Section D Biological Crystallography*, 66(4), pp. 486–501.

Estela, R. & Luis, J. (2013) 'Hydrolysis of Biomass Mediated by Cellulases for the Production of Sugars', in *Sustainable Degradation of Lignocellulosic Biomass - Techniques, Applications and Commercialization*. [Online]. InTech.

Evans, P. (2014) 'CCP4 Tools for X-ray Integration and Data Processing', *Nihon Kessho Gakkaishi*, 56(Supplement), pp. s24–s24.

Evans, P.R. & Murshudov, G.N. (2013) 'How good are my data and what is the resolution?', *Acta Crystallographica Section D Biological Crystallography*, 69(7), pp. 1204–1214.

Fehlner-Peach, H., Magnabosco, C., Raghavan, V., Scher, J.U., Tett, A., Cox, L.M., Gottsegen, C., Watters, A., Wiltshire-Gordon, J.D., Segata, N., Bonneau, R. & Littman, D.R. (2019) 'Distinct Polysaccharide Utilization Profiles of Human Intestinal Prevotella copri Isolates', *Cell Host & Microbe*, 26(5), pp. 680-690.e5.

Fernandez-Julia, P., Commane, D.M., van Sinderen, D. & Munoz-Munoz, J. (2022) 'Cross-feeding interactions between human gut commensals belonging to the Bacteroides and Bifidobacterium genera when grown on dietary glycans', *Microbiome Research Reports*,

Ficko-Blean, E. & Boraston, A.B. (2012) 'Insights into the recognition of the human glycome by microbial carbohydrate-binding modules', *Current Opinion in Structural Biology*, 22(5), pp. 570–577.

Ficko-Blean, E. & Boraston, A.B. (2009) 'N-Acetylglucosamine Recognition by a Family 32 Carbohydrate-Binding Module from Clostridium perfringens NagH', *Journal of Molecular Biology*, 390(2), pp. 208–220.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G. & Lionetti, P. (2010) 'Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa', *Proceedings of the National Academy of Sciences*, 107(33), pp. 14691–14696.

De Filippo, C., Di Paola, M., Ramazzotti, M., Albanese, D., Pieraccini, G., Banci, E., Miglietta, F., Cavalieri, D. & Lionetti, P. (2017) 'Diet, Environments, and Gut Microbiota. A Preliminary Investigation in Children Living in Rural and Urban Burkina Faso and Italy', *Frontiers in Microbiology*, 8.

Fischer, F., Romero, R., Hellhund, A., Linne, U., Bertrams, W., Pinkenburg, O., Eldin, H.S., Binder, K., Jacob, R., Walker, A., Stecher, B., Basic, M., Luu, M., Mahdavi, R., Heintz-Buschart, A., Visekruna, A. & Steinhoff, U. (2020) 'Dietary cellulose induces anti-inflammatory immunity and transcriptional programs via maturation of the intestinal microbiota', *Gut Microbes*, 12(1), p. 1829962.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R. & White, B.A. (2008) 'Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis', *Nature Reviews Microbiology*, 6(2), pp. 121–131.

Flint, H.J., Scott, K.P., Duncan, S.H., Louis, P. & Forano, E. (2012) 'Microbial degradation of complex carbohydrates in the gut', *Gut Microbes*, 3(4), pp. 289–306.

Flint, H.J., Scott, K.P., Louis, P. & Duncan, S.H. (2012) 'The role of the gut microbiota in nutrition and health', *Nature Reviews Gastroenterology & Hepatology*, 9(10), pp. 577–589.

Foley, M.H., Cockburn, D.W. & Koropatkin, N.M. (2016) 'The Sus operon: a model system for starch uptake by the human gut Bacteroidetes', *Cellular and Molecular Life Sciences*, 73(14), pp. 2603–2617.

Fontes, C.M.G.A. & Gilbert, H.J. (2010) 'Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates', *Annual Review of Biochemistry*, 79(1), pp. 655–681.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152.

Fujimoto, Z., Kishine, N., Suzuki, N., Suzuki, R., Mizushima, D., Momma, M., Kimura, K. & Funane, K. (2017) 'Isomaltooligosaccharide-binding structure of *Paenibacillus* sp. 598K cycloisomaltooligosaccharide glucanotransferase', *Bioscience Reports*, 37(2).

Fujimoto, Z., Suzuki, N., Kishine, N., Ichinose, H., Momma, M., Kimura, A. & Funane, K. (2017) 'Carbohydrate-binding architecture of the multi-modular α-1,6-glucosyltransferase from *Paenibacillus* sp. 598K, which produces α-1,6-glucosyl-α-glucosaccharides from starch', *Biochemical Journal*, 474(16), pp. 2763–2778.

Fultz, R., Ticer, T., Ihekweazu, F.D., Horvath, T.D., Haidacher, S.J., Hoch, K.M., Bajaj, M., Spinler, J.K., Haag, A.M., Buffington, S.A. & Engevik, M.A. (2021) 'Unraveling the Metabolic Requirements of the Gut Commensal Bacteroides ovatus', *Frontiers in Microbiology*, 12.

Gálvez, E.J.C., Iljazovic, A., Amend, L., Lesker, T.R., Renault, T., Thiemann, S., Hao, L., Roy, U., Gronow, A., Charpentier, E. & Strowig, T. (2020) 'Distinct Polysaccharide Utilization Determines Interspecies Competition between Intestinal Prevotella spp.', *Cell Host & Microbe*, 28(6), pp. 838-852.e6.

Gardner, J.G. (2016) 'Polysaccharide degradation systems of the saprophytic bacterium Cellvibrio japonicus', *World Journal of Microbiology and Biotechnology*, 32(7), p. 121.

Gaulke, C.A. & Sharpton, T.J. (2018) 'The influence of ethnicity and geography on human gut microbiome composition', *Nature Medicine*, 24(10), pp. 1495–1496.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R. & Whalen, K.L. (2015) 'Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks', *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1854(8), pp. 1019–1037.

Ghosh, A., Verma, A.K., Gautam, S., Gupta, M.N. & Goyal, A. (2014) 'Structure and functional investigation of ligand binding by a family 35 carbohydrate binding module (CtCBM35) of β-mannanase of family 26 glycoside hydrolase from Clostridium thermocellum', *Biochemistry (Moscow)*, 79(7), pp. 672–686.

Gilbert, H.J., Knox, J.P. & Boraston, A.B. (2013) 'Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules', *Current Opinion in Structural Biology*, 23(5), pp. 669–677.

Gírio, F.M., Fonseca, C., Carvalheiro, F., Duarte, L.C., Marques, S. & Bogel-Łukasik, R. (2010) 'Hemicelluloses for fuel ethanol: A review', *Bioresource Technology*, 101(13), pp. 4775–4800.

Giuseppe, P.O., Bonfim, I.M. & Murakami, M.T. (2023) 'Enzymatic systems for carbohydrate utilization and biosynthesis in *Xanthomonas* and their role in pathogenesis and tissue specificity', *Essays in Biochemistry*, 67(3), pp. 455–470.

Glenwright, A.J., Pothula, K.R., Bhamidimarri, S.P., Chorev, D.S., Baslé, A., Firbank, S.J., Zheng, H., Robinson, C. V., Winterhalter, M., Kleinekathöfer, U., Bolam, D.N. & van den Berg, B. (2017) 'Structural basis for nutrient acquisition by dominant members of the human gut microbiota', *Nature*, 541(7637), pp. 407–411.

Glowacki, R.W.P. & Martens, E.C. (2021a) 'If You Eat It or Secrete It, They Will Grow: The Expanding List of Nutrients Utilized by Human Gut Bacteria', *Journal of Bacteriology*, 203(9).

Glowacki, R.W.P. & Martens, E.C. (2021b) 'If You Eat It or Secrete It, They Will Grow: The Expanding List of Nutrients Utilized by Human Gut Bacteria', *Journal of Bacteriology*, 203(9).

Gray, D.A., White, J.B.R., Oluwole, A.O., Rath, P., Glenwright, A.J., Mazur, A., Zahn, M., Baslé, A., Morland, C., Evans, S.L., Cartmell, A., Robinson, C. V., Hiller, S., Ranson, N.A., Bolam, D.N. & van den Berg, B. (2021) 'Insights into SusCD-mediated glycan import by a prominent gut symbiont', *Nature Communications*, 12(1), p. 44.

Gregg, K.J., Finn, R., Abbott, D.W. & Boraston, A.B. (2008) 'Divergent Modes of Glycan Recognition by a New Family of Carbohydrate-binding Modules', *Journal of Biological Chemistry*, 283(18), pp. 12604–12613.

Grondin, J.M., Tamura, K., Déjean, G., Abbott, D.W. & Brumer, H. (2017) 'Polysaccharide Utilization Loci: Fueling Microbial Communities', *Journal of Bacteriology*, 199(15),

Han, S.O., Cho, H.-Y., Yukawa, H., Inui, M. & Doi, R.H. (2004) 'Regulation of Expression of Cellulosomes and Noncellulosomal (Hemi)Cellulolytic Enzymes in *Clostridium cellulovorans* during Growth on Different Carbon Sources', *Journal of Bacteriology*, 186(13), pp. 4218–4227.

Han, S.O., Yukawa, H., Inui, M. & Doi, R.H. (2003) 'Regulation of Expression of Cellulosomal Cellulase and Hemicellulase Genes in *Clostridium cellulovorans*', *Journal of Bacteriology*, 185(20), pp. 6067–6075.

Hehemann, J.-H., Boraston, A.B. & Czjzek, M. (2014) 'A sweet new wave: structures and mechanisms of enzymes that digest polysaccharides from marine algae', *Current Opinion in Structural Biology*, 28pp. 77–86.

Henrissat, B. & Davies, G. (1997) 'Structural and sequence-based classification of glycoside hydrolases', *Current Opinion in Structural Biology*, 7(5), pp. 637–644.

Henshaw, J.L., Bolam, D.N., Pires, V.M.R., Czjzek, M., Henrissat, B., Ferreira, L.M.A., Fontes, C.M.G.A. & Gilbert, H.J. (2004) 'The Family 6 Carbohydrate Binding Module CmCBM6-2 Contains Two Ligand-binding Sites with Distinct Specificities', *Journal of Biological Chemistry*, 279(20), pp. 21552–21559.

Hervé, C., Rogowski, A., Blake, A.W., Marcus, S.E., Gilbert, H.J. & Knox, J.P. (2010) 'Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects', *Proceedings of the National Academy of Sciences*, 107(34), pp. 15293–15298.

Higgins, M.A., Whitworth, G.E., El Warry, N., Randriantsoa, M., Samain, E., Burke, R.D., Vocadlo, D.J. & Boraston, A.B. (2009) 'Differential Recognition and Hydrolysis of Host Carbohydrate Antigens by Streptococcus pneumoniae Family 98 Glycoside Hydrolases', *Journal of Biological Chemistry*, 284(38), pp. 26161–26173.

Hitch, T.C.A., Bisdorf, K., Afrizal, A., Riedel, T., Overmann, J., Strowig, T. & Clavel, T. (2022) 'A taxonomic note on the genus Prevotella: Description of four novel genera and emended description of the genera Hallella and Xylanibacter', *Systematic and Applied Microbiology*, 45(6), p. 126354.

Illiano, P., Brambilla, R. & Parolini, C. (2020) 'The mutual interplay of gut microbiota, diet and human disease', *The FEBS Journal*, 287(5), pp. 833–855.

Jana, U.K., Kango, N. & Pletschke, B. (2021) 'Hemicellulose-Derived Oligosaccharides: Emerging Prebiotics in Disease Alleviation', *Frontiers in Nutrition*, 8.

de Jonge, N., Carlsen, B., Christensen, M.H., Pertoldi, C. & Nielsen, J.L. (2022) 'The Gut Microbiome of 54 Mammalian Species', *Frontiers in Microbiology*, 13.

Jovel, J., Dieleman, L.A., Kao, D., Mason, A.L. & Wine, E. (2018) 'The Human Gut Microbiome in Health and Disease', in *Metagenomics*. [Online]. Elsevier. pp. 197–213.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589.

Kahn, A., Moraïs, S., Galanopoulou, A.P., Chung, D., Sarai, N.S., Hengge, N., Hatzinikolaou, D.G., Himmel, M.E., Bomble, Y.J. & Bayer, E.A. (2019) 'Creation of a functional hyperthermostable designer cellulosome', *Biotechnology for Biofuels*, 12(1), p. 44.

Kaoutari, A. El, Armougom, F., Gordon, J.I., Raoult, D. & Henrissat, B. (2013) 'The abundance and variety of carbohydrate-active enzymes in the human gut microbiota', *Nature Reviews Microbiology*, 11(7), pp. 497–504.

Kelly, W.J., Leahy, S.C., Altermann, E., Yeoman, C.J., Dunne, J.C., Kong, Z., Pacheco, D.M., Li, D., Noel, S.J., Moon, C.D., Cookson, A.L. & Attwood, G.T. (2010) 'The Glycobiome of the Rumen Bacterium Butyrivibrio proteoclasticus B316T Highlights Adaptation to a Polysaccharide-Rich Environment', *PLoS ONE*, 5(8), p. e11942.

Klemm, D., Heublein, B., Fink, H. & Bohn, A. (2005) 'Cellulose: Fascinating Biopolymer and Sustainable Raw Material', *Angewandte Chemie International Edition*, 44(22), pp. 3358–3393.

Koebnik, R. (2005) 'TonB-dependent trans-envelope signalling: the exception or the rule?', *Trends in Microbiology*, 13(8), pp. 343–347.

Koropatkin, N.M., Martens, E.C., Gordon, J.I. & Smith, T.J. (2008) 'Starch Catabolism by a Prominent Human Gut Symbiont Is Directed by the Recognition of Amylose Helices', *Structure*, 16(7), pp. 1105–1115.

KOSHLAND, D.E. (1953) 'STEREOCHEMISTRY AND THE MECHANISM OF ENZYMATIC REACTIONS', *Biological Reviews*, 28(4), pp. 416–436.

Kosugi, A., Murashima, K., Tamaru, Y. & Doi, R.H. (2002) 'Cell-Surface-Anchoring Role of N-Terminal Surface Layer Homology Domains of *Clostridium cellulovorans* EngE', *Journal of Bacteriology*, 184(4), pp. 884–888.

Koutsos, A., Lima, M., Conterno, L., Gasperotti, M., Bianchi, M., Fava, F., Vrhovsek, U., Lovegrove, J. & Tuohy, K. (2017) 'Effects of Commercial Apple Varieties on Human Gut Microbiota Composition and Metabolic Output Using an In Vitro Colonic Model', *Nutrients*, 9(6), p. 533.

Kulkarni, A.R., Peña, M.J., Avci, U., Mazumder, K., Urbanowicz, B.R., Pattathil, S., Yin, Y., O'Neill, M.A., Roberts, A.W., Hahn, M.G., Xu, Y., Darvill, A.G. & York, W.S. (2012) 'The ability of land plants to synthesize glucuronoxylans predates the evolution of tracheophytes', *Glycobiology*, 22(3), pp. 439–451.

Labourel, A., Crouch, L.I., Brás, J.L.A., Jackson, A., Rogowski, A., Gray, J., Yadav, M.P., Henrissat, B., Fontes, C.M.G.A., Gilbert, H.J., Najmudin, S., Baslé, A. & Cuskin, F. (2016) 'The Mechanism by Which Arabinoxylanases Can Recognize Highly Decorated Xylans', *Journal of Biological Chemistry*, 291(42), pp. 22149–22159.

LAEMMLI, U.K. (1970) ‘Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4’, *Nature*, 227(5259), pp. 680–685.

Lairson, L.L., Henrissat, B., Davies, G.J. & Withers, S.G. (2008) ‘Glycosyltransferases: Structures, Functions, and Mechanisms’, *Annual Review of Biochemistry*, 77(1), pp. 521–555.

Lamed, R., Setter, E. & Bayer, E.A. (1983) ‘Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum’, *Journal of Bacteriology*, 156(2), pp. 828–836.

Lamote, B., da Fonseca, M.J.M., Vanderstraeten, J., Meert, K., Elias, M. & Briers, Y. (2023) ‘Current challenges in designer cellulosome engineering’, *Applied Microbiology and Biotechnology*, 107(9), pp. 2755–2770.

Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. (2019) ‘Bacteroidetes use thousands of enzyme combinations to break down glycans’, *Nature Communications*, 10(1), p. 2043.

Larsbrink, J., Rogers, T.E., Hemsworth, G.R., McKee, L.S., Tauzin, A.S., Spadiut, O., Klinter, S., Pudlo, N.A., Urs, K., Koropatkin, N.M., Creagh, A.L., Haynes, C.A., Kelly, A.G., Cederholm, S.N., Davies, G.J., Martens, E.C. & Brumer, H. (2014a) ‘A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes’, *Nature*, 506(7489), pp. 498–502.

Larsbrink, J., Rogers, T.E., Hemsworth, G.R., McKee, L.S., Tauzin, A.S., Spadiut, O., Klinter, S., Pudlo, N.A., Urs, K., Koropatkin, N.M., Creagh, A.L., Haynes, C.A., Kelly, A.G., Cederholm, S.N., Davies, G.J., Martens, E.C. & Brumer, H. (2014b) ‘A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes’, *Nature*, 506(7489), pp. 498–502.

Larson, S.B., Day, J., Barba de la Rosa, A.P., Keen, N.T. & McPherson, A. (2003) ‘First Crystallographic Structure of a Xylanase from Glycoside Hydrolase Family 5: Implications for Catalysis,’ *Biochemistry*, 42(28), pp. 8411–8422.

Latif, A., Shehzad, A., Niazi, S., Zahid, A., Ashraf, W., Iqbal, M.W., Rehman, A., Riaz, T., Aadil, R.M., Khan, I.M., Özogul, F., Rocha, J.M., Esatbeyoglu, T. & Korma, S.A. (2023) ‘Probiotics: mechanism of action, health benefits and their application in food industries’, *Frontiers in Microbiology*, 14.

Leth, M.L., Pichler, M.J. & Abou Hachem, M. (2023) ‘Butyrate-producing colonic clostridia: picky glycan utilization specialists’, *Essays in Biochemistry*, 67(3), pp. 415–428.

Letunic, I. & Bork, P. (2018) ‘20 years of the SMART protein domain annotation resource’, *Nucleic Acids Research*, 46(D1), pp. D493–D496.

Letunic, I. & Bork, P. (2021) ‘Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation’, *Nucleic Acids Research*, 49(W1), pp. W293–W296.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R. & Gordon, J.I. (2008) ‘Evolution of Mammals and Their Gut Microbes’, *Science*, 320(5883), pp. 1647–1651.

Light, S.H., Cahoon, L.A., Halavaty, A.S., Freitag, N.E. & Anderson, W.F. (2016) ‘Structure to function of an α-glucan metabolic pathway that promotes Listeria monocytogenes pathogenesis’, *Nature Microbiology*, 2(2), p. 16202.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P.M. & Henrissat, B. (2010) 'A hierarchical classification of polysaccharide lyases for glycogenomics', *Biochemical Journal*, 432(3), pp. 437–444.

Lopetuso, L.R., Scaldaferri, F., Petito, V. & Gasbarrini, A. (2013) 'Commensal Clostridia: leading players in the maintenance of gut homeostasis', *Gut Pathogens*, 5(1), p. 23.

Lowe, E.C., Baslé, A., Czjzek, M., Firbank, S.J. & Bolam, D.N. (2012) 'A scissor blade-like closing mechanism implicated in transmembrane signaling in a *Bacteroides* hybrid two-component system', *Proceedings of the National Academy of Sciences*, 109(19), pp. 7298–7303.

Ma, X.-Q., Wang, B., Wei, W., Tan, F.-C., Su, H., Zhang, J.-Z., Zhao, C.-Y., Zheng, H.-J., Feng, Y.-Q., Shen, W., Yang, J.-B. & Li, F.-L. (2024) 'Alginate oligosaccharide assimilation by gut microorganisms and the potential role in gut inflammation alleviation', *Applied and Environmental Microbiology*, 90(5), .

Makki, K., Deehan, E.C., Walter, J. & Bäckhed, F. (2018) 'The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease', *Cell Host & Microbe*, 23(6), pp. 705–715.

Malgas, S., Mafa, M.S., Mkabayi, L. & Pletschke, B.I. (2019) 'A mini review of xylanolytic enzymes with regards to their synergistic interactions during hetero-xylan degradation', *World Journal of Microbiology and Biotechnology*, 35(12), p. 187.

Manor, O., Dai, C.L., Kornilov, S.A., Smith, B., Price, N.D., Lovejoy, J.C., Gibbons, S.M. & Magis, A.T. (2020) 'Health and disease markers correlate with gut microbiome composition across thousands of people', *Nature Communications*, 11(1), p. 5206.

Martens, E.C., Chiang, H.C. & Gordon, J.I. (2008a) 'Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont', *Cell Host & Microbe*, 4(5), pp. 447–457.

Martens, E.C., Chiang, H.C. & Gordon, J.I. (2008b) 'Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont', *Cell Host & Microbe*, 4(5), pp. 447–457.

Martens, E.C., Koropatkin, N.M., Smith, T.J. & Gordon, J.I. (2009) 'Complex Glycan Catabolism by the Human Gut Microbiota: The Bacteroidetes Sus-like Paradigm', *Journal of Biological Chemistry*, 284(37), pp. 24673–24677.

Martens, E.C., Lowe, E.C., Chiang, H., Pudlo, N.A., Wu, M., McNulty, N.P., Abbott, D.W., Henrissat, B., Gilbert, H.J., Bolam, D.N. & Gordon, J.I. (2011) 'Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts', *PLoS Biology*, 9(12), p. e1001221.

Martens, E.C., Neumann, M. & Desai, M.S. (2018) 'Interactions of commensal and pathogenic microorganisms with the intestinal mucosal barrier', *Nature Reviews Microbiology*, 16(8), pp. 457–470.

Martens, E.C., Roth, R., Heuser, J.E. & Gordon, J.I. (2009) 'Coordinate Regulation of Glycan Degradation and Polysaccharide Capsule Biosynthesis by a Prominent Human Gut Symbiont', *Journal of Biological Chemistry*, 284(27), pp. 18445–18457.

Matsui, H., Ogata, K., Tajima, K., Nakamura, M., Nagamine, T., Aminov, R.I. & Benno, Y. (2000a) 'Phenotypic Characterization of Polysaccharidases Produced by Four Prevotella Type Strains', *Current Microbiology*, 41(1), pp. 45–49.

Matsui, H., Ogata, K., Tajima, K., Nakamura, M., Nagamine, T., Aminov, R.I. & Benno, Y. (2000b) 'Phenotypic Characterization of Polysaccharidases Produced by Four Prevotella Type Strains', *Current Microbiology*, 41(1), pp. 45–49.

Matsuyama, K., Kishine, N., Fujimoto, Z., Sunagawa, N., Kotake, T., Tsumuraya, Y., Samejima, M., Igarashi, K. & Kaneko, S. (2020a) 'Unique active-site and subsite features in the arabinogalactan-degrading GH43 exo-β-1,3-galactanase from Phanerochaete chrysosporium', *Journal of Biological Chemistry*, 295(52), pp. 18539–18552.

Matsuyama, K., Kishine, N., Fujimoto, Z., Sunagawa, N., Kotake, T., Tsumuraya, Y., Samejima, M., Igarashi, K. & Kaneko, S. (2020b) 'Unique active-site and subsite features in the arabinogalactan-degrading GH43 exo-β-1,3-galactanase from Phanerochaete chrysosporium', *Journal of Biological Chemistry*, 295(52), pp. 18539–18552.

McCarter, J.D. & Stephen Withers, G. (1994) 'Mechanisms of enzymatic glycoside hydrolysis', *Current Opinion in Structural Biology*, 4(6), pp. 885–892.

McCartney, L., Blake, A.W., Flint, J., Bolam, D.N., Boraston, A.B., Gilbert, H.J. & Knox, J.P. (2006) 'Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules', *Proceedings of the National Academy of Sciences*, 103(12), pp. 4765–4770.

McCoy, A.J. (2007) 'Solving structures of protein complexes by molecular replacement with *Phaser*', *Acta Crystallographica Section D Biological Crystallography*, 63(1), pp. 32–41.

McKee, L.S., La Rosa, S.L., Westereng, B., Eijsink, V.G., Pope, P.B. & Larsbrink, J. (2021) 'Polysaccharide degradation by the Bacteroidetes: mechanisms and nomenclature', *Environmental Microbiology Reports*, 13(5), pp. 559–581.

McNulty, N.P., Wu, M., Erickson, A.R., Pan, C., Erickson, B.K., Martens, E.C., Pudlo, N.A., Muegge, B.D., Henrissat, B., Hettich, R.L. & Gordon, J.I. (2013) 'Effects of Diet on Resource Utilization by a Model Human Gut Microbiota Containing Bacteroides cellulosilyticus WH2, a Symbiont with an Extensive Glycobiome', *PLoS Biology*, 11(8), p. e1001637.

Milani, C., Alessandri, G., Mancabelli, L., Mangifesta, M., Lugli, G.A., Viappiani, A., Longhi, G., Anzalone, R., Duranti, S., Turroni, F., Ossiprandi, M.C., van Sinderen, D. & Ventura, M. (2020) 'Multi-omics Approaches To Decipher the Impact of Diet and Host Physiology on the Mammalian Gut Microbiome', *Applied and Environmental Microbiology*, 86(23), .

MOHNEN, D. (2008) 'Pectin structure and biosynthesis', *Current Opinion in Plant Biology*, 11(3), pp. 266–277.

Montanier, C., van Bueren, A.L., Dumon, C., Flint, J.E., Correia, M.A., Prates, J.A., Firbank, S.J., Lewis, R.J., Grondin, G.G., Ghinet, M.G., Gloster, T.M., Herve, C., Knox, J.P., Talbot, B.G., Turkenburg, J.P., Kerovuo, J., Brzezinski, R., Fontes, C.M.G.A., Davies, G.J., *et al.* (2009) 'Evidence that family 35 carbohydrate binding modules display conserved specificity but divergent function', *Proceedings of the National Academy of Sciences*, 106(9), pp. 3065–3070.

Moon, Y.H., Iakiviak, M., Bauer, S., Mackie, R.I. & Cann, I.K.O. (2011) 'Biochemical Analyses of Multiple Endoxylanases from the Rumen Bacterium Ruminococcus albus 8

and Their Synergistic Activities with Accessory Hemicellulose-Degrading Enzymes', *Applied and Environmental Microbiology*, 77(15), pp. 5157–5169.

Moraïs, S., Barak, Y., Caspi, J., Hadar, Y., Lamed, R., Shoham, Y., Wilson, D.B. & Bayer, E.A. (2010) 'Contribution of a Xylan-Binding Module to the Degradation of a Complex Cellulosic Substrate by Designer Cellulosomes', *Applied and Environmental Microbiology*, 76(12), pp. 3787–3796.

Moraïs, S., David, Y. Ben, Bensoussan, L., Duncan, S.H., Koropatkin, N.M., Martens, E.C., Flint, H.J. & Bayer, E.A. (2016) 'Enzymatic profiling of cellulosomal enzymes from the human gut bacterium, <scp> *R* </scp> *uminococcus champanellensis* , reveals a fine-tuned system for cohesin-dockerin recognition', *Environmental Microbiology*, 18(2), pp. 542–556.

Moraïs, S., Winkler, S., Zorea, A., Levin, L., Nagies, F.S.P., Kapust, N., Lamed, E., Artan-Furman, A., Bolam, D.N., Yadav, M.P., Bayer, E.A., Martin, W.F. & Mizrahi, I. (2024) 'Cryptic diversity of cellulose-degrading gut bacteria in industrialized humans', *Science*, 383(6688), .

Morotomi, M., Nagai, F., Sakon, H. & Tanaka, R. (2009) 'Paraprevotella clara gen. nov., sp. nov. and Paraprevotella xylaniphila sp. nov., members of the family "Prevotellaceae" isolated from human faeces', *International journal of systematic and evolutionary microbiology*, 59(8), pp. 1895–1900.

Mortensen, P.B. & Clausen, M.R. (1996) 'Short-Chain Fatty Acids in the Human Colon: Relation to Gastrointestinal Health and Disease', *Scandinavian Journal of Gastroenterology*, 31(sup216), pp. 132–148.

Mullis, K.B. & Faloona, F.A. (1987) *[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction*, in [Online]. pp. 335–350.

Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F. & Vagin, A.A. (2011) '*REFMAC* 5 for the refinement of macromolecular crystal structures', *Acta Crystallographica Section D Biological Crystallography*, 67(4), pp. 355–367.

N I, M. (1984) 'The contribution of the large intestine to energy supplies in man', *The American Journal of Clinical Nutrition*, 39(2), pp. 338–342.

Naas, A.E., Mackenzie, A.K., Mravec, J., Schückel, J., Willats, W.G.T., Eijsink, V.G.H. & Pope, P.B. (2014) 'Do Rumen *Bacteroidetes* Utilize an Alternative Mechanism for Cellulose Degradation?', *mBio*, 5(4),

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

Ndeh, D., Rogowski, A., Cartmell, A., Luis, A.S., Baslé, A., Gray, J., Venditto, I., Briggs, J., Zhang, X., Labourel, A., Terrapon, N., Buffetto, F., Nepogodiev, S., Xiao, Y., Field, R.A., Zhu, Y., O'Neill, M.A., Urbanowicz, B.R., York, W.S., *et al.* (2017) 'Complex pectin metabolism by gut bacteria reveals novel catalytic functions', *Nature*, 544(7648), pp. 65–70.

Neelamegham, S., Aoki-Kinoshita, K., Bolton, E., Frank, M., Lisacek, F., Lütteke, T., O'Boyle, N., Packer, N.H., Stanley, P., Toukach, P., Varki, A., Woods, R.J., Darvill, A., Dell, A., Henrissat, B., Bertozzi, C., Hart, G., Narimatsu, H., Freeze, H., *et al.* (2019)

‘Updates to the Symbol Nomenclature for Glycans guidelines’, *Glycobiology*, 29(9), pp. 620–624.

Neves, A.L.A., Yu, J., Suzuki, Y., Baez-Magana, M., Arutyunova, E., O’Hara, E., McAllister, T., Ominski, K.H., Lemieux, M.J. & Guan, L.L. (2021) ‘Accelerated discovery of novel glycoside hydrolases using targeted functional profiling and selective pressure on the rumen microbiome’, *Microbiome*, 9(1), p. 229.

Noinaj, N., Guillier, M., Barnard, Travis J. & Buchanan, S.K. (2010) ‘TonB-Dependent Transporters: Regulation, Structure, and Function’, *Annual Review of Microbiology*, 64(1), pp. 43–60.

O. Sheridan, P., Martin, J.C., Lawley, T.D., Browne, H.P., Harris, H.M.B., Bernalier-Donadille, A., Duncan, S.H., O’Toole, P.W., P. Scott, K. & J. Flint, H. (2016) ‘Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes’, *Microbial Genomics*, 2(2), .

Okazawa, Y., Miyazaki, T., Yokoi, G., Ishizaki, Y., Nishikawa, A. & Tonozuka, T. (2015) ‘Crystal Structure and Mutational Analysis of Isomalto-dextranase, a Member of Glycoside Hydrolase Family 27’, *Journal of Biological Chemistry*, 290(43), pp. 26339–26349.

Pannu, N.S., Waterreus, W.-J., Skubák, P., Sikharulidze, I., Abrahams, J.P. & de Graaff, R.A.G. (2011) ‘Recent advances in the *CRANK* software suite for experimental phasing’, *Acta Crystallographica Section D Biological Crystallography*, 67(4), pp. 331–337.

Park, S., Baker, J.O., Himmel, M.E., Parilla, P.A. & Johnson, D.K. (2010) ‘Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance’, *Biotechnology for Biofuels*, 3(1), p. 10.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D.H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Orengo, C.A., Pandurangan, A.P., Rivoire, C., *et al.* (2023) ‘InterPro in 2022’, *Nucleic Acids Research*, 51(D1), pp. D418–D427.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H. & Ferrin, T.E. (2021) ‘<scp>UCSF ChimeraX</scp>: Structure visualization for researchers, educators, and developers’, *Protein Science*, 30(1), pp. 70–82.

Pires, V.M.R., Henshaw, J.L., Prates, J.A.M., Bolam, D.N., Ferreira, L.M.A., Fontes, C.M.G.A., Henrissat, B., Planas, A., Gilbert, H.J. & Czjzek, M. (2004a) ‘The Crystal Structure of the Family 6 Carbohydrate Binding Module from Cellvibrio mixtus Endoglucanase 5A in Complex with Oligosaccharides Reveals Two Distinct Binding Sites with Different Ligand Specificities’, *Journal of Biological Chemistry*, 279(20), pp. 21560–21568.

Pires, V.M.R., Henshaw, J.L., Prates, J.A.M., Bolam, D.N., Ferreira, L.M.A., Fontes, C.M.G.A., Henrissat, B., Planas, A., Gilbert, H.J. & Czjzek, M. (2004b) ‘The Crystal Structure of the Family 6 Carbohydrate Binding Module from Cellvibrio mixtus Endoglucanase 5A in Complex with Oligosaccharides Reveals Two Distinct Binding Sites with Different Ligand Specificities’, *Journal of Biological Chemistry*, 279(20), pp. 21560–21568.

Png, C.W., Lindén, S.K., Gilshenan, K.S., Zoetendal, E.G., McSweeney, C.S., Sly, L.I., McGuckin, M.A. & Florin, T.H.J. (2010) 'Mucolytic Bacteria With Increased Prevalence in IBD Mucosa Augment In Vitro Utilization of Mucin by Other Bacteria', *American Journal of Gastroenterology*, 105(11), pp. 2420–2428.

Pollet, R.M., Martin, L.M. & Koropatkin, N.M. (2021) 'TonB-dependent transporters in the Bacteroidetes: Unique domain structures and potential functions', *Molecular Microbiology*, 115(3), pp. 490–501.

Rajeshwari, R., Jha, G. & Sonti, R. V. (2005) 'Role of an In Planta-Expressed Xylanase of *Xanthomonas oryzae* pv. *oryzae* in Promoting Virulence on Rice', *Molecular Plant-Microbe Interactions*®, 18(8), pp. 830–837.

Rakotoarivonina, H., Terrie, C., Chambon, C., Forano, E. & Mosoni, P. (2009) 'Proteomic identification of CBM37-containing cellulases produced by the rumen cellulolytic bacterium Ruminococcus albus 20 and their putative involvement in bacterial adhesion to cellulose', *Archives of Microbiology*, 191(4), pp. 379–388.

Raman, B., McKeown, C.K., Rodriguez, M., Brown, S.D. & Mielenz, J.R. (2011) 'Transcriptomic analysis of Clostridium thermocellumATCC 27405 cellulose fermentation', *BMC Microbiology*, 11(1), p. 134.

Ravachol, J., de Philip, P., Borne, R., Mansuelle, P., Maté, M.J., Perret, S. & Fierobe, H.-P. (2016) 'Mechanisms involved in xyloglucan catabolism by the cellulosome-producing bacterium Ruminiclostridium cellulolyticum', *Scientific Reports*, 6(1), p. 22770.

La Reau, A.J. & Suen, G. (2018) 'The Ruminococci: key symbionts of the gut ecosystem', *Journal of Microbiology*, 56(3), pp. 199–208.

Reeves, A.R., Wang, G.R. & Salyers, A.A. (1997) 'Characterization of four outer membrane proteins that play a role in utilization of starch by Bacteroides thetaiotaomicron', *Journal of Bacteriology*, 179(3), pp. 643–649.

Rincon, M.T., Cepeljnik, T., Martin, J.C., Barak, Y., Lamed, R., Bayer, E.A. & Flint, H.J. (2007) 'A Novel Cell Surface-Anchored Cellulose-Binding Protein Encoded by the *sca* Gene Cluster of *Ruminococcus flavefaciens*', *Journal of Bacteriology*, 189(13), pp. 4774–4783.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G., Gasbarrini, A. & Mele, M. (2019) 'What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases', *Microorganisms*, 7(1), p. 14.

Robert, C. & Bernalier-Donadille, A. (2003) 'The cellulolytic microflora of the human colon: evidence of microcrystalline cellulose-degrading bacteria in methane-excreting subjects', *FEMS Microbiology Ecology*, 46(1), pp. 81–89.

Robert, C., Chassard, C., Lawson, P.A. & Bernalier-Donadille, A. (2007) 'Bacteroides cellulosilyticus sp. nov., a cellulolytic bacterium from the human gut microbial community', *International Journal of Systematic and Evolutionary Microbiology*, 57(7), pp. 1516–1520.

Rogowski, A., Briggs, J.A., Mortimer, J.C., Tryfona, T., Terrapon, N., Lowe, E.C., Baslé, A., Morland, C., Day, A.M., Zheng, H., Rogers, T.E., Thompson, P., Hawkins, A.R., Yadav, M.P., Henrissat, B., Martens, E.C., Dupree, P., Gilbert, H.J. & Bolam, D.N. (2015) 'Glycan complexity dictates microbial resource allocation in the large intestine', *Nature Communications*, 6(1), p. 7481.

Rongpipi, S., Ye, D., Gomez, E.D. & Gomez, E.W. (2019) 'Progress and Opportunities in the Characterization of Cellulose – An Important Regulator of Cell Wall Growth and Mechanics', *Frontiers in Plant Science*, 9.

La Rosa, S.L., Kachrimanidou, V., Buffetto, F., Pope, P.B., Pudlo, N.A., Martens, E.C., Rastall, R.A., Gibson, G.R. & Westereng, B. (2019) 'Wood-Derived Dietary Fibers Promote Beneficial Human Gut Microbiota', *mSphere*, 4(1),

La Rosa, S.L., Leth, M.L., Michalak, L., Hansen, M.E., Pudlo, N.A., Glowacki, R., Pereira, G., Workman, C.T., Arntzen, M.Ø., Pope, P.B., Martens, E.C., Hachem, M.A. & Westereng, B. (2019) 'The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary β-mannans', *Nature Communications*, 10(1), p. 905.

La Rosa, S.L., Ostrowski, M.P., Vera-Ponce de León, A., McKee, L.S., Larsbrink, J., Eijsink, V.G., Lowe, E.C., Martens, E.C. & Pope, P.B. (2022a) 'Glycan processing in gut microbiomes', *Current Opinion in Microbiology*, 67p. 102143.

La Rosa, S.L., Ostrowski, M.P., Vera-Ponce de León, A., McKee, L.S., Larsbrink, J., Eijsink, V.G., Lowe, E.C., Martens, E.C. & Pope, P.B. (2022b) 'Glycan processing in gut microbiomes', *Current Opinion in Microbiology*, 67p. 102143.

Russell, J.B., Muck, R.E. & Weimer, P.J. (2009) 'Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen', *FEMS Microbiology Ecology*, 67(2), pp. 183–197.

Sainz-Polo, M.A., Valenzuela, S.V., González, B., Pastor, F.I.J. & Sanz-Aparicio, J. (2014) 'Structural Analysis of Glucuronoxylan-specific Xyn30D and Its Attached CBM35 Domain Gives Insights into the Role of Modularity in Specificity*', *Journal of Biological Chemistry*, 289(45), pp. 31088–31101.

Sakka, M., Goto, M., Fujino, T., Fujino, E., Karita, S., Kimura, T. & Sakka, K. (2010) 'Analysis of a *Clostridium josui* Cellulase Gene Cluster Containing the *man5A* Gene and Characterization of Recombinant Man5A', *Bioscience, Biotechnology, and Biochemistry*, 74(10), pp. 2077–2082.

Salyers, A.A., West, S.E., Vercellotti, J.R. & Wilkins, T.D. (1977) 'Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon', *Applied and Environmental Microbiology*, 34(5), pp. 529–533.

Samanta, A.K., Jayapal, N., Jayaram, C., Roy, S., Kolte, A.P., Senani, S. & Sridhar, M. (2015) 'Xylooligosaccharides as prebiotics from agricultural by-products: Production and applications', *Bioactive Carbohydrates and Dietary Fibre*, 5(1), pp. 62–71.

Santos, C.R., Hoffmam, Z.B., de Matos Martins, V.P., Zanphorlin, L.M., de Paula Assis, L.H., Honorato, R.V., Lopes de Oliveira, P.S., Ruller, R. & Murakami, M.T. (2014) 'Molecular Mechanisms Associated with Xylan Degradation by Xanthomonas Plant Pathogens', *Journal of Biological Chemistry*, 289(46), pp. 32186–32200.

Scheller, H.V. & Ulvskov, P. (2010a) 'Hemicelluloses', *Annual Review of Plant Biology*, 61(1), pp. 263–289.

Scheller, H.V. & Ulvskov, P. (2010b) 'Hemicelluloses', *Annual Review of Plant Biology*, 61(1), pp. 263–289.

Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G.,

Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry, A.G. & Crittenden, A.N. (2014) 'Gut microbiome of the Hadza hunter-gatherers', *Nature Communications*, 5(1), p. 3654.

Scott, K.P., Duncan, S.H. & Flint, H.J. (2008) 'Dietary fibre and the gut microbiota', *Nutrition Bulletin*, 33(3), pp. 201–211.

Slavin, J.L. (1981) *Apparent Digestibility of Dietary Fiber and its Components in Human Subjects*. [Online]. University of Wisconsin.

Solé, M., Scheibner, F., Hoffmeister, A.-K., Hartmann, N., Hause, G., Rother, A., Jordan, M., Lautier, M., Arlat, M. & Büttner, D. (2015) 'Xanthomonas campestris pv. vesicatoria Secretes Proteases and Xylanases via the Xps Type II Secretion System and Outer Membrane Vesicles', *Journal of Bacteriology*, 197(17), pp. 2879–2893.

Sonnenburg, E.D., Sonnenburg, J.L., Manchester, J.K., Hansen, E.E., Chiang, H.C. & Gordon, J.I. (2006) 'A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing *in vivo* to carbohydrate metabolism', *Proceedings of the National Academy of Sciences*, 103(23), pp. 8834–8839.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J., Buhler, J.D. & Gordon, J.I. (2005a) 'Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont', *Science*, 307(5717), pp. 1955–1959.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J., Buhler, J.D. & Gordon, J.I. (2005b) 'Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont', *Science*, 307(5717), pp. 1955–1959.

Sprockett, D.D. & Coyte, K.Z. (2023) 'When microbes go missing: Understanding the impact of diversity loss within the gut microbiome', *Cell Host & Microbe*, 31(8), pp. 1249–1251.

Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R. & Watson, M. (2019) 'Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery', *Nature Biotechnology*, 37(8), pp. 953–961.

Stokes, M.R. & Chen, J. (1994) 'Effects of an Enzyme-Inoculant Mixture on the Course of Fermentation of Corn Silage', *Journal of Dairy Science*, 77(11), pp. 3401–3409.

Studier, F.W. (1991) 'Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system', *Journal of Molecular Biology*, 219(1), pp. 37–44.

Su, G., Morris, J.H., Demchak, B. & Bader, G.D. (2014) 'Biological Network Exploration with Cytoscape 3', *Current Protocols in Bioinformatics*, 47(1),

Sun, H.-N., Yu, C.-M., Fu, H.-H., Wang, P., Fang, Z.-G., Zhang, Y.-Z., Chen, X.-L. & Zhao, F. (2021) 'Diversity of Marine 1,3-Xylan-Utilizing Bacteria and Characters of Their Extracellular 1,3-Xylanases', *Frontiers in Microbiology*, 12.

Suzuki, N., Fujimoto, Z., Kim, Y.-M., Momma, M., Kishine, N., Suzuki, R., Suzuki, S., Kitamura, S., Kobayashi, M., Kimura, A. & Funane, K. (2014) 'Structural Elucidation of the Cyclization Mechanism of α-1,6-Glucan by Bacillus circulans T-3040 Cycloisomaltooligosaccharide Glucanotransferase', *Journal of Biological Chemistry*, 289(17), pp. 12040–12051.

Swart, C. (2013) *Production of Libraries to Study Biopolymer Metabolism in Arabidopsis thaliana and Tylosema esculentum.* [Online]. University of Stellenbosch.

Szabó, L., Jamal, S., Xie, H., Charnock, S.J., Bolam, D.N., Gilbert, H.J. & Davies, G.J. (2001) 'Structure of a Family 15 Carbohydrate-binding Module in Complex with Xylopentaose', *Journal of Biological Chemistry*, 276(52), pp. 49061–49065.

Tamura, K. & Brumer, H. (2021) 'Glycan utilization systems in the human gut microbiota: a gold mine for structural discoveries', *Current Opinion in Structural Biology*, 68pp. 26–40.

Tamura, K., Stecher, G. & Kumar, S. (2021) 'MEGA11: Molecular Evolutionary Genetics Analysis Version 11', *Molecular Biology and Evolution*, 38(7), pp. 3022–3027.

Tan, H. & Nie, S. (2020) 'Deciphering diet-gut microbiota-host interplay: Investigations of pectin', *Trends in Food Science & Technology*, 106pp. 171–181.

Tancula, E., Feldhaus, M.J., Bedzyk, L.A. & Salyers, A.A. (1992) 'Location and characterization of genes involved in binding of starch to the surface of Bacteroides thetaiotaomicron', *Journal of Bacteriology*, 174(17), pp. 5609–5616.

Tauzin, A.S., Kwiatkowski, K.J., Orlovsky, N.I., Smith, C.J., Creagh, A.L., Haynes, C.A., Wawrzak, Z., Brumer, H. & Koropatkin, N.M. (2016) 'Molecular Dissection of Xyloglucan Recognition in a Prominent Human Gut Symbiont', *mBio*, 7(2),

Terrapon, N., Lombard, V., Drula, É., Lapébie, P., Al-Masaudi, S., Gilbert, H.J. & Henrissat, B. (2018) 'PULDB: the expanded database of Polysaccharide Utilization Loci', *Nucleic Acids Research*, 46(D1), pp. D677–D683.

Terrapon, N., Lombard, V., Gilbert, H.J. & Henrissat, B. (2015) 'Automatic prediction of polysaccharide utilization loci in Bacteroidetes species', *Bioinformatics*, 31(5), pp. 647–655.

Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., von Heijne, G. & Nielsen, H. (2022) 'SignalP 6.0 predicts all five types of signal peptides using protein language models', *Nature Biotechnology*, 40(7), pp. 1023–1025.

Thursby, E. & Juge, N. (2017) 'Introduction to the human gut microbiota', *Biochemical Journal*, 474(11), pp. 1823–1836.

Timilsina, S., Potnis, N., Newberry, E.A., Liyanapathiranage, P., Iruegas-Bocardo, F., White, F.F., Goss, E.M. & Jones, J.B. (2020) 'Xanthomonas diversity, virulence and plant–pathogen interactions', *Nature Reviews Microbiology*, 18(8), pp. 415–427.

Torres, A.G., Redford, P., Welch, R.A. & Payne, S.M. (2001) 'TonB-Dependent Systems of Uropathogenic *Escherichia coli*: Aerobactin and Heme Transport and TonB Are Required for Virulence in the Mouse', *Infection and Immunity*, 69(10), pp. 6179–6185.

Tunnicliffe, R.B., Bolam, D.N., Pell, G., Gilbert, H.J. & Williamson, M.P. (2005) 'Structure of a Mannan-specific Family 35 Carbohydrate-Binding Module: Evidence for Significant Conformational Changes upon Ligand Binding', *Journal of Molecular Biology*, 347(2), pp. 287–296.

Ventura, M., Turroni, F., Motherway, M.O., MacSharry, J. & van Sinderen, D. (2012) 'Host–microbe interactions that facilitate gut colonization by commensal bifidobacteria', *Trends in Microbiology*, 20(10), pp. 467–476.

Viborg, A.H., Katayama, T., Arakawa, T., Abou Hachem, M., Lo Leggio, L., Kitaoka, M., Svensson, B. & Fushinobu, S. (2017) ‘Discovery of α-l-arabinopyranosidases from human gut microbiome expands the diversity within glycoside hydrolase family 42’, *Journal of Biological Chemistry*, 292(51), pp. 21092–21101.

Vinolo, M.A.R., Rodrigues, H.G., Nachbar, R.T. & Curi, R. (2011) ‘Regulation of Inflammation by Short Chain Fatty Acids’, *Nutrients*, 3(10), pp. 858–876.

Vodovnik, MaÅ¡a, Duncan, S.H., Reid, M.D., Cantlay, L., Turner, K., Parkhill, J., Lamed, R., Yeoman, C.J., Miller, M.E.Berg., White, B.A., Bayer, E.A., MarinÅ¡ek-Logar, R. & Flint, H.J. (2013) ‘Correction: Expression of Cellulosome Components and Type IV Pili within the Extracellular Proteome of Ruminococcus flavefaciens 007’, *PLoS ONE*, 8(12), .

Vodovnik, Maša, Duncan, S.H., Reid, M.D., Cantlay, L., Turner, K., Parkhill, J., Lamed, R., Yeoman, C.J., Miller, M.E.Berg., White, B.A., Bayer, E.A., Marinšek-Logar, R. & Flint, H.J. (2013) ‘Expression of Cellulosome Components and Type IV Pili within the Extracellular Proteome of Ruminococcus flavefaciens 007’, *PLoS ONE*, 8(6), p. e65333.

Wang, K., Pereira, G. V., Cavalcante, J.J. V., Zhang, M., Mackie, R. & Cann, I. (2016a) ‘Bacteroides intestinalis DSM 17393, a member of the human colonic microbiome, upregulates multiple endoxylanases during growth on xylan’, *Scientific Reports*, 6(1), p. 34360.

Wang, K., Pereira, G. V., Cavalcante, J.J. V., Zhang, M., Mackie, R. & Cann, I. (2016b) ‘Bacteroides intestinalis DSM 17393, a member of the human colonic microbiome, upregulates multiple endoxylanases during growth on xylan’, *Scientific Reports*, 6(1), p. 34360.

Wang, S., Xiao, Y., Tian, F., Zhao, J., Zhang, H., Zhai, Q. & Chen, W. (2020) ‘Rational use of prebiotics for gut microbiota alterations: Specific bacterial phylotypes and related mechanisms’, *Journal of Functional Foods*, 66p. 103838.

Wang, Y. & LaPointe, G. (2020) ‘Arabinogalactan Utilization by Bifidobacterium longum subsp. longum NCC 2705 and Bacteroides caccae ATCC 43185 in Monoculture and Coculture’, *Microorganisms*, 8(11), p. 1703.

Wang, Y., Ma, M., Dai, W., Shang, Q. & Yu, G. (2024) ‘Bacteroides salyersiae is a potent chondroitin sulfate-degrading species in the human gut microbiota’, *Microbiome*, 12(1), p. 41.

Weiser, J.N., Ferreira, D.M. & Paton, J.C. (2018) ‘Streptococcus pneumoniae: transmission, colonization and invasion’, *Nature Reviews Microbiology*, 16(6), pp. 355–367.

Wen, S., Yuan, G., Li, C., Xiong, Y., Zhong, X. & Li, X. (2022) ‘High cellulose dietary intake relieves asthma inflammation through the intestinal microbiome in a mouse model’, *PLOS ONE*, 17(3), p. e0263762.

Wertheim, H.F.L., Nghia, H.D.T., Taylor, W. & Schultsz, C. (2009) ‘*Streptococcus suis:* An Emerging Human Pathogen’, *Clinical Infectious Diseases*, 48(5), pp. 617–625.

White, J.B.R., Silale, A., Feasey, M., Heunis, T., Zhu, Y., Zheng, H., Gajbhiye, A., Firbank, S., Baslé, A., Trost, M., Bolam, D.N., van den Berg, B. & Ranson, N.A. (2023a) ‘Outer membrane utilisomes mediate glycan uptake in gut Bacteroidetes’, *Nature*, 618(7965), pp. 583–589.

White, J.B.R., Silale, A., Feasey, M., Heunis, T., Zhu, Y., Zheng, H., Gajbhiye, A., Firbank, S., Baslé, A., Trost, M., Bolam, D.N., van den Berg, B. & Ranson, N.A. (2023b) 'Outer membrane utilisomes mediate glycan uptake in gut Bacteroidetes', *Nature*, 618(7965), pp. 583–589.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A. & Wilson, K.S. (2011) 'Overview of the *CCP* 4 suite and current developments', *Acta Crystallographica Section D Biological Crystallography*, 67(4), pp. 235–242.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F.D. & Lewis, J.D. (2011a) 'Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes', *Science*, 334(6052), pp. 105–108.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F.D. & Lewis, J.D. (2011b) 'Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes', *Science*, 334(6052), pp. 105–108.

Wu, H., Crost, E.H., Owen, C.D., van Bakel, W., Martínez Gascueña, A., Latousakis, D., Hicks, T., Walpole, S., Urbanowicz, P.A., Ndeh, D., Monaco, S., Sánchez Salom, L., Griffiths, R., Reynolds, R.S., Colvile, A., Spencer, D.I.R., Walsh, M., Angulo, J. & Juge, N. (2021) 'The human gut symbiont Ruminococcus gnavus shows specificity to blood group A antigen during mucin glycan foraging: Implication for niche colonisation in the gastrointestinal tract', *PLOS Biology*, 19(12), p. e3001498.

Xia, Z., Cui, Y., Zhang, A., Tang, T., Peng, L., Huang, C., Yang, C. & Liao, X. (2022) 'A Review of Parallel Implementations for the Smith–Waterman Algorithm', *Interdisciplinary Sciences: Computational Life Sciences*, 14(1), pp. 1–14.

Xie, T., Jin, F., Jia, X., Mao, H., Xu, Y. & Zhang, S. (2022a) 'High cellulose diet promotes intestinal motility through regulating intestinal immune homeostasis and serotonin biosynthesis', *Biological Chemistry*, 403(3), pp. 279–292.

Xie, T., Jin, F., Jia, X., Mao, H., Xu, Y. & Zhang, S. (2022b) 'High cellulose diet promotes intestinal motility through regulating intestinal immune homeostasis and serotonin biosynthesis', *Biological Chemistry*, 403(3), pp. 279–292.

Xu, C., Huang, R., Teng, L., Jing, X., Hu, J., Cui, G., Wang, Y., Cui, Q. & Xu, J. (2015) 'Cellulosome stoichiometry in Clostridium cellulolyticum is regulated by selective RNA processing and stabilization', *Nature Communications*, 6(1), p. 6900.

Xu, C., Huang, R., Teng, L., Wang, D., Hemme, C.L., Borovok, I., He, Q., Lamed, R., Bayer, E.A., Zhou, J. & Xu, J. (2013) 'Structure and regulation of the cellulose degradome in Clostridium cellulolyticum', *Biotechnology for Biofuels*, 6(1), p. 73.

Xu, H., Reuhs, B.L., Cantu-Jungles, T.M., Tuncil, Y.E., Kaur, A., Terekhov, A., Martens, E.C. & Hamaker, B.R. (2022a) 'Corn arabinoxylan has a repeating structure of subunits of high branch complexity with slow gut microbiota fermentation', *Carbohydrate Polymers*, 289p. 119435.

Xu, H., Reuhs, B.L., Cantu-Jungles, T.M., Tuncil, Y.E., Kaur, A., Terekhov, A., Martens, E.C. & Hamaker, B.R. (2022b) 'Corn arabinoxylan has a repeating structure of subunits of

high branch complexity with slow gut microbiota fermentation', *Carbohydrate Polymers*, 289p. 119435.

Xu, H., Reuhs, B.L., Cantu-Jungles, T.M., Tuncil, Y.E., Kaur, A., Terekhov, A., Martens, E.C. & Hamaker, B.R. (2022c) 'Corn arabinoxylan has a repeating structure of subunits of high branch complexity with slow gut microbiota fermentation', *Carbohydrate Polymers*, 289p. 119435.

Zeybek, N., Rastall, R.A. & Buyukkileci, A.O. (2020) 'Utilization of xylan-type polysaccharides in co-culture fermentations of Bifidobacterium and Bacteroides species', *Carbohydrate Polymers*, 236p. 116076.

Zhang, B., Zhong, Y., Dong, D., Zheng, Z. & Hu, J. (2022a) 'Gut microbial utilization of xylan and its implication in gut homeostasis and metabolic response', *Carbohydrate Polymers*, 286p. 119271.

Zhang, B., Zhong, Y., Dong, D., Zheng, Z. & Hu, J. (2022b) 'Gut microbial utilization of xylan and its implication in gut homeostasis and metabolic response', *Carbohydrate Polymers*, 286p. 119271.

Zhang, H. & Wang, S. (2013) 'Rice versus Xanthomonas oryzae pv. oryzae: a unique pathosystem', *Current Opinion in Plant Biology*, 16(2), pp. 188–195.

Zhang, M., Chekan, J.R., Dodd, D., Hong, P.-Y., Radlinski, L., Revindran, V., Nair, S.K., Mackie, R.I. & Cann, I. (2014) 'Xylan utilization in human gut commensal bacteria is orchestrated by unique modular organization of polysaccharide-degrading enzymes', *Proceedings of the National Academy of Sciences*, 111(35), .

Zheng, J., Hu, B., Zhang, X., Ge, Q., Yan, Y., Akresi, J., Piyush, V., Huang, L. & Yin, Y. (2023) 'dbCAN-seq update: CAZyme gene clusters and substrates in microbiomes', *Nucleic Acids Research*, 51(D1), pp. D557–D563.

9. Appendix

Table 9-1: Cloning primers and E. coli expression vectors used in this study

Gene	Gene Construct	Vector	Forward Primer	Reverse Primer
BACOVA_03433	Bo98	pET21a	CTC GCT AGC CCC GGT ACG GCT GTC	CTC CTC GAG TTT TCT TTC GAT AAC AAT GTT G
BACOVA_03433	Bo98ΔDUF	pET21a	CTC GCT AGC CTT TTT AAA GGA GAC AAT CCC	CTC CTC GAG TTT TCT TTC GAT AAC AAT GTT G
BACOVA_03433	Bo98ΔCBM	pET21a	CTC GCT AGC CCC GGT ACG GCT GTC	CTC CTC GAG GCC CTG ATA AAT TTG TGG AG
BACOVA_03433	Bo98CBM35	pGEX-6P-1	CTC GGA TCC CAG GGC GCT TAT CAG	CTC GAA TTC TTT TCT TTC GAT AAC AAT GTT G
RUM_RS09975	Rc98	pET21a	CTC GCT AGC TCG CCG AAG GGC AC	CTC CTC GAG GCC GGG ATC ATC TGC
RUM_03400	RC_GH10A	pET21a	CTC GCT AGC GAG GTG GAG TTA GAG GAT ATC G	CTC CTC GAG GGT TAC CGG CTC CAG ACC
RUM_23130	RC_GH11A-CE	pET21a	CTC GCT AGC GCC ACG GTC ATT ACC GAA AAC	CTC CTC GAG GTT GCA GGC GTT GTA CAG G

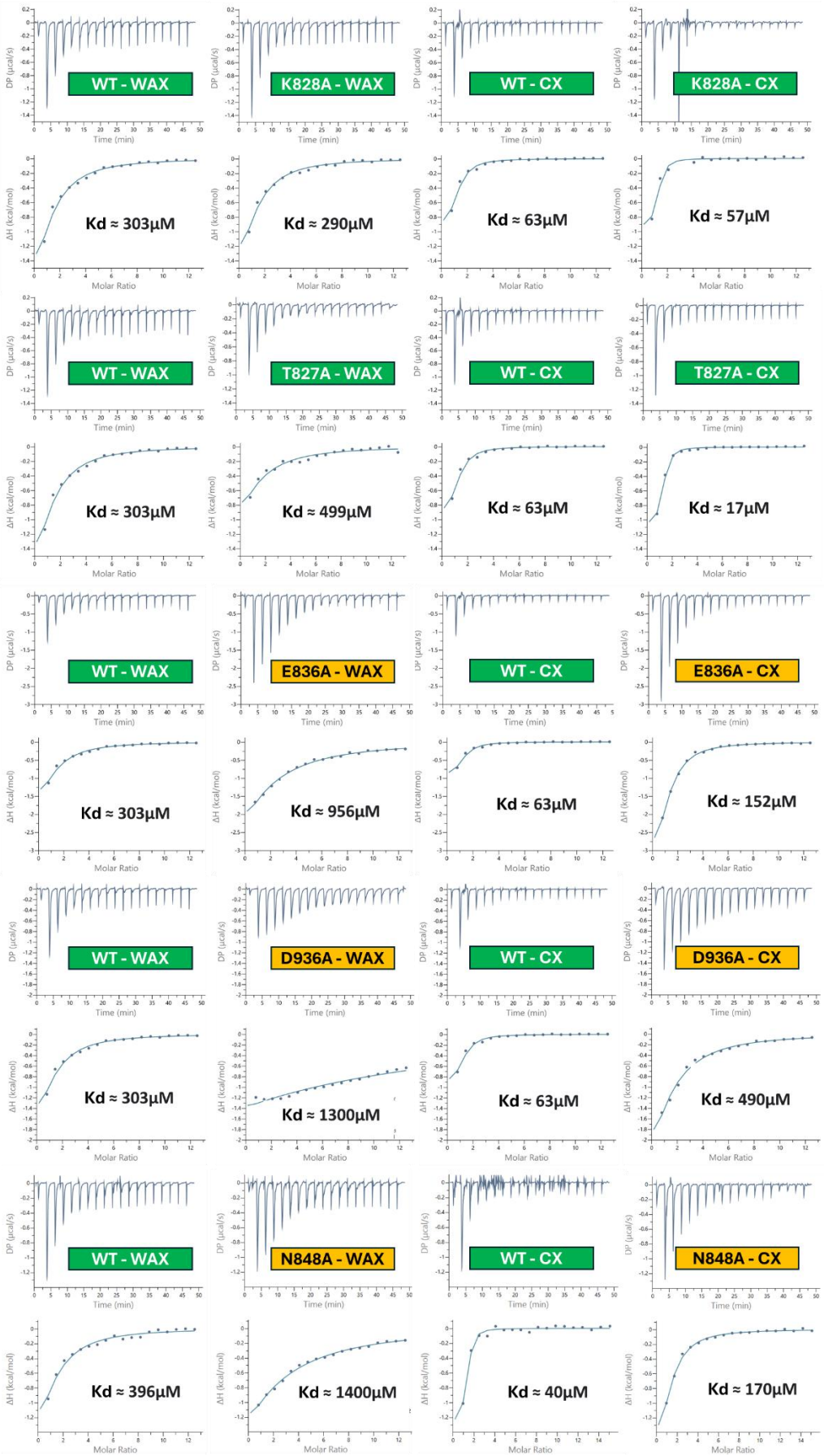
Table 9-2: Primers used for site directed mutagenesis in this study

Gene Construct	Residue Change	Forward Primer	Reverse Primer
Bo98	E361A	GTG TTC AAT TAT TGT GCG CAG TTT TGG GGG TAT G	C ATA CCC CCA AAA CTG CGC ACA ATA ATT GAA CAC
Bo98	D467A	GGC ATC CGT TTC GCG CAA TGC GGC TGG	CCA GCC GCA TTG CGC GAA ACG GAT GCC
Bo98	W253A	GTT CAT ATT GAT TCA GCG AAT TAT GCC GAT CCT C	G AGG ATC GGC ATA ATT CGC TGA ATC AAT ATG AAC
Bo98	E285A	GC CAT GAT GCG GCT ACT GGG	CCC AGT AGC CGC ATC ATG GC
Bo98	W364A	GT GAG CAG TTT GCG GGG TAT GAT GAC	GTC ATC ATA CCC CGC AAA CTG CTC AC
Bo98	W404A	GT GGT AAT ACT GCG TCG GCT AAT ATC	GAT ATT AGC CGA CGC AGT ATT ACC AC
Bo98	S320A	CAG CCG TCA GCC GGC GGT TTT AG	CT AAA ACC GCC GGC TGA CGG CTG
Bo98	S282A	CG TTG TCT GTA GCC CAT GAT GAG	CTC ATC ATG GGC TAC AGA CAA CG
Bo98	S280A	C TCG TTG GCT GTA AGC CAT G	C ATG GCT TAC AGC CAA CGA G
Bo98	W511A	GGA CCT GAA CTT ATA GCG CAG CAA TGT TTT AAG	CTT AAA ACA TTG CTG CGC TAT AAG TTC AGG TCC
Bo98	R465A	CAG TAC GGC ATC GCT TTC GAC CAA TGC	GCA TTG GTC GAA AGC GAT GCC GTA CTG
Bo98	E508A	GTT ATT GAC GGA CCT GCA CTT ATA TGG CAG CAA TG	CA TTG CTG CCA TAT AAG TGC AGG TCC GTC AAT AAC
Bo98	K436A	C ATT ATG TGT GAG GCA TAC ACC ACG CAA AG	CT TTG CGT GGT GTA TGC CTC ACA CAT AAT G
Bo98	E473A	GGC TGG ACA GCG GAG AAA GG	CC TTT CTC CGC TGT CCA GCC
Bo98	W253A	GTT CAT ATT GAT TCA GCG AAT TAT GCC GAT CCT C	G AGG ATC GGC ATA ATT CGC TGA ATC AAT ATG AAC
Bo98	T438A	G TGT GAG AAA TAC GCC ACG CAA AGT GGC	GCC ACT TTG CGT GGC GTA TTT CTC ACA C
Bo98	T827A	GAT TTT AAG AAT GTG GCT AAG AGA GTG ACA AAA GG	CC TTT TGT CAC TCT CTT AGC CAC ATT CTT AAA ATC
Bo98	K828A	GAT TTT AAG AAT GTG ACT GCG AGA GTG ACA AAA GGA G	C TCC TTT TGT CAC TCT CGC AGT CAC ATT CTT AAA ATC
Bo98	E836A	CA AAA GGA GAT TCG GCG CCG ATA CGT AAC	GTT ACG TAT CGG CGC CGA ATC TCC TTT TG
Bo98	N848A	GCA CAG GGA TAT ATA GCT TTT GGT GCT AGT TCT G	C AGA ACT AGC ACC AAA AGC TAT ATA TCC CTG TGC
Bo98	D936A	CA TCA GGT GCC GGT GCC TTG TAT CTC GAC	GTC GAG ATA CAA GGC ACC GGC ACC TGA TG
Bo98CBM35	T827A	GAT TTT AAG AAT GTG GCT AAG AGA GTG ACA AAA GG	CC TTT TGT CAC TCT CTT AGC CAC ATT CTT AAA ATC
Bo98CBM35	K828A	GAT TTT AAG AAT GTG ACT GCG AGA GTG ACA AAA GGA G	C TCC TTT TGT CAC TCT CGC AGT CAC ATT CTT AAA ATC
Bo98CBM35	E836A	CA AAA GGA GAT TCG GCG CCG ATA CGT AAC	GTT ACG TAT CGG CGC CGA ATC TCC TTT TG
Bo98CBM35	N848A	GCA CAG GGA TAT ATA GCT TTT GGT GCT AGT TCT G	C AGA ACT AGC ACC AAA AGC TAT ATA TCC CTG TGC
Bo98CBM35	D936A	CA TCA GGT GCC GGT GCC TTG TAT CTC GAC	GTC GAG ATA CAA GGC ACC GGC ACC TGA TG
Bo98CBM35	Y938A	GCC GGT GAC TTG GCT CTC GAC AAC ATT G	C AAT GTT GTC GAG AGC CAA GTC ACC GGC
Bo98CBM35	K824A	GC TTT GAT TTT GCG AAT GTG ACT AAG	CTT AGT CAC ATT CGC AAA ATC AAA GC
Bo98CBM35	V826A	GAT TTT AAG AAT GCG ACT AAG AGA GTG	CAC TCT CTT AGT CGC ATT CTT AAA ATC
Bo98CBM35	P837A	GA GAT TCG GAG GCG ATA CGT AAC TAC	GTA GTT ACG TAT CGC CTC CGA ATC TC
Bo98CBM35	I838A	GCG GAG CCG GCA CGT AAC TAC AC	GT GTA GTT ACG TGC CGG CTC CGC
Bo98CBM35	Y846A	CA GCA CAG GGA GCT ATA AAT TTT GG	CC AAA ATT TAT AGC TCC CTG TGC TG
Bo98CBM35	K920A	GTT TCC TTA CGA GCA GGG GCA AAT ACG	CGT ATT TGC CCC TGC TCG TAA GGA AAC
Bo98CBM35	L864A	CT GTT ACT GCA GCG GAA GAT GGT G	C ACC ATC TTC CGC TGC AGT AAC AG
Bo98CBM35	R946A	C ATT GTT ATC GAA GCA AAA GAA TTC CCG	CGG GAA TTC TTT TGC TTC GAT AAC AAT G
Bo98CBM35	Y814A	CAG GGC GCT GCT CAG TAT GAG	CTC ATA CTG AGC AGC GCC CTG
Bo98CBM35	Y869A	GAA GAT GGT GTA GCT ACA ATC CGT ATT C	G AAT ACG GAT TGT AGC TAC ACC ATC TTC
Bo98CBM35	D859A	GCC GTA CGT GCT GCT GTT ACT G	C AGT AAC AGC AGC ACG TAC GGC
Bo98CBM35	Y816A	GCT TAT CAG GCT GAG GCT GAA TG	CA TTC AGC CTC AGC CTG ATA AGC
Bo98CBM35	E926A	GGG GCA AAT ACG TTT GCA TTG AAA GCC AAT TC	GA ATT GGC TTT CAA TGC AAA CGT ATT TGC CCC
Bo98CBM35	Y888A	GTC AAT ACC GTA GAT ATG GCT ATA AAC AAT ACC AAA GTA	C TAC TTT GGT ATT GTT TAT AGC CAT ATC TAC GGT ATT GA
Bo98CBM35	R858A	CA GCT GCC GTA GCT GAT GCT GTT AC	GT AAC AGC ATC AGC TAC GGC AGC TG
Bo98CBM35	E898A	GTA GGT ACT CCG GCG TTT GCA CAA ACT G	C AGT TTG TGC AAA CGC CGG AGT ACC TAC
Bo98CBM35	L913A	GTA TGG AAT ACA GCC GCA ATG TCG GTT TCC TTA C	G TAA GGA AAC CGA CAT TGC GGC TGT ATT CCA TAC
Bo98CBM35	N910A	GAC AAT ACT GTA TGG GCT ACA GCC TTA ATG TCG	CGA CAT TAA GGC TGT AGC CCA TAC AGT ATT GTC
Bo98CBM35	K928A	CA AAT ACG TTT GAA TTG GCA GCC AAT TCA TCA GG	CC TGA TGA ATT GGC TGC CAA TTC AAA CGT ATT TG
Bo98CBM35	N891A	GTA GAT ATG TAT ATA AAC GCT ACC AAA GTA GGT ACT CCG	CGG AGT ACC TAC TTT GGT AGC GTT TAT ATA CAT ATC TAC
Bo98CBM35	D886A	GCT ACA GTC AAT ACC GTA GCT ATG TAT ATA AAC AAT ACC	GGT ATT GTT TAT ATA CAT AGC TAC GGT ATT GAC TGT AGC

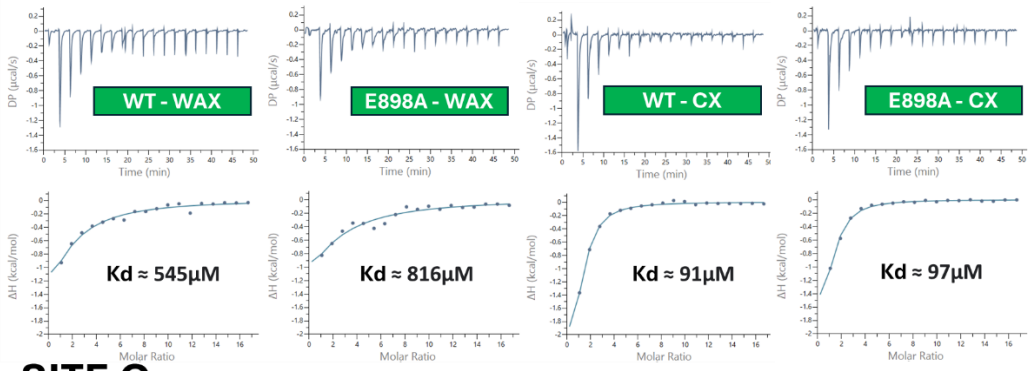
Table 9-3: Components used in *Bacteroides* Minimal Media

	Component	Amount
4 L 10X *Bacteroides* Salts – pH 7.2 filter sterilise		
	KH_2PO_4	544 g
	NaCl	35 g
	$(NH_4)_2SO_4$	45 g
1 L Balch's Vitamins – pH 7.0, filter sterilise, store in dark at 4 ºC		
	p-Aminobenzoic acid	5 mg
	Folic acid	2 mg
	Biotin	2 mg
	Nicotinic acid	5mg
	Calcium pantothenate	5 mg
	Riboflavin	5 mg
	Thiamine HCl	5 mg
	Pyridoxine HCl (vitamin B6)	10 mg
	Cyanocobalamin (vitamin B12)	0.1mg
	Thioctic acid	5 mg
250 mL Amino Acid Solution – filter sterilise, store at room temperature		
	Alanine	62.5 mg
	Arginine	62.5 mg
	Asparagine	62.5 mg
	Aspartic Acid	62.5 mg
	Cysteine	62.5 mg
	Glutamic Acid	62.5 mg
	Glutamine	62.5 mg
	Glycine	62.5 mg
	Histidine	62.5 mg
	Isoleucine	62.5 mg
	Leucine	62.5 mg
	Lysine	62.5 mg
	Methionine	62.5 mg
	Phenylalanine	62.5 mg
	Proline	62.5 mg
	Serine	62.5 mg
	Threonine	62.5 mg
	Tryptophan	62.5 mg
	Tyrosine	62.5 mg
	Valine	62.5 mg
1 L Purine/ Pyrimidine Solution – pH 7.0, filter sterilise, store at room temperature		
	Adenine	200 mg
	Guanine	200 mg
	Thymine	200 mg
	Cytosine	200 mg
	Uracil	200 mg
1 L Trace Mineral Supplement – pH 7.0, filter sterilise, store at room temperature		
	EDTA	500 mg
	$MgSO_4*7H_2O$	3 g
	$MnSO_4*H_2O$	500 mg
	NaCl	1 g
	$FeSO_4*7H_2O$	100 mg
	$CaCl_2$	100 mg
	$ZnSO_4*7H_2O$	100 mg
	$CuSO_4*5H_2O$	10 mg
	H_3BO_3	10 mg
	$Na_2MoO_4*2H_2O$	10 mg
	$NiCl_2*6H_2O$	20 mg

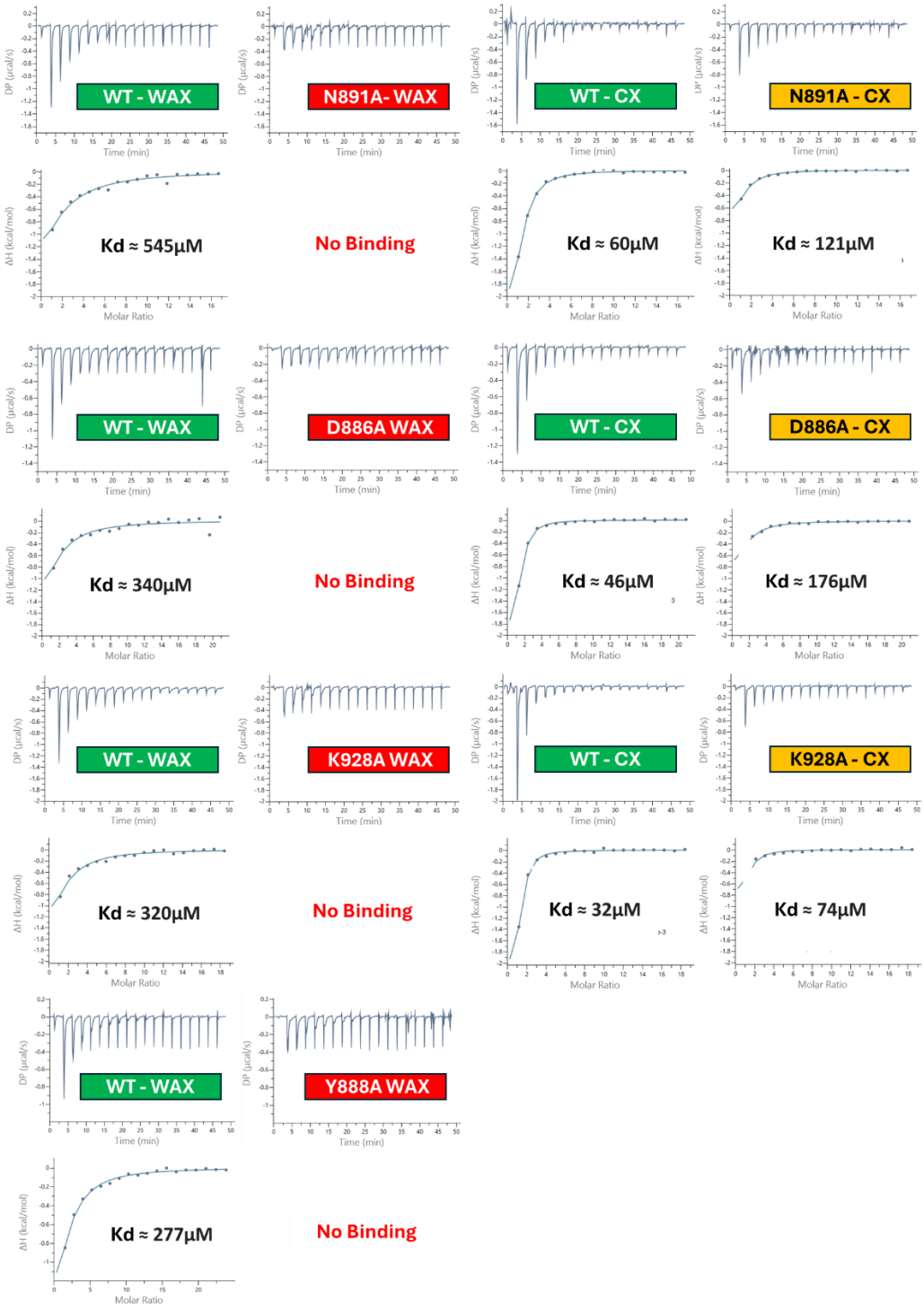
SITE A


WT - WAX
K828A - WAX
WT - CX
K828A - CX
Kd ≈ 303µM
Kd ≈ 290µM
Kd ≈ 63µM
Kd ≈ 57µM
WT - WAX
T827A - WAX
WT - CX
T827A - CX
Kd ≈ 303µM
Kd ≈ 499µM
Kd ≈ 63µM
Kd ≈ 17µM
WT - WAX
E836A - WAX
WT - CX
E836A - CX
Kd ≈ 303µM
Kd ≈ 956µM
Kd ≈ 63µM
Kd ≈ 152µM
WT - WAX
D936A - WAX
WT - CX
D936A - CX
Kd ≈ 303µM
Kd ≈ 1300µM
Kd ≈ 63µM
Kd ≈ 490µM
WT - WAX
N848A - WAX
WT - CX
N848A - CX
Kd ≈ 396µM
Kd ≈ 1400µM
Kd ≈ 40µM
Kd ≈ 170µM
Time (min)
Molar Ratio
DP (µcal/s)
ΔH (kcal/mol)

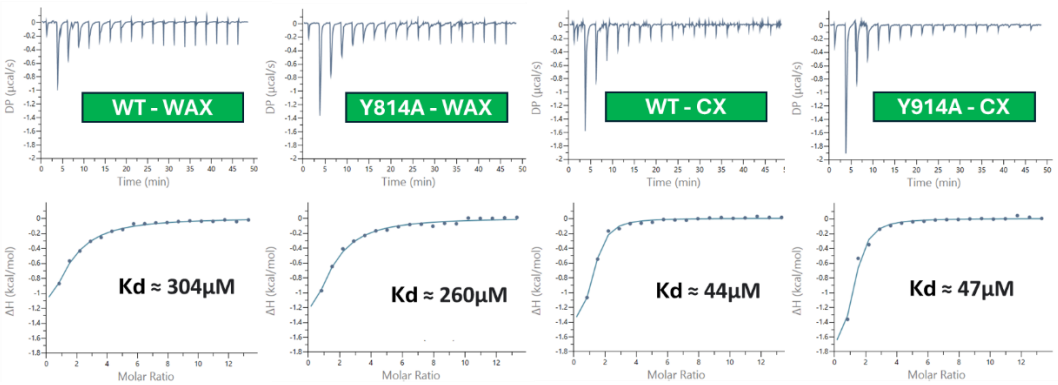
SITE B


WT - WAX
E898A - WAX
WT - CX
E898A - CX
Kd ≈ 545µM
Kd ≈ 816µM
Kd ≈ 91µM
Kd ≈ 97µM

SITE C


WT - WAX
N891A- WAX
WT - CX
N891A - CX
Kd ≈ 545µM
No Binding
Kd ≈ 60µM
Kd ≈ 121µM
WT - WAX
D886A WAX
WT - CX
D886A - CX
Kd ≈ 340µM
No Binding
Kd ≈ 46µM
Kd ≈ 176µM
WT - WAX
K928A WAX
WT - CX
K928A - CX
Kd ≈ 320µM
No Binding
Kd ≈ 32µM
Kd ≈ 74µM
WT - WAX
Y888A WAX
Kd ≈ 277µM
No Binding

SITE D



Appendix Figure 9-1: ITC data for mutants of Bo98-CBM35. Impact of 11 single point mutants from 4 potential binding sites on binding activity was studied. The top frames show the raw heats, and the bottom show the integrated peak areas fitted to a one site model using MicroCal PEAQ-ITC Analysis Software v1.41