Harnessing Natural Diversity for the Discovery of Glycoside Hydrolases and Design of New Glycosynthases

by

Zachary Armstrong

B.Sc., The University of British Columbia, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Harnessing Natural Diversity for the Discovery of Glycoside Hydrolases and Design of New Glycosynthases.

Submitted By Zachary Armstrong in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Genome Science and Technology

Examining Committee:

Professor Stephen G. Withers
Co-supervisor
Professor Steven J. Hallam
Co-supervisor
Professor Lindsay D. Eltis
Supervisory Committee Member
Professor Harry Brummer
University Examiner
Professor Martin Tanner
University Examiner

Abstract

Plant biomass offers a sustainable source for energy and materials and an alternative to fossil fuels. However, the industrial scale production or biorefining of fermentable sugars from plant biomass is currently limited by the lack of cost effective and efficient biocatalysts. Microbes, the earth's master chemists – employing biocatalytic solutions to harvest energy, and transform this energy into useful molecules – offer a potential solution to this problem. However, a majority of microbes remain uncultured, limiting our access to the genetic potential encoded within their genomes. This has spurred the development of culture independent methods, termed metagenomics.

In this thesis I harnessed high-throughput functional metagenomic screening to discover biomass deconstructing biocatalysts from uncultured microbial communities. Towards this goal, twenty-two clone libraries containing DNA sourced from diverse microbial communities inhabiting terrestrial and aquatic ecosystems were screened with 4-methylumbelliferyl cellobioside to detect glycoside hydrolase activity. This revealed 178 active clones containing glycoside hydrolases, often in gene clusters. This set of active clones was consolidated and further characterized through sequencing and rapid, plate-based, biochemical assays. Additionally, libraries sourced from beaver fecal and gut microbiomes were screened with four fluorogenic probes (6-chloro-4-methylumbelliferyl derivatives of cellobiose, xylobiose, xylose and mannose) for glycoside hydrolase activity. This revealed a total of 247 active fosmid-harbouring clones, that encoded many polysaccharide-degrading genes and gene cassettes. Specific candidate genes from the fecal library were sub-cloned, and the resulting purified enzymes were shown to be involved in synergistic degradation of arabinoxylan oligomers. The clone libraries that were generated through functional metagenomic screening were then employed to reveal the promiscuity of glycoside hydrolases towards unnatural azido- and aminoglycosides. Promiscuous enzymes identified from metagenomic and synthetic clone libraries were then used as a starting point for the generation of new glycosynthases capable of incorporating modified glucosides

and galactosides. The resulting set of eight new glycosynthases are capable of synthesizing di- and trisaccharides, glycolipids and inhibitors such as 2,4-dinitrophenyl 4'-amino-2,4'-dideoxy-2-fluoro-cellobioside. Taken together this work has exploited the power of functional metagenomics to reveal new modes of biocatalysis and develop new synthetic tools.

Lay Summary

Microbes are ubiquitous; they are in soil, air, water and inside our bodies. They also make enzymes to promote chemical transformations in our environment, including plant matter degradation. Some microbes are difficult to study as they can't be grown in the laboratory. For these microbes we use a collection of growth free methods, termed metagenomics. This thesis investigates plant-degrading genes – the DNA that codes for enzymes – present in soil, ocean water, bioreactors, coal beds, and beaver digestive tracts using metagenomics. Many of the uncovered genes had not been seen before. I also investigated how some of these enzymes work together to degrade specific sugars present in plants. Additionally, I made some of these enzymes capable of creating unnatural molecules that would be difficult to create otherwise. This work used metagenomics to discover catalysts, including those that break down plants, and create catalysts to synthesize unnatural molecules.

Preface

A number of sections of this work are partly or wholly published in press. Much of this research was conducted as a collaborative effort and contributions to each section are detailed below.

- Portions of Chapter 1 and Chapter 5 drew references and ideas from previous publications but contain wording original to this thesis. These publications, for which I am an author, follow below:

  Zachary Armstrong, Keith Mewis, Cameron Strachan, and Steven J. Hallam. "Biocatalysts for biomass deconstruction from environmental genomics." *Current Opinion in Chemical Biology* 29: 18-25. (2015)

  Zachary Armstrong, Stephen G Withers. "Synthesis of glycans and glycopolymers through engineered enzymes." *Biopolymers* 99(10): 666-674

  Zachary Armstrong, Peter Rahfeld, Stephen G Withers. "Discovery of New Glycosidases From Metagenomic Libraries." *Methods in enzymology* 597, 3-23

- Chapter 2: The functional screening method was developed and applied for study of the anaerobic bioreactor and forest soils by Dr. Keith Mewis. Other screening efforts were undertaken by Sam Kheirandish under supervision of Dr. Keith Mewis and Zachary Armstrong. Fosmid libraries for environments screened were created by Dr. Marcus Taupp, Dr. Sangwon Lee, Payal Sipahimalani, and Melanie Scofield.

- Chapter 3: Sampling of beaver feces was performed by Zachary Armstrong and Dr. Kevin Mehr, and DNA isolation and purification was performed by Zachary Armstrong. The fosmid library was created by Zachary Armstrong with assistance from Melanie Scofield. Screening and hit validation was performed by Zachary Armstrong. Cell lysate initial rate characteriza-

tion was preformed by Zachary Armstrong and Dr. Feng Liu. Beaver intestinal samples were collected by Dr. Keith Mewis and Zachary Armstrong, and DNA isolation and purification was performed by Zachary Armstrong. DNA and fosmid sequencing was performed by Dr. Keith Mewis at the UBC Pharmaceutical Sciences Sequencing Center (PSSC) with help from Dr. Sunita Sinha and Jennifer Chiang. Molecular cloning and protein expression, purification and characterization was performed by Zachary Armstrong.

- Chapter 4: All molecular cloning, protein purification and characterization was performed by Zachary Armstrong. Both Zachary Armstrong and Dr. Feng Liu were responsible for large-scale purifications. All NMR assignment was performed by Dr. Feng Liu.

The UBC Office of Research Ethics was consulted related to work with dissected beavers in Chapter 3, but no ethical applications or approval was required.

Throughout this work, the term "we" refers to Zachary Armstrong, unless otherwise stated.

Table of Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

Abbreviations

$\Delta\Delta G^{o\ddagger}$	Change in the Gibbs energy of activation
AG	Arabinogalactan
Ara	Arabinan
BLAST	Basic Local Alignment Search Tool
bp	base pair
CAZy	Carbohydrate Active enZyme
CE	Carbohydrate Esterase
Cel	Cellulose
DNA	Deoxyribonucleic acid
GH	Glycoside Hydrolase
GM	Glucomannan
GX	Glucuronoxylan
HG	Homogalacturonan
Kbp	thousand base pairs
LPMO	Lytic Polysaccharide Mono-Oxygenase
Mbp	Million base pairs
ORF	Open reading frame
OTU	Operational Taxonomic Unit
PL	Polysaccharide Lyase
PUL	Polysaccharide Utilization Loci
RG	Rhamnogalacturonan
RNA	Ribonucleic acid
SSU rRNA	Small Subunit Ribosomal Ribonucleic Acid
T_m	Denaturation midpoint
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
XG	Xyloglucan
XGUL	Xyloglucan Utilization Loci

Amino acids

Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic Acid
Cys	C	Cysteine
Glu	E	Glutamic Acid
Gln	Q	Glutamine
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine

Substrates

αF-3-N_3-Glc	3-azido-3-deoxy-α-D-glucopyranosyl fluoride
αF-3-NH_2-Glc	3-amino-3-deoxy-α-D-glucopyranosyl fluoride
αF-4-N_3-Glc	4-azido-4-deoxy-α-D-glucopyranosyl fluoride
αF-4-NH_2-Glc	4-amino-4-deoxy-α-D-glucopyranosyl fluoride
αF-6-N_3-Gal	6-azido-6-deoxy-α-D-galactopyranosyl fluoride
αF-6-N_3-Glc	6-azido-6-deoxy-α-D-glucopyranosyl fluoride
αF-6-NH_2-Glc	6-amino-6-deoxy-α-D-glucopyranosyl fluoride
αF-Gal	α-D-galactopyranosyl fluoride
αF-Glc	α-D-glucopyranosyl fluoride
CMU	6-chloro-4-methylumbelliferyl
CMU-C	6-chloro-4-methylumbelliferyl cellobioside
CMU-X	6-chloro-4-methylumbelliferyl β-D-xylopyranoside
CMU-X2	6-chloro-4-methylumbelliferyl xylobioside
DNP	2,4-dinitrophenol
DNP 2-F-Gal	2,4-dinitrophenyl 2-deoxy-2-β-D-fluoro-galactoside
DNP 2-F-Glc	2,4-dinitrophenyl 2-deoxy-2-β-D-fluoro-glucoside
DNP-C	2,4-dinitrophenyl cellobioside
Gal	Galactose
Glc	Glucose
MU	4-methylumbelliferone
MU-3-N_3-Glc	4-methylumbelliferyl 3-azido-3-deoxy-β-D-glucopyranoside
MU-4-N_3-Glc	4-methylumbelliferyl 4-azido-4-deoxy-β-D-glucopyranoside
MU-6-N_3-Gal	4-methylumbelliferyl 6-azido-6-deoxy-β-D-galactopyranoside
MU-6-N_3-Glc	4-methylumbelliferyl 6-azido-6-deoxy-β-D-glucopyranoside
MU-3-NH_2-Glc	4-methylumbelliferyl 3-amino-3-deoxy-β-D-glucopyranoside
MU-4-NH_2-Glc	4-methylumbelliferyl 4-amino-4-deoxy-β-D-glucopyranoside
MU-6-NH_2-Glc	4-methylumbelliferyl 6-amino-6-deoxy-β-D-glucopyranoside
MU-3-O-Me-Gal	3-methoxy-β-D-galactopyranoside
MU-3-O-Me-Glc	3-methoxy-β-D-glucopyranoside
MU-Ara	4-methylumbelliferyl α-L-arabinofuranoside
MU-C	4-methylumbelliferyl cellobioside
MU-Gal	4-methylumbelliferyl β-D-galactoside
MU-Glc	4-methylumbelliferyl β-D-glucopyranoside
MU-Lac	4-methylumbelliferyl Lactoside
MU-Man	4-methylumbelliferyl mannoside
MU-X	4-methylumbelliferyl β-D-xylopyranoside
MU-X2	4-methylumbelliferyl xylobioside
pNP	*p*-nitrophenol
pNP 6-PO_4-Glc	*p*-nitrophenyl 6-phospho-β-D-glucopyranoside
pNP-Ara	*p*-nitrophenyl α-L-arabinofuranoside
pNP-Gal	*p*-nitrophenyl β-D-galactopyranoside
pNP-Glc	*p*-nitrophenyl β-D-glucopyranoside
pNP-X	*p*-nitrophenyl β-D-xylopyranoside
Xyl	Xylose

Acknowledgements

Firstly, I would like to thank my two Ph.D. supervisors Dr. Steven Hallam and Dr. Stephen Withers for their years of support, mentorship, enthusiasm and vision. I would also like to thank all the past and present members of both the Hallam and Withers labs for their support and friendship. In particular I would like to thank my partner on many of these projects: Dr. Keith Mewis, for his drive and enthusiasm. I would also like to thank Sam Keirandish for his robotic expertise; Dr. Feng Liu to his determined NMR assignments and robot wrangling; lab managers Emily Kwan, Diane Fairley, Melanie Scofield, and Jade Schiller without whom nothing would ever get done; Spence Macdonald, Dr. Peter Rahfield, Dr. David Kwan and Dr. Ethan Goddard-Bodger for helpful conversations, perspective and inspiration; Connor Morgan-Lang, Dr. Aria Hahn, and Dr. Niels Hanson for bioinformatic support; and Dr. Harry Brumer for the use of his HPAEC.

Finally, and most gratefully, I would like to thank my wife Sarah for her unending patience and support.

Chapter 1

Introduction

Nature is replete with a diversity of biocatalysts possessing the potential to solve current industrial and medical challenges. To convert this potential into tangible solutions new biocatalysts must be discovered, characterized and adapted to the particular application. Discovery of new biocatalysts has historically relied on the screening of cultured organisms for the activity of interest. This, however, neglects the majority of microbes which belong to phyla lacking a cultured representative [133, 256]. To pass beyond the limitations of culture dependence, metagenomic techniques have been developed which allow us to investigate the genomic information of an environmental sample without prior culturing of the present microorganisms [112, 255, 256]. Furthermore, functional metagenomic screening allows us to tap into the catalytic power of these microbes without prior sequence annotation, enabling the discovery of catalysts which may have little or no sequence similarity to previously known catalysts.

Ready access to sustainable energy and materials is one challenge which may be solved through biocatalytic means. Plant biomass is a renewable resource that can be converted into energy and materials as an alternative to fossil fuel [47, 100]. The structure of plant biomass, has however, evolved to be highly recalcitrant, thus complicating the realization of its potential value [122]. To harvest the fermentable sugars and aromatics provided in plant biomass, mechanical, chemical and biological processes have been developed [199, 232]. However, a major limitation for the industrial deconstruction of plant biomass polymers continues to be a lack of cost-effective and efficient biocatalysts. For over 3.5 billion years, cooperative microbial communities have been driving energy and material transformations that create and sustain planetary living conditions. As a result, although the vast majority of microbes in nature remain uncultured, they represent an almost unbounded reservoir of genetic information and metabolic potential [256, 293]. High-

throughput functional metagenomic screening offers a method to tap into this reservoir and identify new catalysts for the deconstruction of plant biomass.

Within this introductory chapter I review existing literature and motivate the creation of active clone libraries through functional metagenomic screening and the development of biocatalysts from these libraries. Firstly, the molecular structure of plant biomass (with particular emphasis on the polysaccharide component), and its variation within primary and secondary cell walls is reviewed. Next, carbohydrate active enzymes are reviewed, with particular emphasis placed on classes of enzymes and enzyme systems that degrade plant biomass. This chapter then reviews a class of engineered enzymes (glycosynthases) which can be generated from plant biomass-degrading enzymes and used for synthesis. Functional metagenomic methods for the discovery of biomass-degrading enzymes are then reviewed. Finally, the structure of this dissertation is detailed.

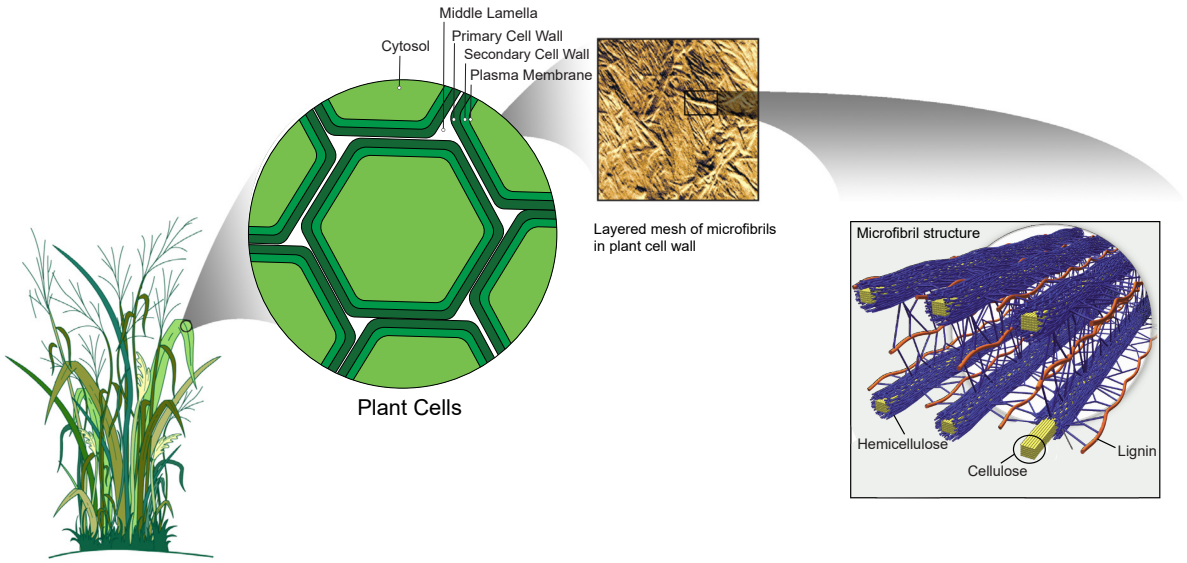

Figure 1.1: *Polymer Constituents of Lignocellulose.* Cellulose, hemicellulose and lignin form structures which are organized into macrofibrils that mediate the structural stability of plant cell walls. Pectic fraction is not shown. Adapted from Rubin [259].

1.1 Plant Biomass

One of the hallmarks of plant cells is the presence of a cell wall composed of the polysaccharides cellulose, hemicelluloses and pectin and the polyaromatic lignin. This cell wall can be divided into three layers: the primary, and secondary cell walls and the middle lamella, Figure 1.1. Primary cell walls are synthesized during growth and typically are relatively thin, flexible, highly hydrated structures [66]. Secondary cell walls provide strength and rigidity in plant tissues that have ceased growing [66]. The middle lamella is a thin pectin-rich region between adjacent primary cell walls [334]. The concentrations and structures of cell wall polysaccharides vary between the primary and secondary cell walls and with the plant taxa.

Marine algae have similar cell walls to land plants, containing crystalline cellulose, hemicellulose and matrix polysaccharides. However, algae contain several hemicelluloses and matrix polysaccharides that are not found within the land plants, including the sulfated glucan and glucuronan hemicelluloses found in red and brown algae [**?**]. Futhermore algae do not contain pectins but instead the green algae contain ulvans, red algae contain fucans and brown algae contain agars, carageenans and prophyrans [**?**]. Although algal polysaccharide structures are undoubtedly important to understanding biomass degradation, the following description of plant polysaccharides will be confined to those found within terrestrial plants.

1.1.1 Structure of Polysaccharides

The majority of plant cell wall polymers are polysaccharides. The functional role of these polysaccharides is dictated by the monosaccharides present (see Figure 1.3) and their linkages, which may be highly repetitive or extremely diverse. In order to determine the structure of a polysaccharide the following must be determined:

- **Which monosaccharides are present.** The most common monosaccharides contain either five or six carbons and are known as pentoses and hexoses respectively. These monosaccharides contain either an aldehyde or ketone at one end and typically hydroxyls at each of the other carbons. For a hexose this means that in the linear form there are four stereocentres (carbons 2-5 in Figure 1.2). Instead of referring to monosaccharides by their absolute R- and S-

configuration, as this would be cumbersome, each of the possible monosaccharides has a trivial name.

- **Whether the sugar present is the D- or L- isomer.** These isomers are mirror images of each other and can be determined by the configuration of the carbon furthest from the aldehyde or ketone functionality. This is carbon 5 for glucose, shown in Figure 1.2. This naming is based on the analogy to D- and L-glyceraldehyde.

- **Whether the sugar is in the pyranose or furanose form.** Free sugars often exist as a mixture of the linear, 5-membered ring (furanose) and 6-membered ring (pyranose) structures. When present as acetals or ketals in polysaccharides the monomers are no longer able to interconvert between these isomers, unless they are at the reducing end of a polysaccharide. See Figure 1.2 for cyclic forms of glucose.

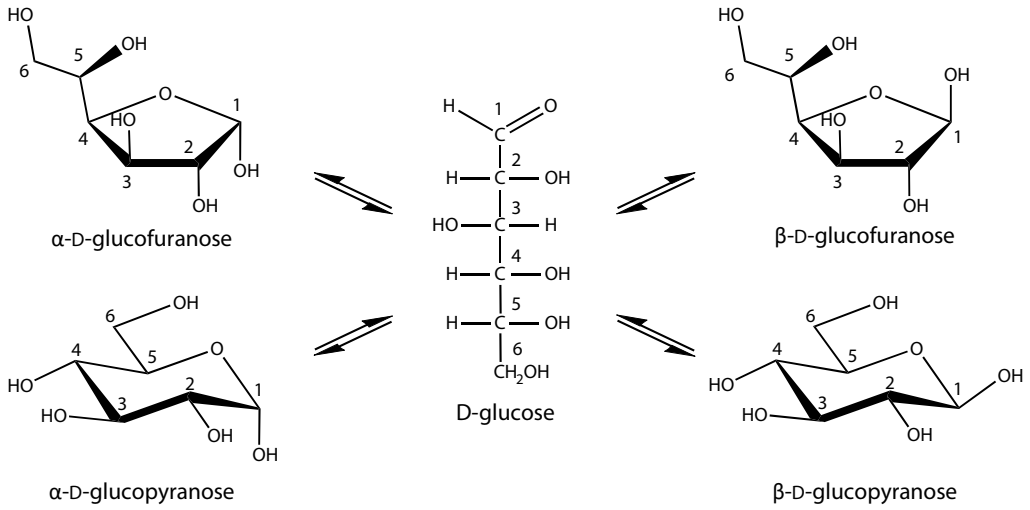

Figure 1.2: *Forms of D-Glucose.* D-Glucose is shown in its open chain form as a Fischer projection The cyclic forms of glucose are shown as Haworth projections for the furanose forms and in the chair conformation for the pyranose forms. Carbons are numbered.

- **Whether the anomeric position (C_1 in the ring form, in most cases) is in the α- or β- configuration.** The anomeric configuration of a sugar is determined by reference to the carbon that determines the D- or L- configuration. In a Fischer projection, if the substituent

off the anomeric centre is on the same side as the oxygen of the configurational (D- or L-) carbon, then it is the α-anomer. If it is directed in the opposite direction it is the β-anomer. See Figure 1.2 for anomeric configurations of glucose.

- **Which hydroxyls form the linkage between two monosaccharides.** Typically, glycosidic linkages are formed between the anomeric center of one monosaccharide and a non-anomeric hydroxyl in another monosaccharide. For two β-D-glucopyranose residues this would mean four separate linkages could be formed (1,2-, 1,3- 1,4- or 1,6-linkages).
- **The position and presence of any non-carbohydrate modifications.** Common modifications include acetylations and methylations of hydroxyls.

1.1.2 Cellulose

Approximately 35 to 50 % of dry plant matter is composed of cellulose [184], a polymer of repeating 1,4-linked β-D-glucopyranose subunits, making it the most abundant terrestrial polymer. Within plants, macromolecular complexes synthesize several cellulose strands simultaneously [65]. Hydrogen bonding and hydrophobic interactions between these strands cause them to form an insoluble crystalline cellulose microfibril [279]. These microfibrils may be 3-5 nm in diameter, several micrometers in length, and contain several hundred glucose molecules [65]. The insoluble nature of cellulose confers structural stability and causes practical problems for microbial degradation [121]. Many taxa, other than the land plants, also synthesize cellulose, including: green algae (Chlorophyta and Charophyta)[79], red algae (Rhodophyta) [297], brown algae (Phaeophyceae) [185], Oomycetes [118], animals (Urochordates which are marine invertebrates) [212], Amoebozoa [85] and Cyanobacteria [221].

1.1.3 Hemicellulose

Hemicellulose is an overarching term used to describe the non-ionic polysaccharides, other than cellulose, which are present in the plant cell wall. This includes mixed-linkage glucans, xylans, xyloglucans, glucomannan and mannans. Hemicelluloses are thought to play roles as signalling molecules [335] and in strengthening cell walls through interactions with cellulose and lignin [264].

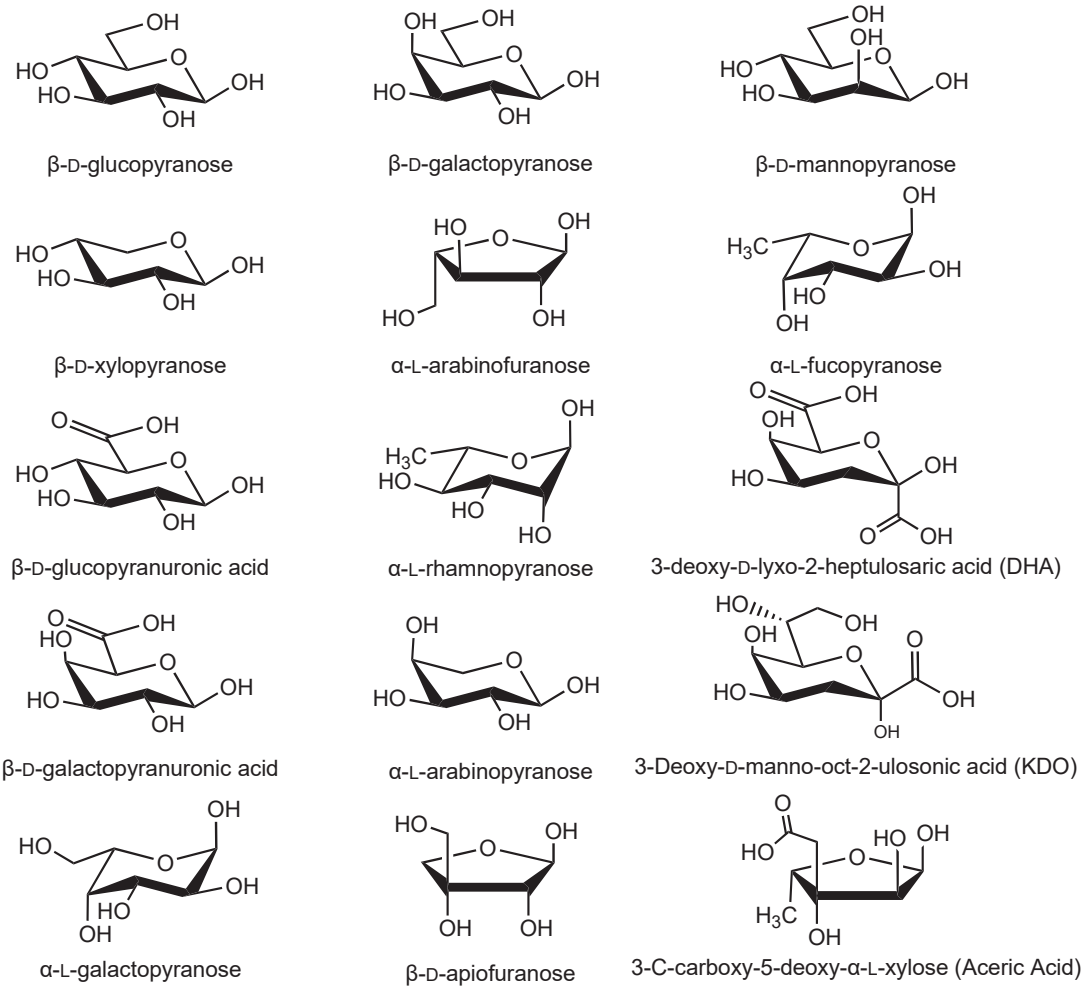

Figure 1.3: *Monosaccharides Present in Plant Biomass.* Cellulose, hemicelluloses and pectin are composed of the variety of monosaccharides shown.

The composition and abundance of each of these polysaccharides is variable between species, and often differs between primary and secondary cell walls within the same species [264].

Mixed linkage glucans

Like cellulose, mixed linkage glucans (MLGs) consist entirely of β-glucose residues, however unlike cellulose MLGs contain 1,3-linkages in addition to the 1,4-linkages seen in cellulose. Typically, the 1,3-linkages are located every 3 to 4 residues, linking cellotriosyl and cellotetraosyl subunits [141]. MLGs are considered to be limited to grasses (Poales) and one isolated land plant genus

(*Equisetum*) [96] and are thought to be more abundant in the primary cell wall than in the lignified secondary cell wall [305]. MLGs are also present in many food sources including oats and barley [195]

Xylans

Xylans are characterized by a backbone chain of β-1,4-linked D-xylose residues. This backbone is the site for several decorations, the most common of these being: acetylation at the 2 or 3 position, attachment of α-D-glucuronic acid or 4-*O*-methyl-α-glucuronic acid at the 2-position and attachment of α-L-arabinofuranose at the 2- or 3-position [264]. Additionally, a majority of xylans have a characteristic reducing end sequence consisting of (xylose-(1,3)-α-L-rhamnopyranose-(1,2)-α-D-galacturonate-(1,4)-D-xylopyranose) [142]. Further decorations on the α-1,2-linked glucuronic acid residues are present in the monocot orders Alismatales, Asparagales and the dicot *Eucalyptis grandis* [234]. Within these taxa α-L-arabinopyranose is attached via 1,2-linkage to glucuronic acid, which may or may not be methylated [234]. The xylans of *Eucalyptis* also contain β-galactose units 1,2-linked to glucuronic acid [234, 298]. Xylans present in grasses (family Poaceae) may also have ferulic acid esters attached to the C-5 hydroxyl [75] and further decorations on C-2 hydroxyl of the α-1,3-linked arabinofuranose substituents [234]. Corn bran xylan is one such polymer, having multiple separate tetrasaccharides containing D-galactose, L-galactose and ferulic acid and D-xylose branching from an arabinofuranose sidechain [6, 11].

The decorations observed on the xylan backbone change with taxonomy [42, 234]. Dicots, including hardwood trees, contain primarily glucuronoxylan and lack arabinofuranose decorations. On the other hand, monocots of the order Poales, which includes grasses, contain higher concentrations 2- and 3-linked arabinofuranose decorations in addition to α-glucuronyl sidechains [234]. Xylans from softwood (Gymnosperms) such as Douglas fir [*Pseudotsuga menziesii*], or Spruce [*Picea abies*] are also decorated with both glucuronic acid and 1,3-linked arabinofuranose, but lack 1,2-linked arabinofuranose [42, 86]. Grasses (Poaceae) have higher concentrations of xylans than do either dicots or gymnosperms.

Xyloglucan

Xyloglucans are mainly present in the primary cell walls of plants, and form strong interactions with cellulose microfibrils through hydrogen bonding [233]. The backbone of this polysaccharide is a chain of β-1,4-linked glucose residues. The most common decorations of this polymer are xylosyl residues attached via α-linkages at the 6-hydroxyl position of the glucose backbone. Both the backbone glucosyl and the xylosyl residues can be further substituted with D- and L-galactosyl, L-fucosyl, D-galacturonosyl, L-arabinopyranosyl, L-arabinofuranosyl and acetyl moieties at specific locations in specific linkages, resulting in the 24 unique structures identified to date [233].

As for xylans, the diversity of xyloglucan substitutions varies with taxonomy. The most common form of xyloglucan contains a repeating structural unit consisting of three xylose-decorated glucoses followed by one undecorated glucose. This sequence is often galactosylated and fucosylated, resulting in fucogalactoxyloglucan which is observed to be present in most tissues of most dicots [233]. The majority of xyloglucans in the primary cell walls of gymnosperms are also fucogalactoxyloglucans with similar structures to those of the xyloglucans in the primary walls of most dicots [130]. Monocots have diverse xyloglucan structures, with non-grass monocots having structures similar to dicots, whereas grasses have fewer decorations and reduced xyloglucan concentrations [129].

Mannan and Glucomannan

Mannans and glucomannans contain β-1,4-linked D-mannose residues in their structural backbone with glucomannans also containing backbone β-1,4-linked D-glucose residues. This backbone can be decorated with acetyl groups at the 2- and 3-positions or with α-linked galactosyl groups at the 6-positions, forming galactoglucomannans [207]. Acetylated galactoglucomannans are present in gymnosperms, such as conifers, as the main hemicellulose, although xylans are also present [44]. Dicots and grasses also contain glucomannans, though in smaller amounts [264]. Mannans are also highly abundant in seeds as a storage polymer [37].

1.1.4 Pectins

Pectins, used as the gelling agent used in the preparation of jams and jellies, form gel like structures in cell walls which help hold the layers of the cell wall together [139]. Pectins contain a wider variety of monomers and linkages than those seen in either cellulose or hemicellulose. The identifying feature of pectins is a backbone containing the charged sugar galacturonic acid. This monomer is the sole backbone sugar in homogalacturonan (HG), rhamnogalacturonan II (RG-II) and xylogalacturonan, where it is linked via α-1,4-bonds. Rhamnogalacturonan I (RG-I), on the other hand, contains alternating α-L-rhamnose and α-D-galacturonic acid monomers with galactose attached to the 2-position and rhamnose attached to the 4-position of galacturonic acid.

The type of branches and decorations linked to the backbone polymer vary with the type of pectin. In HG, decorations of the backbone consist of methylesterifications of the C-6 carboxylate and acetylations at the O-2 and O-3 positions [43]. Xylogalacturonan also contains β-linked xylose at the 3-hydroxyl and occasionally the 4-hydroxyl of the galacturonan backbone [205]. In RG-I, branching polysaccharides occur at the 4-position of the rhamnose backbone residues. Arabinans, galactans and arabinogalactans have all been observed as branches from RG-I. The arabinans branching from RG-I contain a polymer of α-L-arabinose residues with a 1,5-linkage, which may be further decorated at the 3-position with additional α-L-arabinose residues. The galactans branching from RG-I consist of β-1,4-linked D-galactose residues, which may contain β-D-galactose decorations at the 6-position The branching arabinogalactans contain a backbone of either β-1,4-linked D-galactose (type I), or β-1,3-linked D-galactose (Type II). These arabinogalactan backbones serve as further branching points that can contain a variety of decorations [43].

RG-II, is the most complex of the pectin polymers as it contains 12 different sugar monomers with 20 different linkages (see Figure 1.7 for structure). These include the following rare sugars: D-apiose, L-aceric acid, 2-O-methyl L-fucose, 2-O-methyl D-xylose, L-galactose, 2-keto-3-deoxy-D-lyxo-heptulosaric acid (DHA) and 2-keto-3-deoxy-D-manno-octulosonic acid (Kdo) [228]. Additionally, RG-II is bound to borate through D-apiose residues, causing cross-linkages between RG-II strands.

In terms of abundance, HG is the major component of pectins, constituting approximately 65 % of pectin [205]. The next most abundant pectin polymer, RG-I, represents 20-35 % while RG-II

represents approximately 10 % of the pectin present in primary cell walls [205]. Xylogalacturonan concentrations are typically low as these polymers are mainly found in reproductive cells [205]. Pectins are abundant in the growing primary cell walls and middle lamella, but are present at much lower levels in secondary cell walls. RG-II is also present in the primary cell wall, but not detected in the middle lamella [228]. RG-II presence also varies with taxonomy, making up between 1-4 % of primary cell walls of dicots and gymnosperms but less than 0.1 % in grasses [228].

1.1.5 Lignin

Lignin, which after cellulose is the most abundant terrestrial biopolymer, [31] can constitute up to 30 % of secondary cell walls [264]. It is an aromatic polymer created through radical oxidative coupling of monolignols. The three most common lignin monomers are: *p*-coumaryl alcohol, sinapyl alcohol and coniferyl alcohol and their relative abundance in lignin varies with species. Dicots contain lignin derived from sinapyl and coniferyl alcohol, while grasses incorporate higher amounts *p*-coumaryl alcohol and gymnosperms (such as conifers) lack sinapyl alcohol [31].

The radical, oxidative mechanism of lignin formation causes highly diverse structural linkages to be formed between monomers and between monomers and carbohydrates present in the cell wall. In fact, the linkages are so diverse it has been hypothesized that no two lignin molecules are identical [3]. Lignin stiffens the cell wall by cross-linking with the polysaccharide fraction and it provides a barrier between potential pathogens and the energy-rich polysaccharides [304].

Table 1.1: Plant Cell Wall Composition, Amount of Polysaccharide (% w/w)

	Dicot Walls		Grass Walls		Conifer Walls		
Polymer	1^o	2^o	1^o	2^o	1^o	2^o	Reference
Cellulose	15-30	45-50	20-30	35-45	20-30	40-50	[39, 78, 252, 309]
Hemicelluloses							
β-Glucans	-	-	2-15	Minor	-	-	[264]
Xyloglucan	20-25	Minor	2-5	Minor	10	-	[264]
Glucuronoxylan	-	20-30	-	-	-	-	[264]
Glucuronoarabinoxlyan	5	-	20-40	40-50	2	5-15	[264]
Glucomannan	3-5	2-5	2	0-5	-	-	[264]
Galactoglucomannan	-	0-3	-	-	Present	10-30	[264]
Pectins	20-35	Minor	5	Minor	20-35	Minor	[309, 334]

*Values vary between different species and tissue types.

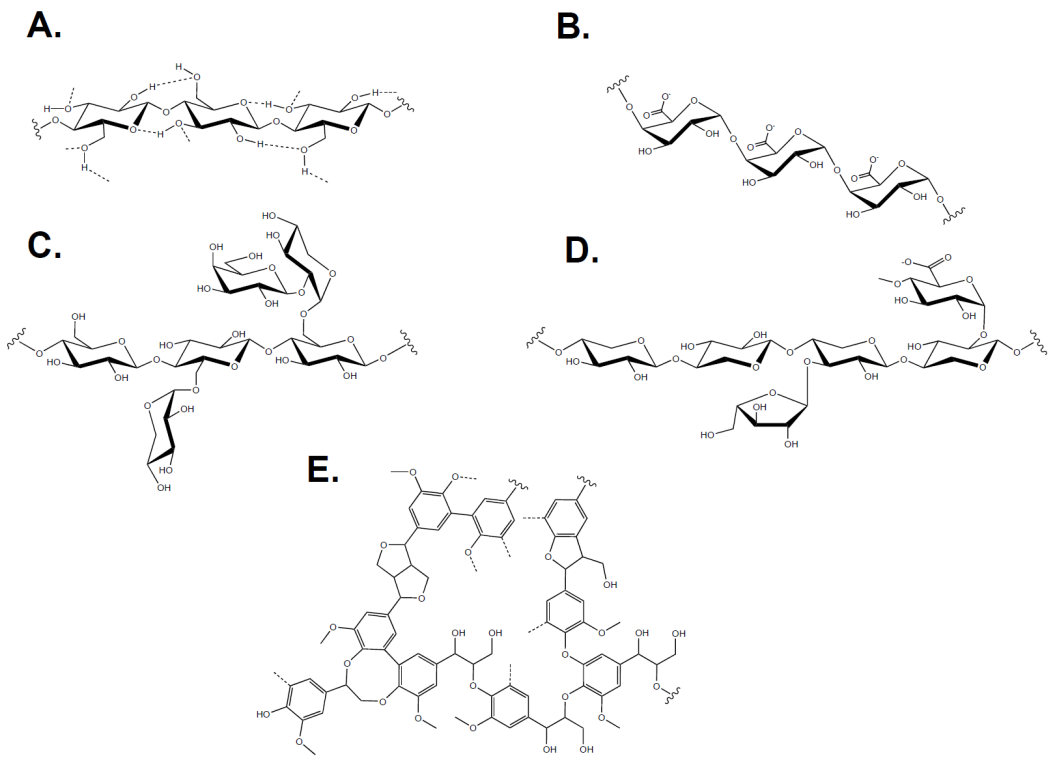

Figure 1.4: *Polymer Constituents of Lignocellulose.*Cellulose (A) solely contains glucose subunits linked through β-1,4 bonds. Polygalacturonan (B) is the main constituent of pectin and forms the backbone of rhamnogalacturonan II and homogalacturonan. Hemicellulose contains, among others, the polymers xyloglucan (C) and glucuronoarabinoxylan (D). Lignin (E) is a polyaromatic, with extremely heterogeneous structure.

1.2 Carbohydrate Active Enzymes

The enzymes that synthesize, modify and degrade carbohydrates are termed carbohydrate active enzymes. The Carbohydrate Active enZymes (CAZy) database (http://www.cazy.org/) has emerged as an integral clearing-house for functional annotation [181]. CAZy categorizes polysaccharide degradation genes, such as glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs) and more recently lytic polysaccharide mono-oxygenases (LPMOs) [172], into sequence-defined families.

1.2.1 Glycoside Hydrolases

Glycoside hydrolases (EC 3.2.1.x) are enzymes that catalyse the hydrolytic cleavage of either glycosidic bonds between saccharides or between a saccharide and a non-sugar molecule (aglycone). Classification within the EC framework fails to take into account enzyme mechanism and many enzymes with the same EC number have unrelated sequences. Sequence-based classification by the CAZy database has delineated over 150 glycoside hydrolase (GH) families [181]. Categorization into sequence-based families gives insight into the conserved mechanisms and active-site residues within families. Several families with multiple activities, including both GH43 and GH5, have also been further classified into subfamilies which provide finer details of the evolution and substrate specificity of specific families [15, 202].

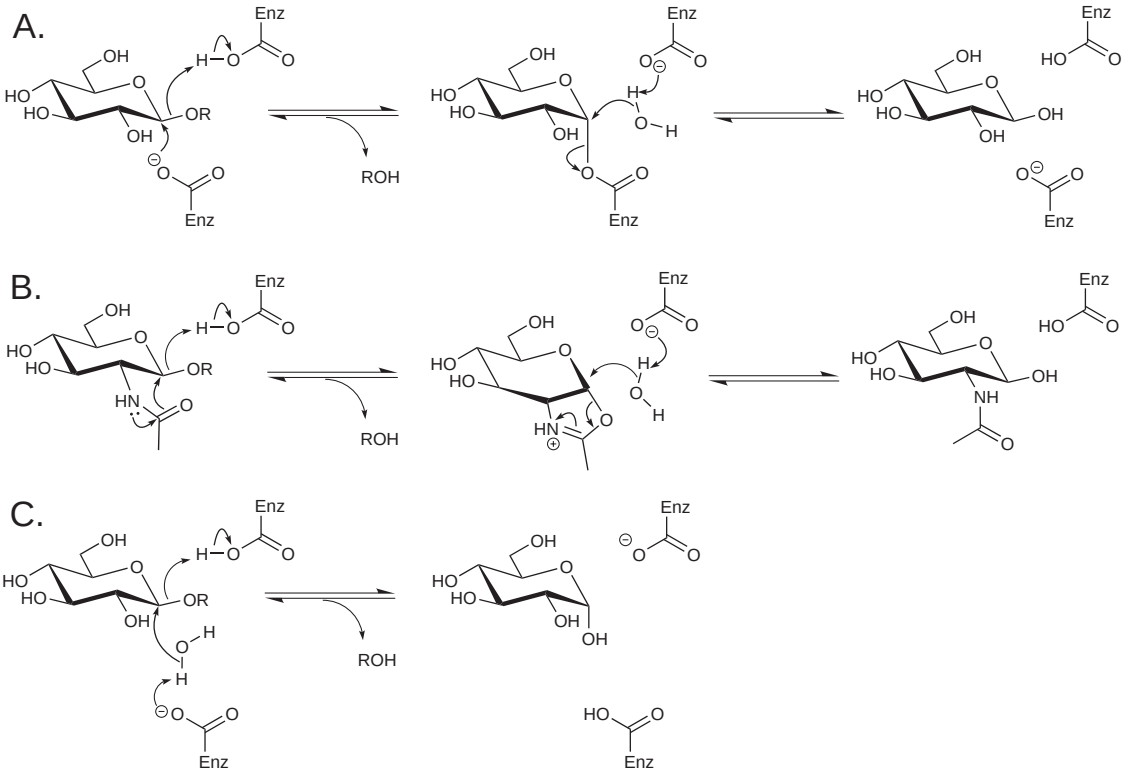

Figure 1.5: *Glycoside Hydrolase Mechanisms.* Catalytic mechanisms of a retaining glycoside hydrolase **(A)**, a substrate assisted, *N*-acetyl glucosaminidase **(B)**, and an inverting glycoside hydrolase **(C)**.

Glycoside hydrolysis can occur with either retention or inversion of stereochemistry at the

anomeric centre (see Figure 1.5). Retaining glycosidases progress through a double displacement mechanism involving a covalent intermediate. In the first step the nucleophilic active site residue (generally an aspartate or glutamate) attacks the anomeric center concomitant with protonation of the aglycone by the active site acid residue (often an aspartate or glutamate as well). This results in cleavage of the glycosidic bond and formation of the covalent glycosyl-enzyme intermediate. The anomeric center of this intermediate is subsequently attacked by a water molecule, with base catalytic assistance from the same active site acid/base residue. Retaining enzymes that hydrolyse 2-acetamido sugars can alternatively employ a substrate-assisted mechanism in which the active site nucleophile is absent [186]. In this case the acetamide oxygen attacks the anomeric centre producing an oxazolinium ion intermediate, which is in turn attacked by water to release the hydrolysis product with net retention of the anomeric stereochemistry (Figure 1.5 B). Inverting enzymes utilise an acid and a base residue to catalyse the direct attack of water at the anomeric centre, facilitating release of the aglycone with inversion of anomeric stereochemistry (Figure 1.5 C).

GHs can be further categorized as either exo- or endo-acting. Exo-acting enzymes cleave monosaccharides from either the reducing (the terminal anomeric center is not involved in bonding) or non-reducing (terminal anomeric center is involved in bonding) termini of a polysaccharide, releasing monosaccharides. Endo-acting enzymes, on the other hand, cleave glycosidic bonds within a polysaccharide releasing two polysaccharide fragments, and creating new termini which can be targeted by exo-cleaving enzymes. Within the context of plant biomass degradation both endo- and exo-acting enzymes are required for efficient degradation of plant polysaccharides. Furthermore, the complex nature of polysaccharides such as xyloglucan and rhamnogalacturonan II dictates requirement of many different GH families to catalyse the complete degradation of these polymers (see Figures 1.6 and 1.7).

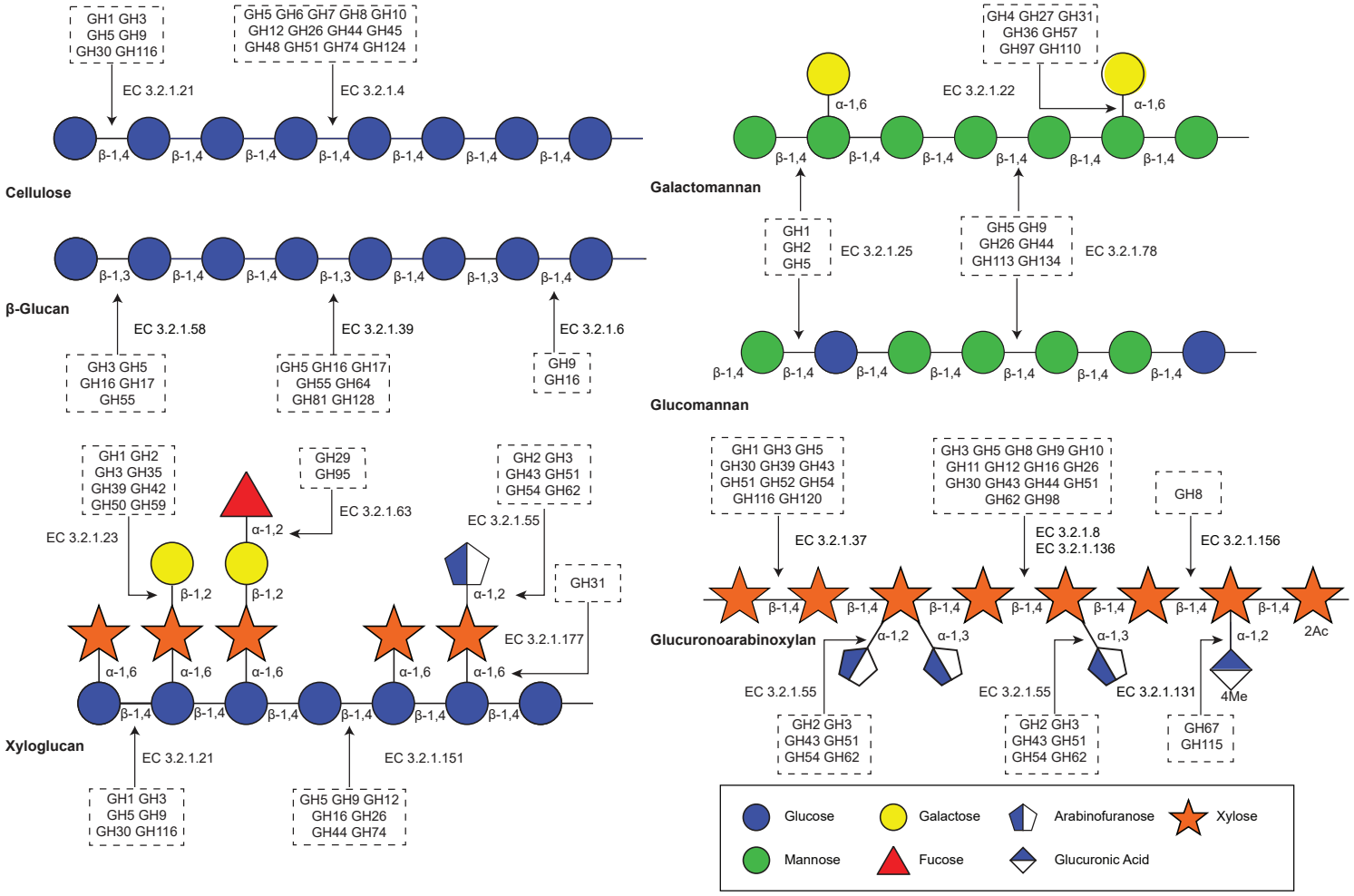

Figure 1.6: *Enzymatic Degradation of Plant Cellulose and Hemicelluloses.* Glycoside hydrolase and polysaccharide lysase families with the required activities to cleave plant polysaccharides. Monosaccharides are abbreviated as symbols and the linkages between them are labeled. Methylations and acetylations are abbreviated as Me and Ac respectively. GH and PL families with the required activity are given within the dashed boxes, and corresponding EC numbers are also indicated. CE families are excluded for simplicity.

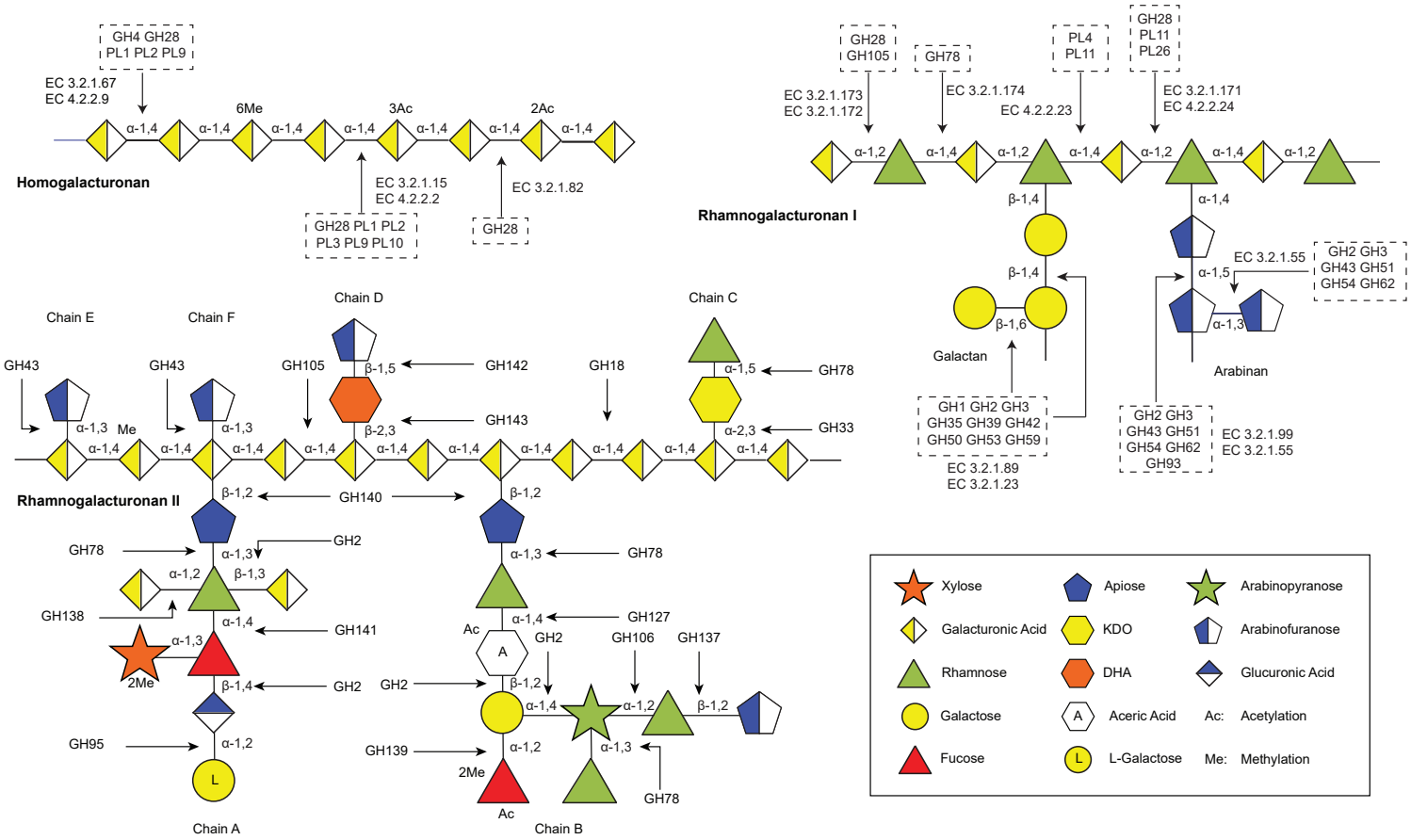

Figure 1.7: *Enzymatic Degradation of Plant Pectins.* Glycoside hydrolase and polysaccharide lysase families with the required activities to cleave plant polysaccharides. Monosaccharides are abbreviated as symbols and the linkages between them are labeled. Methylations and acetylations are abbreviated as Me and Ac respectively. GH and PL families with the required activity are given within the dashed boxes, and corresponding EC numbers are also indicated. The RG-II degradation genes shown has been limited to those families identified by Ndeh et. al. [215]. CE families are excluded for simplicity.

1.2.2 Polysaccharide Utilization Loci

The study of carbohydrate metabolism has resulted in foundational achievements in molecular biology. Study of the *lac* operon and the L-arabinose operon have revealed mechanisms of gene expression and provided powerful molecular tools [107, 137]. More recently, research has focused on co-localized gene clusters in Bacteroidetes genomes that target plant biomass. These carbohydrate targeting gene clusters have been termed Polysacccharide Utilization Loci (PULs) [28]. As Bacteroidetes are found in many diverse environments, including gut microbiomes [14, 208], both marine [89] and fresh water, [290] and soils [166], PULs play a significant role in planetary carbohydrate degradation.

The first PUL to be identified by Salyers et al. [289] was the starch utilization system (SUS), see Figure 1.8. This archetypical PUL contains the outer-membrane binding proteins which bind starch, and a surface-bound hydrolase that produces starch oligomers. These oligomers are then transported via a TonB-dependent transporter into the periplasmic space where they are further degraded by two additional hydrolases. The products of this saccharification can then enter the cell and central metabolism [90]. Not only are all these genes co-expressed, but they are also co-localized within the genome [90].

The hallmark of all PULs is the presence of a sequential pair of SusC-like and SusD-like proteins, encoding a TonB-dependent transporter and a surface binding protein. Otherwise, the variety of enzymes encoded within PULs varies in complexity with the polysaccharide being acted on. PULs have been shown to contain catalytic PLs, CEs, sulfatases and phosphorylases in addition to both endo- and exo-acting GHs [48, 187, 215]. Furthermore, the discovery of PULs that target the components of the plant cell wall (including mixed-linkage glucans [287], xyloglucan [165], xylan [257], galactomannan [16], RG-I [183] and RG-II [215]) has lead to the identification of several new GH families and the identification of new activities within previously known families

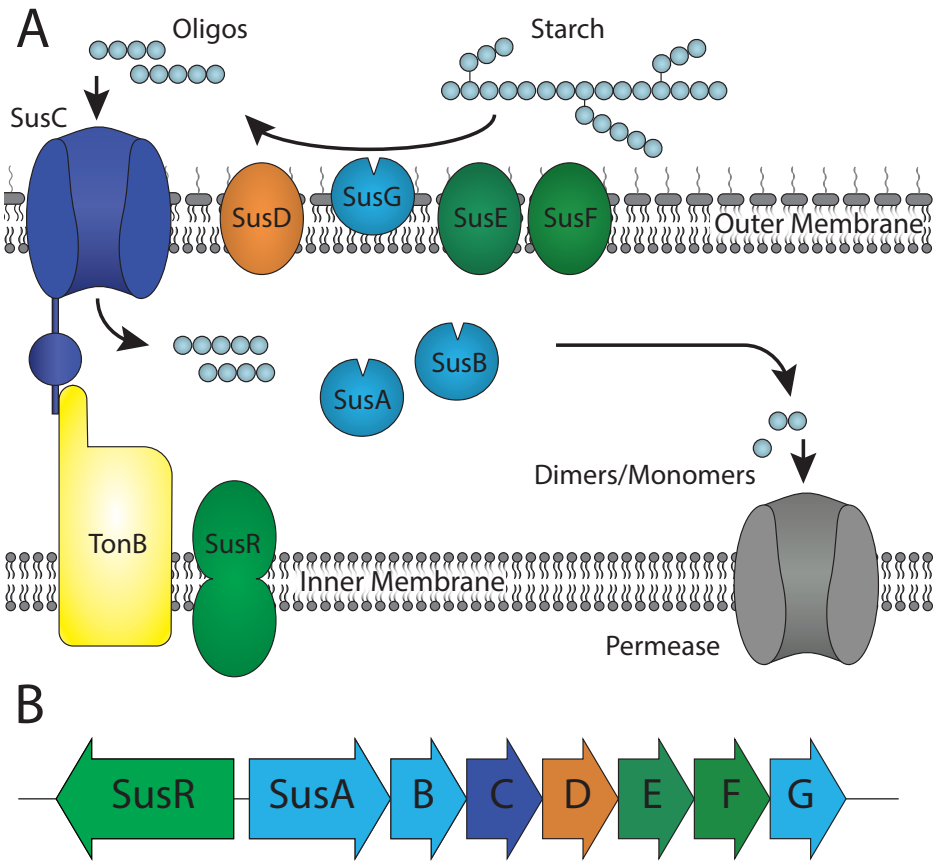

Figure 1.8: *Starch Utilization System (SUS) Operon in B. thetaiotaomicron.* **A:** Extracellular starch is bound by the outermembrane lipoproteins SusDEF and hydrolysed by SusG (GH13). These starch oligomers are then transported to the periplasm via the TonB-dependent transporter SusC. The starch oligos are further degraded to dimers and monomers by the hydrolases SusA (GH13) and SusB (GH97), which then enter the cell. SusR senses maltose and drives expression of *susABCDEFG*. **B:** Genomic organization of the SUS operon, genes are not shown to scale. Figure adapted from Koropatkin et al.[156]

1.2.3 Glycosynthases

Glycoside hydrolases are reversible, and therefore have the capability to be used in the synthesis of glycans. Reversal by altering the equilibrium position, however, is challenging and requires the use of very high sugar concentrations to counteract the presence of 55 M water [116]. Attempts to perform transglycosylations in non-aqueous solutions are not generally useful since the sugars themselves typically become insoluble, though in certain cases worthwhile products can be obtained

[164]. More fruitful has been the formation of products under kinetic control via transglycosylation. This necessarily uses a retaining glycosidase, typically with an activated donor sugar to form a high steady state concentration of glycosyl enzyme, allowing efficient transglycosylation [98]. However, the products formed from activated glycosides can subsequently be hydrolysed by the glycosidase, limiting yields.

Mutation of the active site nucleophile drastically decreases the hydrolytic activity of a retaining glycoside hydrolase [313]. This also prevents transglycosylation as the necessary covalent glycosyl-enzyme intermediate is no longer formed. If however, a donor substrate possessing an activated leaving group at the anomeric center with the opposite anomeric stereochemistry relative to the hydrolysis product (a mimic of the glycosyl-enzyme intermediate) is employed, transfer to a suitable acceptor can be catalysed without subsequent hydrolysis (Figure 1.9 A). Enzymes of this class have been termed glycosynthases. This method for glycan synthesis was first demonstrated 20 years ago with the E358A nucleophile variant of *Agrobacterium sp.* β-glucosidase (Abg) [188]. This enzyme was chosen to create a glycosynthase as the wild-type enzyme normally catalyses efficient transglycosylation [245] and the substitution of alanine for the glutamate nucleophile resulted in a stable enzyme with severely decreased hydrolysis rates [312]. The use of either α-galactosyl fluoride or α-glucosyl fluoride as donors and para-nitrophenyl glycoside acceptors with this enzyme enabled the production of several different glycans with yields of up to 92 % [188].

A similar method has also been developed for retaining glycoside hydrolases employing substrate assisted catalysis [303]. By utilising an activated oxazoline glycan as a donor, transglycosylation of 2-acetamido-glycans can be catalysed by a wild-type glycosidase [98]. However, the product may still be hydrolysed. Yamamoto and colleagues were able to circumvent this problem by introducing active site mutations which reduced hydrolysis rates without substantially compromising transglycosylation rates, thereby improving yields [303] (Figure 1.9 B).

Though most glycosynthases are derived from retaining glycoside hydrolases, inverting glycosidases have also been converted into glycosynthases by mutating the catalytic base and using an activated glycan with the same anomeric stereochemistry as the normal hydrolysis product. Efficient transglycosylation can be achieved without subsequent hydrolysis by reversal of the normal reaction since fluoride requires no acid activation for departure, yet the normal hydrolytic process

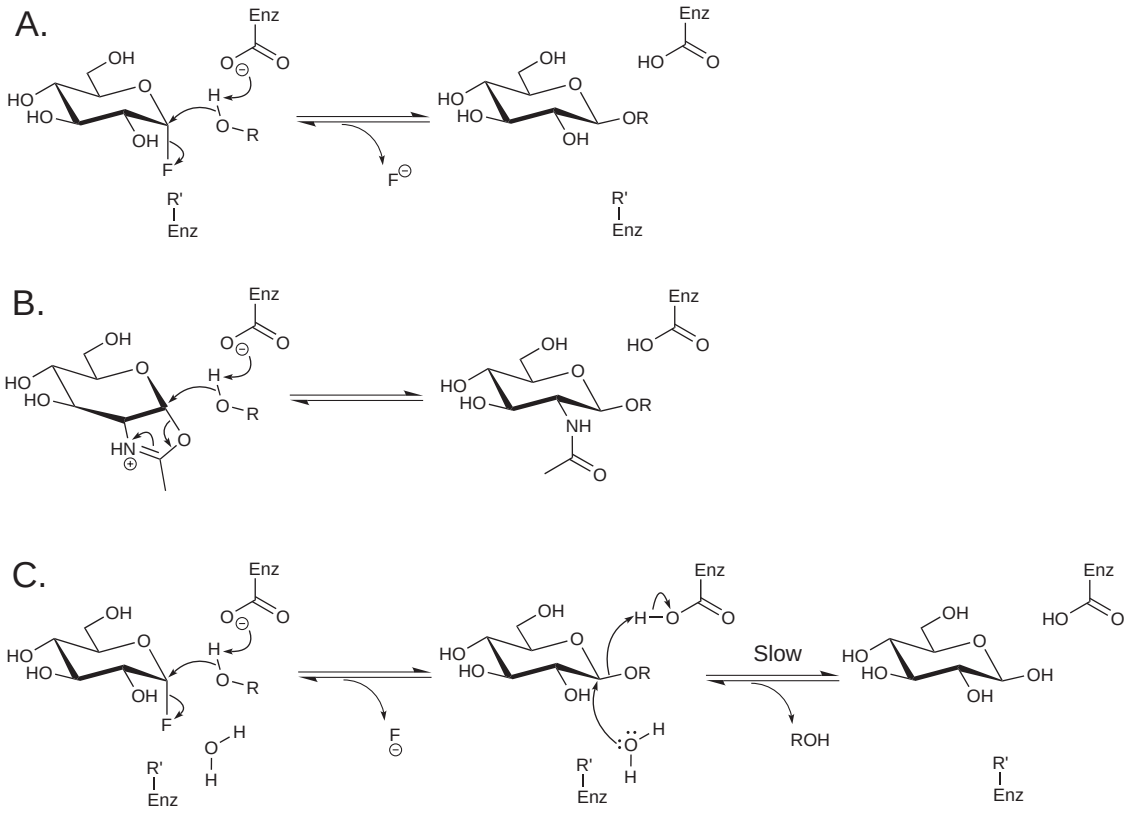

Figure 1.9: *Glycosynthase Mechanisms.* The mechanism of a glycosynthase developed from a retaining glycosidase **(A)**, a glycosynthase utilising an oxazoline donor sugar **(B)**, and an inverting glycosynthase **(C)**.

is substantially slowed (Figure 1.9 C). An example of this type of glycosynthase is an inverting glycosynthase from GH19 [227]. In that case, the S102A variant of the *Bryum coronatum* chitinase can catalyse the synthesis of chitooligosaccharides from α-chitobiosyl fluoride, which acts as both donor and acceptor molecule.

One could envisage using glycosynthases to synthesize almost any polysaccharide with defined regiospecificity and without need for chemical protection. However, for this vision to become a reality, the range of available glycosynthases must be expanded. Glycosynthases have thus far been developed from 17 GH families [60, 227], but this represents only a small fraction of the over 140 active glycoside hydrolase families currently known. Expansion of the range of GH families that have been converted to glycosynthases will enable the production of new glycan linkages. Also, the exploration of hydrolases within a family that may act on similar glycans but with different protein

or lipid specificity is a worthwhile goal. The creation of functional gene libraries should enable the rapid screening for enzymes with specific hydrolytic activities (including non-natural activites) which can then be converted into glycosynthases with a cognate synthase activity.

1.3 Metagenomics

A vast majority of the estimated 10^{30} prokaryotic cells [317] belong to species which have never been cultured in isolation. This confounds the central questions of microbial ecology, namely "who is there?" and "what are they doing?" [314]. To address these questions a number or techniques have been employed to analyse all the genetic material within an environment as a whole. To access the metagenome, a term first coined in 1998 [113], DNA is often isolated directly from the environment, thus bypassing the need for culturing.

Metagenomic research has taken advantage of massively paralleled, high-throughput DNA sequencing techniques to provide insight into environmental DNA. To analyse the functional role of these sequences and their corresponding genes within an environment, a functional prediction must be made. However, this is limited by the number of genes that have been functionally characterized and the reliability of prediction. Furthermore, these predictions are unable to assign new functionalities to novel genes; sequence annotation can only operate within the current paradigm of gene functions. For example, it has been estimated that only 6% of CAZy enzymes have been characterized and it has been estimated that the function of only 20% of the proteins in the sequence database can be predicted with confidence [105].

There is a clear need for the functional characterization of metagenomic DNA. This can be accomplished by functional metagenomic activity screens, coupled with high-throughput enzyme characterization. Functional screens have the ability to provide a direct link between metagenomes and their functional activities. They can also provide the ability to discover enzymes with activities that exist outside the current paradigms of gene annotation, which in turn, can better inform *in silico* approaches.

1.3.1 Functional Metagenomic Screens

Functional metagenomic screens involve the construction and screening of environmental DNA expression libraries. These libraries require a suitable vector for heterologous expression in a compatible host system such as *E. coli* (Figure 1.10). Identifying a suitable source of environmental DNA is a critical consideration when designing a screening strategy. Potential DNA sources include soil [73], water [135], feces [148] and bioreactors [201], all of which present different challenges in their processing. Soil and feces typically contain contaminants that interfere with downstream enzymatic processes, necessitating additional DNA purification steps. Water samples, on the other hand, may be too dilute, in terms of the number of cells per liter, and require the filtration of a large volume to obtain enough cells. Additionally, the choice of environment will likely dictate the viability and method of functional screening. If the targeted activity is known to be abundant in the environmental sample a small insert library is potentially sufficient. Conversely, if the suspected activities are likely to be rare, a small insert library will not be sufficient and a large insert, or fosmid, library will potentially be the better choice [293].

The choice of host strain is another factor that must be considered when designing a screen. Engineered *E. coli* strains are the most commonly used screening hosts for functional metagenomics, as they grow rapidly and are easy to transform with exogenous DNA. However, there are limitations when dealing with exogenous promoters, initiation factors, codon usage or protein folding. Gabor et al. [99] estimated that from a diverse subset of genomes the expression potential for an *E. coli* host system ranges widely, from only 7% to up to 73% of the genes. Additionally, it is important to select a host strain that lacks endogenous activity against the screening substrate.

Resulting libraries are screened for activity on agar [88] or in microtiter plates [200, 201], using a reporter e.g. substrate or transgene, or other form of phenotypic selection e.g. growth. Screening libraries sourced from a range of environmental conditions (e.g. pH, temperature, metal ion concentrations) enables recovery of active clones with alternative substrate specificities and tolerances [19, 24, 27, 147, 193, 223, 306, 318, 328, 332]. Similarly, libraries sourced from xylotrophic or wood-feeding organisms can provide insight into biomass deconstruction. Recently, Ruegg and colleagues screened an isolate fosmid library sourced from the lignocellulolytic bacterium *Enter-*

obacter lignolyticus to identify genes conferring IL tolerance under biorefining conditions in an *E. coli* host [260]. They recovered an active clone encoding a membrane transporter and transcriptional regulator enabling a 20% increase in biofuel production in the presence of 68 mM 1-ethyl-3-methylimidazolium chloride. Similarly, Bastien and colleagues screened fosmid libraries sourced from the termite *Pseudacanthotermes militaris* gut and fecal combs [19]. This species cultivates a termite-specific Basidiomycete fungus, *Termitomyces sp.*, which thrives upon combs made of termite feces. Functional screening recovered 101 clones acting on a range of model substrates containing arabinoxylan and xylan moieties and identified differences in biomass deconstruction potential between microbial communities inhabiting the gut and comb milieus. Functional metagenomic screening has allowed the discovery from a number of environments, however, many remain to be explored.

1.3.2 16S Ribosomal RNA Profiling

To address the question of which species are present within an environment, molecular methods have been developed. This is necessary as it is difficult to determine the taxonomy of prokaryotic cells based on morphology alone. By examining the sequence of marker genes, encoded within the genome, a systematic framework for bacterial taxonomy has been developed. The specific marker gene that is typically used is the small sub-unit ribosomal RNA, also known as 16S rRNA. This is an ideal choice as this gene is ubiquitous, functionally conserved and different regions change at different rates [321]. The 16S rRNA contains nine (9) variable regions [331] which can be targeted with primers, facilitating the amplification of these regions from the genetic background. Amplified DNA can then be sequenced, with short read sequencing technology, producing many thousands of reads. The resulting sequences are then processed with the use of a bioinformatic pipeline, such as QIIME [46]. This pipeline removes low quality sequences, and clusters the sequences (typically at 97% sequence similarity). The resulting bins, refered to as Operational Taxonomic Units (OTUs), can then be assigned a taxonomy based on identity with known 16S sequences.

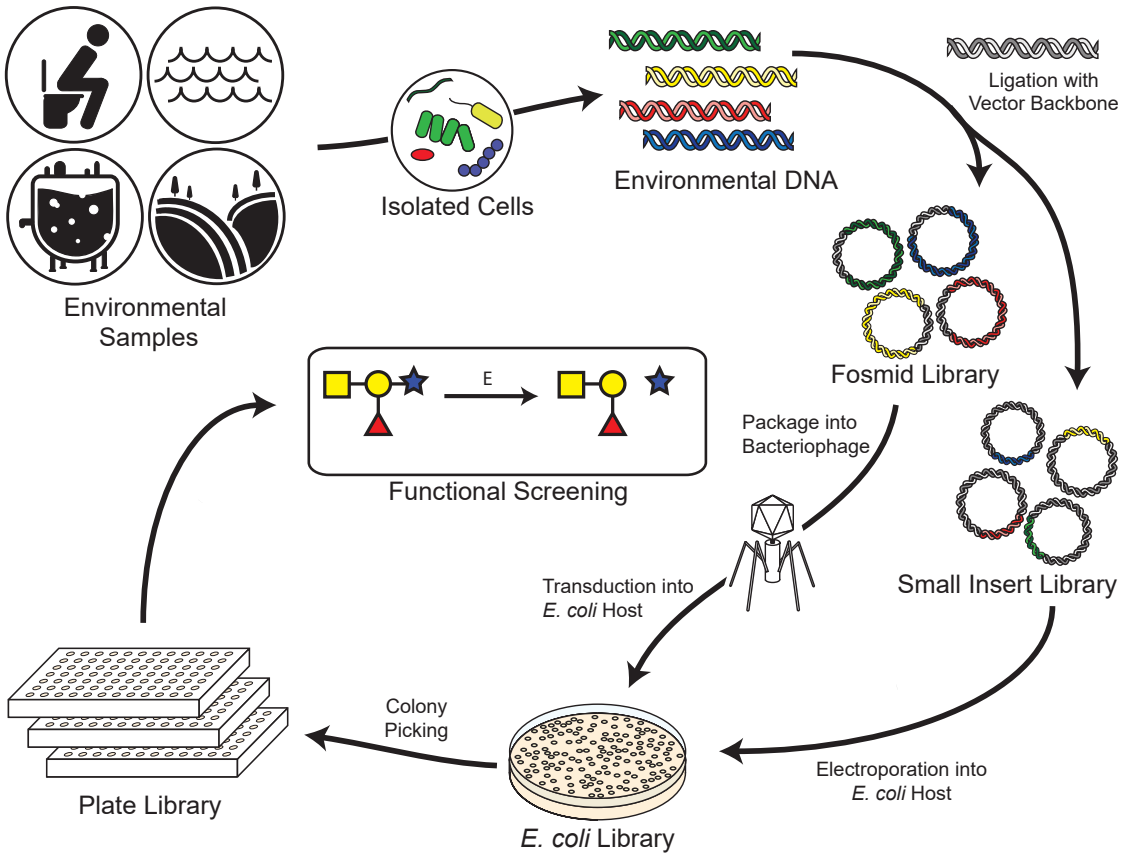

Figure 1.10: *Functional Metagenomic Screening Workflow.* Microbial communities can be interrogated for biological activities through functional metagenomic screening. Environmental DNA can be extracted directly from natural and engineered ecosystems and used to construct screening libraries. A workflow for constructing large insert fosmid libraries and small insert libraries is depicted. Fosmid library production involves high molecular weight environmental DNA preparation, ligation into a vector backbone and head-full packaging of ligated DNA into a phage delivery system. Small insert libraries can similarily be ligated with a variety of vector backbones which can be used to transformed via electroporation. Host cells are then transfected, plated and arrayed in 384-well plate libraries and can be interrogated with a variety of functional screens.

1.4 Dissertation Overview

The aim of this thesis is to analyze the functional aspects of microbial communities that degrade plant polysaccharides and to investigate unexamined environments with high-throughput functional screens. This will lead to the creation of a library of functional clones which can be rapidly interrogated under a variety of conditions with a variety of substrates. Additionally, mutation of these catalysts can produce enzymes that are capable of synthesizing defined glycans containing chemi-

cally modified sugars. This thesis contributes a better understanding of the enzymatic conversion of plant polysaccharides, and to new catalysts for both the deconstruction of plant biomass and synthesis of chemically modified polysaccharides.

Chapter 2 details the use of high-throughput functional metagenomics to screen 22 different environments for cellobioside-degrading activities. This enabled the creation of a panel of active fosmid-harbouring clones, which were further characterized by rapid, plate-based, assessment of the biochemical parameters, and sequencing. This has revealed hundreds of glycoside hydrolases, many of which show low identity to any previously discovered gene.

Chapter 3 describes the application of functional metagenomic screening to the *Castor canadensis* fecal and gut microbiomes. Four fosmid libraries were created from different sites within the digestive tract and fecal matter. These were subjected to functional screening with new and highly-activated substrates specific for cellulose- and hemicellulose-cleaving enzymes. This resulted in the identification of many previously unknown PULs and characterization of enzymes that synergistically degrade arabinoxylans.

Chapter 4 uses the clone libraries generated in Chapters 2 and 3 and a synthetic gene library to detail the promiscuity of glycoside hydrolases. Genes identified with activity towards modified glycosides were then mutated in the hopes of creating glycosynthases that could use modified acceptor sugars. The efficiency and products produced by the created glycosythases are described. This has led to the generation of eight new synthetically useful glycosynthases.

Chapter 5 gives an overall analysis and integration of the research and conclusions of the thesis in light of current research in the field. This chapter also comments on strengths and limitations of the thesis research and presents possible future research directions in the field drawing on the work of this thesis.

Finally, Chapter 6 details the materials and methods used to conduct the research contained within this thesis.

Chapter 2

Large-Scale Functional Metagenomic Screening for Glycoside Hydrolases

2.1 Summary

This chapter presents the high-throughput functional screening of 22 large insert metagenomic libraries and the characterization of active clones. Screening was performed in 384-well plate format with a model substrate (4-methylumbelliferyl cellobioside) that releases a fluorescent molecule when cleaved by β-glucosidases or cellulases, and resulted in 178 verifiably active clones. The substrate specificity, thermal stability and optimal pH of the glycosidase(s) expressed on these clones was investigated in a high-throughput, plate-based format. The insert DNA, harboured within each of these clones, was sequenced and functional annotation revealed a cornucopia of carbohydrate-degrading enzymes. The discovered genes were compared to those of previously characterized glycoside hydrolases, which revealed several genes belonging to clades that have not previously been characterized. The large insert sequences were investigated for the presence of operons and gene clusters, which revealed synteny between fosmids. This well characterized collection of clones serves as a future resource for the development of optimized biocatalysts, whether it be for the degradation of biomass or for other specialized functions.

2.2 Background

Plant biomass offers a sustainable source for energy and materials and an alternative to fossil fuels. However, the industrial scale production or biorefining of fermentable sugars from plant biomass is currently limited by the lack of cost effective and efficient biocatalysts [57]. Microbes,

the earth's master chemists – employing biocatalytic solutions to harvest energy, and transform this energy into useful molecules – offer a potential solution to this problem. Microbial degradation of carbohydrates involves the use of glycoside hydrolases (GH), which offer some of the greatest catalytic rate enhancements among enzymes [336]. GHs catalyse the degradation of a profuse variety of polysaccharides, including cellulose, the most abundant terrestrial biopolymer [31], pectins and hemicelluloses. They are also important industrial catalysts [76, 239, 267], and therapeutic targets [151]. Clearly, the identification of new GH genes has the potential to improve upon both the efficacy of current biocatalysts and the generation of new catalysts for new chemistries.

The Carbohydrate-Active Enzymes database (CAZy) is an expertly curated resource which classifies GH genes into over 140 families based on sequence similarity [181]. The genes within a family often display catalytic specificity towards the same broad category of substrate, which enables the predictive annotation of genes that have not been functionally analyzed [315]. However, this predictive ability often breaks down when a diverse range of substrates are cleaved by enzymes within a family. It has been estimated that the function of only 20 % of the proteins in the sequence database can be predicted with confidence [105]. Additionally, only a small subset of the GH genes within the CAZy database have been functionally characterized; as of 2013 only 6 % of GH genes have had any form of functional characterization [181] and this percentage is surely decreasing with the influx of new genomes that are deposited into the database.

Metagenomic research has taken advantage of massively paralleled, high-throughput DNA sequencing techniques to provide insight into the function of environmental DNA [300]. Several studies have focused on the discovery of GH genes from environmental DNA [120, 176, 315]. This approach serves as a promising avenue for the discovery of new catalysis, however, typically only very few enzymes from a metagenome are functionally characterized. This lack of functional characterization further expands the gap between the total number of genes sequenced and those gene products that have been functionally characterized.

Most efforts to increase the diversity of functionally characterized GH genes have focused on studying one or a few enzymes at a time. More recently efforts utilizing large-scale gene synthesis have enabled the exploration of phylogenetic branches within a family that have not been well characterized [117]. This is a worthwhile method that one could envision being applied to many

enzyme families. However, until the cost of gene synthesis comes down this type of study remains out of reach for a majority of research groups.

Function-based metagenomic activity screens, coupled with high-throughput enzyme characterization, can enable the functional annotation of genes without the bias introduced when annotation is done by sequence comparison and without the need for costly gene synthesis. Functional screens have the ability to provide a direct link between metagenomes and their functional activities. They can also provide the ability to discover enzymes with activities that exist outside the current paradigms of gene annotation, which in turn can better inform *in silico* approaches.

The aim of this study was to produce a library of fosmid clones containing environmental DNA encoding cellobiohydrolase activity, as this function is key to the degradation of plant polysaccharides [105]. Furthermore, we hoped to profile how presence of cellobiohydrolase genes varied across environments expected to either be enriched or depleted in plant biomass. To this end 309,504 clones containing DNA extracted from 22 diverse sites were interrogated with a fluorogenic activity probe. The resulting resource, a panel of 178 clones, enabled us to rapidly investigate the substrate specificity, acid tolerance and thermal tolerance of enzymes expressed by these clones and revealed a diverse set of genes and activities.

2.3 Results and Discussion

A set of twenty-two (22) fosmid libraries were chosen for functional metagenomic screening. These libraries were sourced from a variety of natural and engineered ecosystems, as described in Table 2.1. Ocean water samples were sourced from the North-Eastern sub-Arctic Pacific Ocean at depths ranging from surface to 2000 m [326]. Soil samples were collected from four different depths from disturbed and undisturbed test plots in Skulow Lake, British Columbia [114]. Coal bed samples were produced from coal bed core cuttings or water withdrawn from the coal beds [8]. Bioreactor samples were sourced from an anaerobic mining bioreactor [201], a methanogenic naptha-degrading culture or a methanogenic toluene-degrading culture [288]. As these DNA sources varied drastically in their physiochemical properties (Table 2.1) and microbial community composition, we hoped that this diversity would potentiate the discovery of new catalysts.

Table 2.1: Fosmid Libraries

Name	Project	Sample Type	Ref.	Depth (m)	Temp. (°C)	pH	Clones
12010	Ocean	Water from Station P12	[326]	10	8.4	7.8	7,680
12200	Ocean	Water from Station P12	[326]	500	4.5	7.3	7,680
12500	Ocean	Water from Station P12	[326]	2000	1.9	7.4	7,680
40010	Ocean	Water from Station P4	[326]	10	9.9	7.8	7,680
40500	Ocean	Water from Station P4	[326]	500	5.6	7.4	7,680
41000	Ocean	Water from Station P4	[326]	1000	3.6	7.3	7,680
41300	Ocean	Water from Station P4	[326]	1300	2.9	7.3	7,680
NO	Soil	Natural; Organic horizon	[114]	0	4.1	5.0	10,752
NA	Soil	Natural; Mineral (eluviation)	[114]	0.1	4.1	5.7	13,440
NB	Soil	Natural; Mineral (transition)	[114]	0.3	4.1	6.0	9,984
NR	Soil	Natural; Mineral (accumulation)	[114]	0.55	4.1	6.7	23,040
CO	Soil	Clearcut; Organic horizon	[114]	0	4.1	6.0	16,512
CA23	Soil	Clearcut; Mineral (eluviation)	[114]	0.1	4.1	5.7	9,216
CB	Soil	Clearcut; Mineral (transition)	[114]	0.3	4.1	6.2	21,888
SCR	Soil	Clearcut; Mineral (accumulation)	[114]	0.55	4.1	6.7	10,752
FOS62	Bioreactor	Bioreactor core sample	[201]	0	18.0	6.9	18,432
TolDC	Bioreactor	Toluene degrading culture	[288]	1.5	25.0	7.5	23,040
NapDC	Bioreactor	Naptha degrading culture	[288]	31	28.0	7.5	20,736
CG23A	Coal Bed	Coal bed produced water	[8]	300-500	32.1	7.9	9,600
CO182	Coal Bed	Coal bed cutting	[8]	686	22.0	N.D.	23,040
CO183	Coal Bed	Coal bed cutting	[8]	730	22.0	N.D.	23,040
PWCG7	Coal Bed	Coal bed produced water	[8]	300-500	32.4	7.7	22,272
						Total	**309,504**

2.3.1 *In-Silico* Screening

All 22 of the chosen libraries have had a portion of the clones end-sequenced, meaning that the ends of the insert DNA were sequenced using Sanger-sequencing technology, Table 2.2. To preliminarily assess the potential of these libraries to catalyse the degradation of cellulosic biomass we turned to these end-sequences, as being representative of genes within the library. A total of 176,472 clones were end-sequenced, 57 % of all clones, producing 235 Mbp of sequence data. Open reading frames (ORFs) were predicted from these end-sequences using Prodigal [134] implemented within the MetaPathways bioinformatic pipeline [155] resulting in a total of 400,561 predicted ORFs. These predicted ORFs were then annotated using LAST [150] implemented in the MetaPathways pipeline based on queries of the CAZy database [181], revealing a total of 3,953 predicted Glycoside Hydrolases(GHs).

Table 2.2: End Sequences Interrogated From Each Library

Library	Project	End Sequences	Predicted ORFs	GH Genes
12010	Ocean	12,477	12,769	50
12200	Ocean	14,740	17,472	106
12500	Ocean	14,886	17,495	96
40010	Ocean	14,275	15,771	90
40500	Ocean	14,705	16,715	107
41000	Ocean	14,701	16,935	116
41300	Ocean	14,488	16,601	111
CO	Soil	15,360	17,086	235
CA	Soil	15,360	16,903	166
CB	Soil	15,360	17,441	185
SCR	Soil	15,360	16,303	188
NO	Soil	15,360	17,577	164
NA	Soil	15,360	17,288	198
NB	Soil	15,360	17,413	212
NR	Soil	15,360	17,259	174
FOS62	Bioreactor	37,632	40,255	837
TolDC	Bioreactor	15,360	16,618	131
NapDC	Bioreactor	15,360	17,126	143
CG23A	Coal Bed	15,360	16,779	195
CO182	Coal Bed	15,360	17,329	225
CO183	Coal Bed	15,360	17,678	209
PWCG7	Coal Bed	15,360	23,748	15
	Total	**352,944**	**400,561**	**3,953**

Of the predicted GHs, 320 (0.080 % of all predicted ORFs) were found to belong to families

that have β-glucosidase activity, but not cellulase activity (GH1, GH3, GH30, GH116) with GH3 being the most abundant (246 ORFs, 0.061 % of predicted ORFs). With respect to cellulases, 256 (0.064 % of predicted ORFs) were found to belong to families that contain members with cellulase activity (GH5, GH6, GH7, GH8, GH9, GH10, GH12, GH26, GH44, GH45, GH48, GH51, GH74 and GH124) with GH5 being the most abundant (105, 0.026 % of predicted ORFs). The distribution of cellulases and β-glucosidases varied greatly between libraries, Figure 2.1. The library derived from an anaerobic mining bioreactor (FOS62) had the highest abundance of predicted cellulases and β-glucosidases (n = 169, 0.42 % of predicted ORFs), likely a reflection of the feed stock for the bioreactor (bacterial biomass and partially degraded and composted cellulose). The soil libraries had the next highest abundance of cellulases and β-glucosidases (24.6 $\pm$ 6.5 ORFs, 0.14 % $\pm$ 0.04 % of predicted ORFs), which one would expect due to the presence of cellulose in the form of decaying plant matter. The ocean and coal bed samples were relatively depleted in cellulases and β-glucosidases (0.08 % $\pm$ 0.03 % and 0.08 % $\pm$ 0.05 % of predicted ORFs respectively) which in turn may be attributed to the paucity of cellulose in these environments. Additionally, coal bed samples show a lack of diversity in the number of cellulase families found; they contain almost exclusively GH5 enzymes. This dearth is also reflected by the counts of cellulases and β-glucosidases predicted for the naptha- and toluene-degrading enrichment cultures (0.08 % and 0.06 %, respectively), furthermore, the end-sequences of these libraries contain almost no enzymes predicted to belong to cellulase families (TolDC = 1, NapDC = 2).

Table 2.3: Highly Repetitive Short ORFs from PWCG7

ORF	Length	Counts	Blast Hit	Identity
1	124	5838	holin [*Pseudomonas stutzeri*]	98%
2	174	5096	phage terminase [*Pseudomonas stutzeri*]	99%
3	112	3459	pyocin R2, holin [*Pseudomonas stutzeri*]	100%

Of the sequences investigated, the PWCG7 library had the fewest predicted cellulases and β-glucosidases (n = 1, 0.004 % of all ORFs). An additional peculiarity is that this library has a substantially higher number of predicted ORFs when compared to other libraries of a similar size, Table 2.2. Further investigation revealed a redundancy among the predicted ORFs; there were only 3,807 unique ORFs within the 23,748 predicted ORFs. Three ORFs in particular were found over

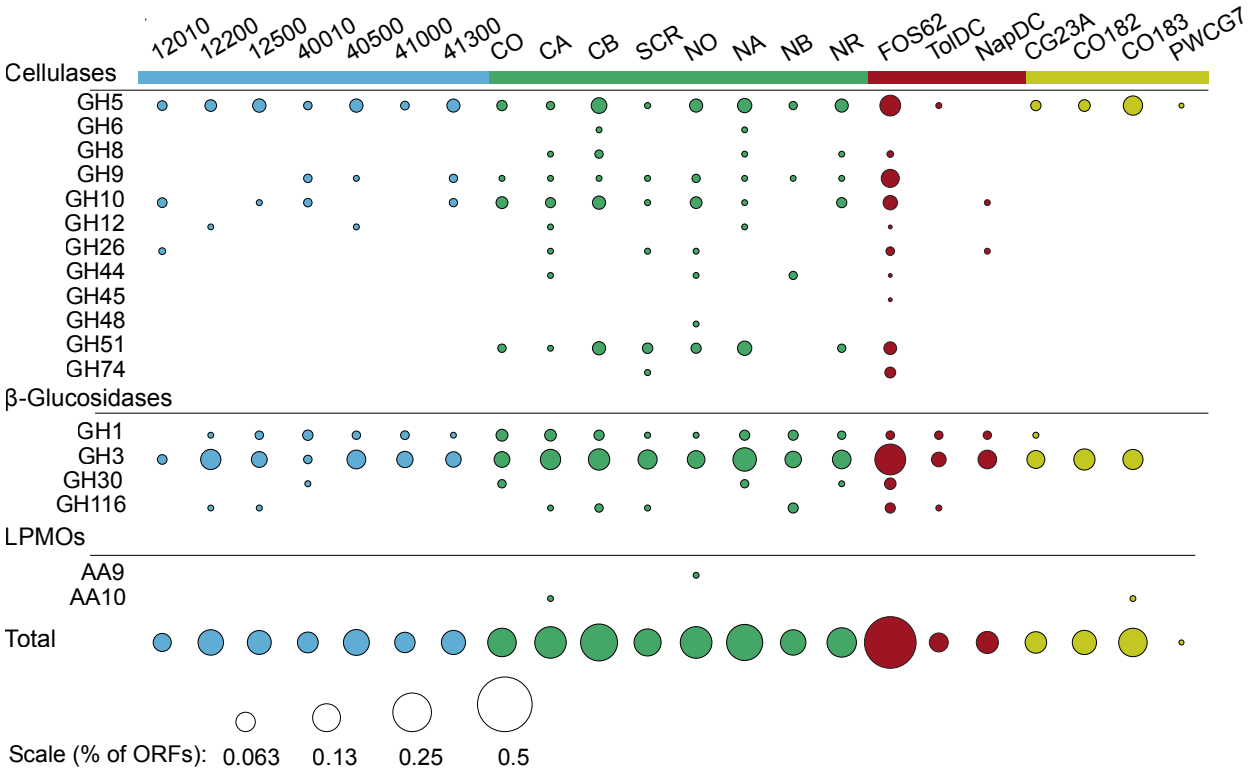

Figure 2.1: *In-Silico Screening of Fosmid End-sequences.* Bubble plot of CAZymes which are predicted to have activity on cellulose or cellooligosaccharides. Families GH7 and GH124 were omitted as there were no predicted genes belonging to these families. Bubble area is proportional to the percentage of all predicted ORFs within a specific familiy

3,000 times within the end-sequences (Table 2.3) and each of these is a small phage protein. This is suggestive of either phage contamination within this portion of the PWCG7 library or that the environment was sampled during a period of viral bloom.

The presence of lytic polysaccharide mono-oxygenases (LPMOs) was also investigated. These enzymes are classified as Auxiliary Activities (AA) in the CAZy database [171] and AA9 and AA10 have been observed to oxidatively cleave cellulose chains and act synergistically with other cellulases [94, 126]. However only 4 AA9 or AA10 proteins were predicted from the fosmid end-sequences (2 AA9s in the NO library and 1 AA10 each in the CA and CO183 libraries). One can speculate that this dearth of LPMOs may be caused by the anoxic or anaerobic nature of a majority of the samples that were used to create libraries.

Taking this information together, the expectation for functional screening would be that the

FOS62 library is likely to produce the greatest number of hits, followed by the soil libraries. Screening of the ocean, coal bed, TolDC and NapDC libraries, on the other hand, is likely to result in a smaller number of functional hits and in the case of the coal bed libraries I would expect a low diversity in the number of families that are functionally identified.

2.3.2 Functional Screening

All 22 libraries were screened by Sam Kheirandish with a fluorogenic substrate, 4-methylumbelliferyl cellobioside (MU-C), designed to detect cellulase, cellobiohydrolase and β-glucosidase activity. This substrate releases the fluorophore 4-methylumbelliferone (MU) which can be detected at concentrations in the nanomolar range, Figure 2.2. MU-C offers an increase in the sensitivity over the previously employed 2,4-dinitrophenyl cellobioside (DNP-C), a chromogenic substrate which had been used to screen a third (6,144 of 18,432) of the FOS62 clones [201], as fluorescent detection is inherently more sensitive. The use of this substrate was adapted to the screening paradigm employed by Mewis et. al. [201] which enabled the rapid screening of over 300,000 clones.

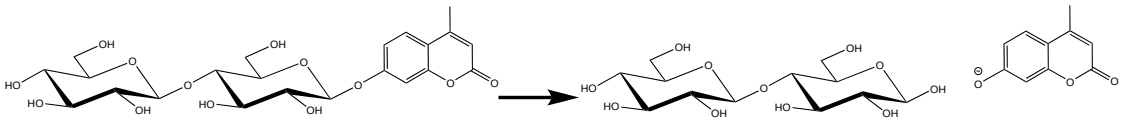

Figure 2.2: *Fluorogenic Reporter 4-Methylumbelliferyl Cellobioside.* Cleavage of the reducing end acetal linkage releases a fluorescent molecule, facilitating the detection of glycoside hydrolases.

Screening revealed 256 hits with a plate-based z-score (the number of standard deviations above the mean) above 10, a hit rate of 1 in 1209 clones, although this varied drastically between libraries, Figure 2.3. Of these hits, 178 were verified after re-streaking and triplicate validation, Table 2.4. As expected from the annotation of fosmid ends the FOS62 library had the highest hit rate of any of the libraries (1 in 222). The library with the next highest hit rate was the TolDC library (1 in 768), followed by the Soil libraries (average hit rate of 1 in 2,627). The coal bed libraries were comparatively poor with an average hit rate less than half that seen for the soil libraries (average hit rate of 1 in 5,996) while the ocean libraries only produced a combined total of 3 hits from the over 50,000 clones screened (average hit rate of 1 in 17,920)

Figure 2.3: *Functional Screening of All Libraries with MU-C.* Z-score values for fluorescence were calculated for each plate. Clones above the a z-score of 10 were chosen for further validation.

The general trend observed for the number of hits fit well with the expectations from fosmid end-annotation (FOS62 > soil > coal bed, ocean, methoanogenic bioreactors). However, there were a few notable exceptions. Firstly it was unexpected that the TolDC library would have the second highest hit rate from all the libraries screened. This enrichment culture was supplied with toluene as its sole carbon source, so it is quite surprising that the library obtained from this source shows such a capacity to degrade cellobiosides. Additionally, the absence of any hits from the CA23 soil library was unexpected, a result which may underline the inherent stochasticity of screening. Another unexpected result was the presence of any hits from the PWCG7 library. As this library had the worst frequency of annotated cellulases and β-glucosidases it was expected to have the fewest number of hits. However, PWCG7 had more hits than all 7 Ocean libraries combined.

Access to both end-sequences and functional screening results also allowed us to empirically estimate the recovery rates for each of the libraries screened. Using the most frequently seen β-

Table 2.4: Functional Screening Hits

Library	Source	Verified Hits	Clones per Hit
12010	Ocean	0	-
12200	Ocean	1	7,680
12500	Ocean	1	7,680
40010	Ocean	0	-
40500	Ocean	1	7,680
41000	Ocean	0	-
41300	Ocean	0	-
NO	Soil	14	768
NA	Soil	8	1,680
NB	Soil	5	1,997
NR	Soil	3	7,680
CO	Soil	5	3,302
CA	Soil	0	-
CB	Soil	7	3,127
CR	Soil	2	5,376
Fos62	Bioreactor	83	222
TolDC	Bioreactor	31	743
NapDC	Bioreactor	4	5,184
CG23A	Coal Bed	3	3,200
CO182	Coal Bed	4	5,760
CO183	Coal Bed	2	11,520
PWCG7	Coal Bed	4	5,568
Total	Ocean	3	17,920
Total	Soil	45	2,569
Total	Bioreactor	118	527
Total	Coal Bed	13	5,996
Total	All Libraries	178	1,738

glucosidase and cellulase families from the end-sequences (GH3 and GH5) we can estimate the expected number of ORFs that belong to these families on all of the clones within each of the libraries 2.5. Comparison of the number of GH3s and GH5s recovered from the fosmids gives us some insight into the hydrolase recovery rates and how this changes across environments. The average recovery rate was approximately 2.5 % for both GH3 and GH5 families, however it was highly variable between libraries. The ocean library had the lowest recovery rates (GH3 = 0.17 %, GH5 = 0.98 %), while the TolDC library had the highest recovery rate seen (GH3 = 9.71 %, GH5 = 13.87) . The differences in recovery percentages seen is likely due to multiple factors, including: regulation and expression of the genes, the ability of *E. coli* to properly translate the genes, and whether the protein products are active under the screening conditions used. One caveat

of interpreting this data is that not all GH3s and GH5s are active glucosides, with some family members targeting other substrates, such as -N-acetylglucosaminides or xylosides in the case of GH3s.

As the FOS62 library had been previously screened with a different, chromogenic, substrate (DNP-C), the performance of MU-C could be compared to this benchmark. For all FOS62 clones screened (n = 18,432) there were 90 colonies determined to be hits with DNP-C (z-score = 6), while 83 were uncovered with MU-C (z-score = 10), and 35 of these clones being found in both screens. These two leaving groups appear to access somewhat different sets of enzymes as 103 of the total 138 fosmids recovered (75 %) were only identified with a single substrate. DNP-C is more reactive, as the pK_a of the 2,4-dinitrophenyl leaving group ($pK_a = 4.09$) is substantially lower than that of MU ($pK_a = 7.79$), resulting in reduced activation energy for bond cleavage. The DNP-C probe however lacks the sensitivity of fluorogenic MU-C. A chemical activity probe bearing a fluorescent leaving group with low pK_a may afford a larger number of clones, and offer an improved hit rate over either DNP-C or MU-C.

Table 2.5: GH3 and GH5 Recovery Rates.

	Expected		Recovered (%)	
Library	GH3	GH5	GH3	GH5
Ocean	577	307	0.17	0.98
Soil	1,861	679	2.58	1.77
Coal Beds	912	510	0.77	0.98
Fos62	762	345	6.04	6.95
TolDC	216	36	9.71	13.87
NapDC	315	0	0.64	N/A
Total	**4,942**	**2,109**	**2.53**	**2.42**

N/A: Could not be calculated.

2.3.3 High-throughput Characterization of Fosmids

To gain further insight into the discovered hits, high-throughput characterization of the fosmid clones was performed by Dr. Feng Liu and Tanya Duo. This characterization exploited the use of a Biomek FX workstation (Beckman Coulter) and plate-based assays to gain insight into the substrate specificity, pH dependence and thermal stability of identified clones without the need for enzyme sub-cloning and purification. Though it should be noted that as there may be more than one gene expressed, this characterization may reflect the activity of more than one enzyme.

Substrate Preference

The 178 fosmid clones identified were assayed against a panel of eight different glycosides bearing a MU leaving group. This panel of substrates consisted of the cellobioside, lactoside, β-D-glucopyranoside, β-D-galactopyranoside, β-D-xyloside, α-L-arabinofuranoside, β-D-mannopyranoside and β-D-*N*-acetylglucosaminide. Many of these monosaccharides and disaccharides are present in the hemicellulosic and pectic fractions of wood. A majority of clones were most active against either the glucoside or cellobioside substrate, however, there were a substantial number of clones that had higher activity against other substrates (see Figure 2.4 and Table A.1). Clones with optimal activity against MU α-L-arabinofuranoside and MU β-D-xyloside were the next most abundant with counts of 34 and 10 clones respectively. These sugar monomers are essential components of hemicellulose [264], thus fosmids active on these substrates may be active against hemicellulose in addition to their activity on glucosides. The presence of either multifunctional enzymes or multiple genes located in gene clusters such as PULs is a likely explanation for the multiple activities seen.

Optimal pH determination

To ascertain the optimal pH for each fosmid clone assays were performed with the optimal substrate for that clone in a number of solutions buffered at a pH ranging between 4.0 and 9.8. The average pH optimum was 5.6 $\pm$ 0.7, with the largest number of clones having an optimal pH of between 5 and 6 (142 of 178 clones), Figure 2.5. Of the clones with pH values greater than 7.5 a disproportionate number were derived from the ocean environment, likely reflecting the slightly

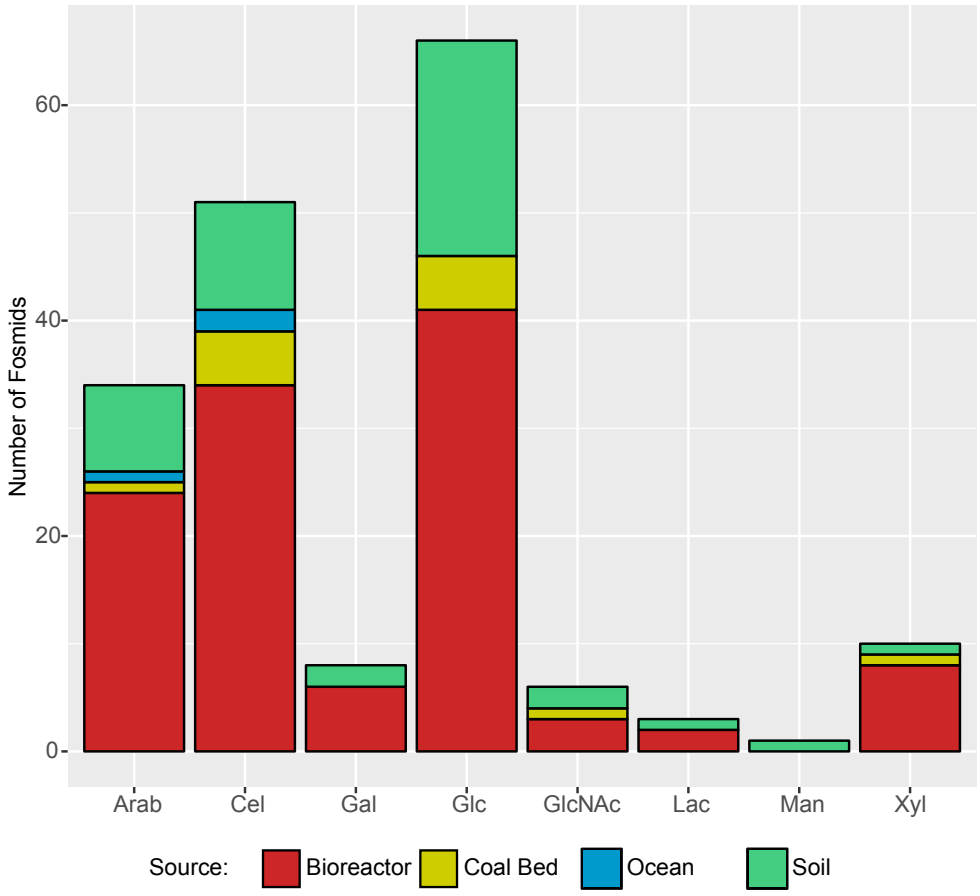

Figure 2.4: *Fosmid Substrate Preference.* Each fosmid containing clone was assayed against eight substrates: MU cellobioside (Cel), MU lactoside (Lac), MU β-D-glucopyranoside (Glc), MU β-D-galactopyranoside (Gal), MU β-D-xyloside (Xyl), MU α-L-arabinofuranoside (Arab), MU β-D-mannopyranoside (Man) and MU β-D-*N*-acetylglucosaminide (GlcNAc). Initial rates were determined using crude cell lysate to determine the optimal substrate for each clone.

alkaline pH of the open ocean. A total of 5 clones were observed to have the lowest pH optima (CB006_04_L11, FOS62_34_K14, NO001_13_N07, PWCG7_49_G20, TolDC_59_K14), being most active in pH 4 buffered solutioins. No clear correlation between the sample pH and the optimal pH of the fosmid clone was observed. One possible explanation for this lack of correlation may be the intracellular use of a subset of these enzymes, causing the pH optima to be a reflection of intracellular pH rather than that of the environment.

The pH range observed for fosmid clones assayed is typical of most β-glucosidases and cellulases which have been studied to date. There are however some notable exceptions. For example, alkaline

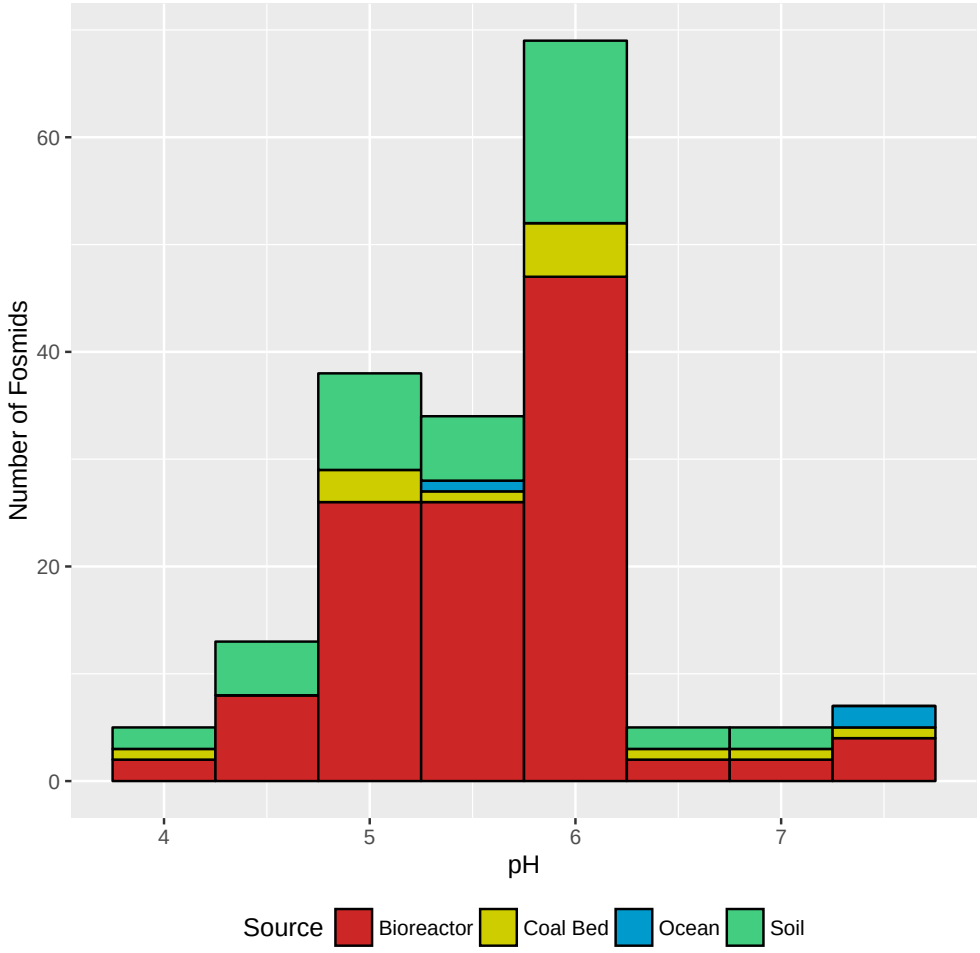

Figure 2.5: *pH Optima of Fosmid Clone Activity.* Initial rates were used to determine the pH at which the fosmid harbouring clones best catalysed the degradation of the optimal substrate.

cellulase K from *Bacillus* sp. strain KSM-635 [273] has optimal activity at a pH of 9.5 almost two units away from the most alkaline tolerant clone found here. On the other end of the spectrum the endo-glucanase SSO1949 from the archaea *Sulfolobus solfataricus* has an very low pH optimum of 1.8 [132], substantially lower than the most acidic fosmid uncovered here. This low optimal pH is likely a reflection of the extremely low pH optima (pH of 2-4) for this species [271]. Extremely acidic or alkaline activity is, however, not necessary for the successful implementation of cellulases or β-glucosidases in commercial cellulase cocktails. The most commonly used cellulase cocktail, Cellic® CTec3 (Novozymes, Copenhagen, Denmark), has a pH optimum of 5.0 - 5.5, a range in

which nearly 30 % of the fosmid clones had optimal activity, and an even greater number were active.

Thermal Stability

Further characterization was performed by Tanya Duo and Dr. Feng Liu to determine the thermal stability of the activity seen for each clone. Assays were performed with the optimal substrate for each clone at its optimal pH, after preheating at a range of temperatures between 37 °C and 90 °C. The resulting rates were used to determine the denaturation midpoint temperature (T_m). The T_m values determined spanned a range from 38 °C to 74 °C and had an average value of 50.7 ± 6.4 °C, Figure 2.6. One noteworthy observation was that three of the four highest T_m values determined were for fosmids from the PWCG7 library (PWCG7_33_K24, PWCG7_19_J20, PWCG7_19_I21 with T_m values of 69, 69 and 74 °C respectively). The PWCG7 library was sourced from coal bed produced water that was at a temperature of 32.4 °C, the highest temperature for any environment screened here (Table 2.1), consistent with this library producing the clones with the highest T_m.

Taken in the context of the scientific literature the T_m values of the recovered clones are modest. Proteins from extremophilic organisms such as *Pyrococcus furiosus* are much more likely to have extremely thermotolerant enzymes. In fact the endoglucanase from *P. furiosus* has a temperature optimum of 100 °C and a T_m of 112 °C [20]. However, the current mixtures of hydrolytic enzymes such as Cellic® CTec3 exhibit optimal activity at moderate temperatures (50 - 55 °C). A total of 27 % of the fosmid clones had T_m values at or above 55 °C, signifying that although the T_m values for recovered clones were not extreme, there are quite a few that are acceptably stable to use as enzyme cocktail additives.

2.3.4 Fosmid Sequencing and Gene Annotation

Validated and characterized fosmids were then fully sequenced and assembled by Dr. Keith Mewis, Sam Kheirandish and myself to reveal the active genes present on each fosmid insert. Sequencing of 178 clones produced 6.2 Mbp of assembled data with an average fosmid insert size of 35 ± 5 kbp, Figure 2.7. Comparison between sequences identified 123 non-redundant clones based on greater

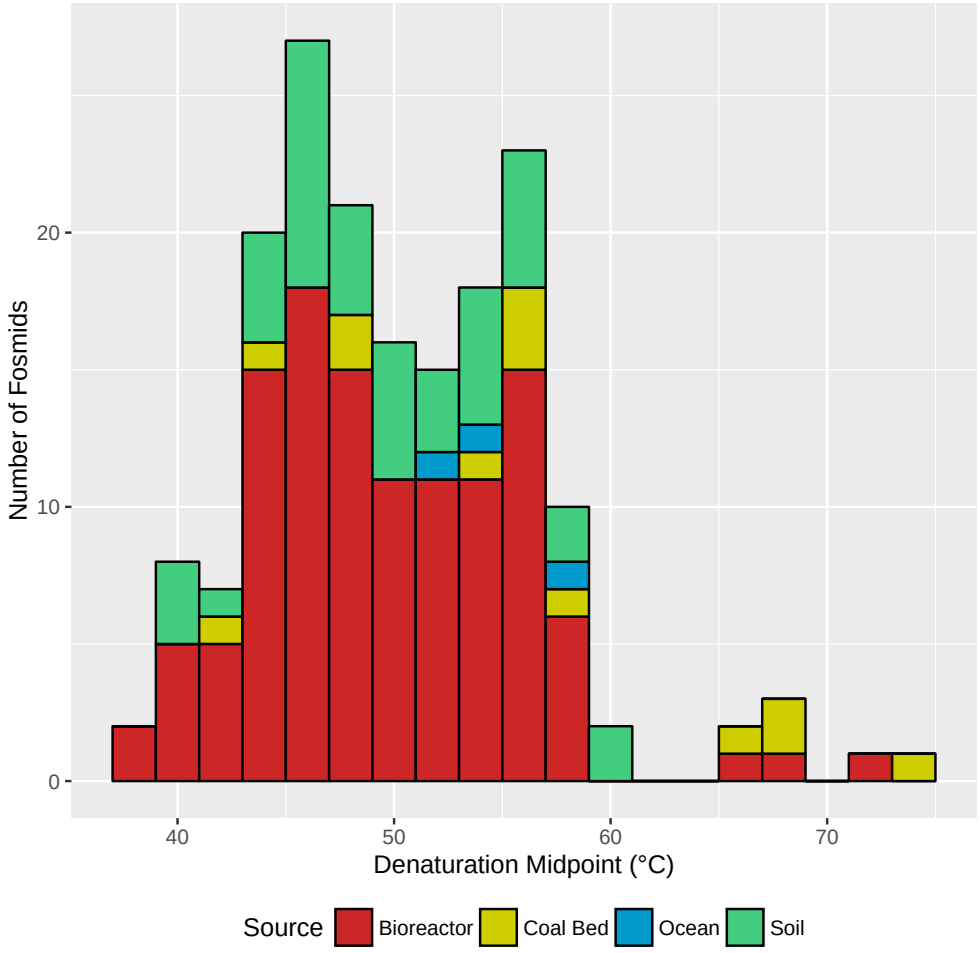

Figure 2.6: *Thermal Stability of Fosmid Clone Activity.* Denaturation midpoints were determined for all clones by first pre-incubating lysate over a range of temperatures and then assaying the clones with the optimal substrate to determine the initial rates of hydrolysis.

than 95 % similarity across more than 90 % of insert length. The redundant clones were most prevalent in the Fos62 and TolDC libraries (60 and 18 redundant clones respectively), while there were no clones meeting the redundancy criteria identified within any of the soil libraries. This suggests that more sequence diversity is captured in the soil libraries.

GH abundance

Across all fosmids 4,653 ORFs were predicted, an average of 26.1 ± 5.5 per fosmid. These ORFs were queried against the CAZy database [181] with LAST [150] implemented in the MetaPathways

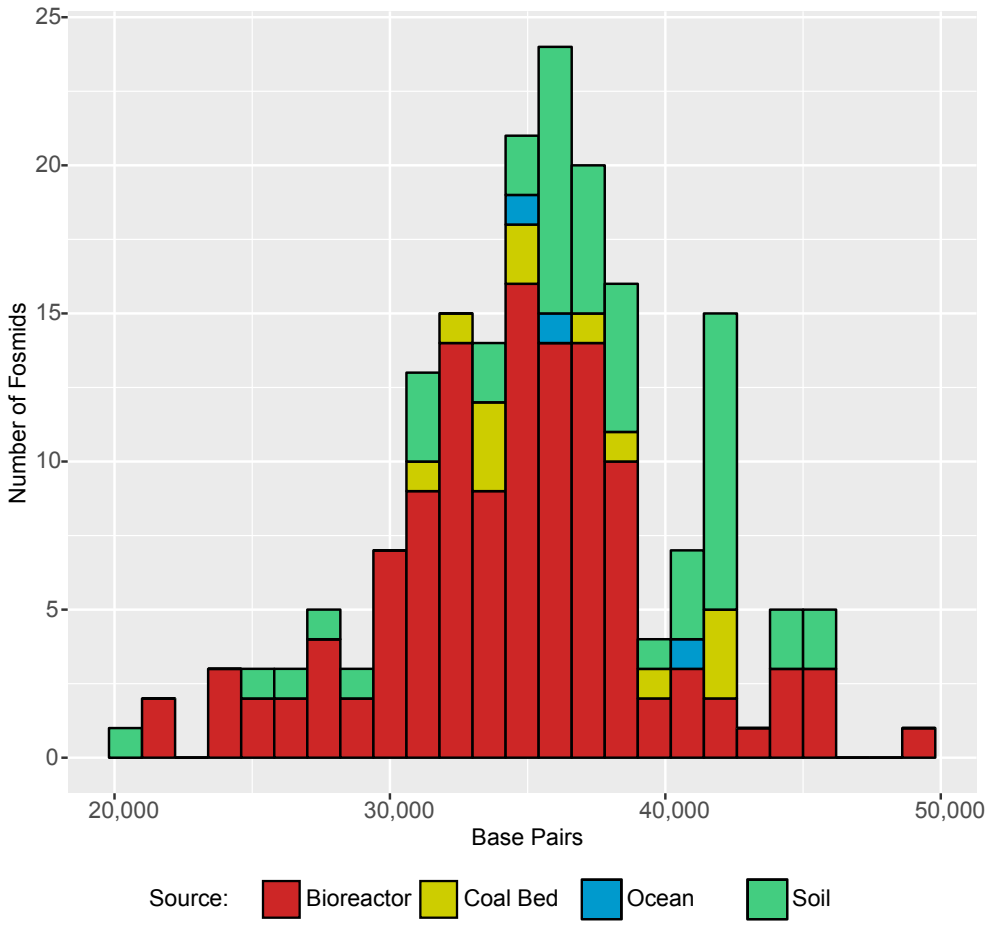

Figure 2.7: *Distribution of Fosmid Insert Length.* Histogram showing the number of sequenced fosmids with a specified length, bars are coloured by the library source.

pipeline [155]. This revealed 516 ORFs annotated as glycoside hydrolases, Figure 2.8. All of the identified fosmids contained a GH belonging to a known β-glucosidase or cellulase family. The annotated GHs spanned 48 families, including all 7 families with β-glucosidase activity (GH1, GH3, GH5, GH9, GH30, GH39 and GH116) and 8 of 14 cellulase families (GH5, GH9, GH8, GH10, GH12, GH26, GH44, and GH51). Of the six cellulase families that were not found (GH6, GH7, GH45, GH48, GH74 and GH124), neither GH7 nor GH124 were identified from the fosmid end-sequences, thus their absence is unsurprising. Although GH45s were identified on FOS62 library end-sequences, the majority (92 %) of GH45 sequences in CAZy are eukaryotic, which may explain the inability to detect any with *E. coli* as a host. The remaining cellulase families that escaped detection were GH6, GH48 and GH74. Of these, GH6 and GH48 are both thought to act processively from the non-reducing end, which, in turn may require tighter binding in the positive enzyme subsites to the cellulose polymer. As MU-C lacks glucose residues in the "+" subsites, this is a feasible justification for the absence of these families. The family GH74 on the other hand is largely composed of endoxyloglucanases, and only one enzyme is seen to have better activity on glucans than xyloglucans [55], the absence of this family from the observed hits can be justified by the scarcity of its action on un-decorated glucans.

The two most abundant hydrolase families recovered from the fosmid inserts, GH3 and GH5 (2.7 and 1.1 % of fosmid ORFs, respectively), were also the most abundant β-glucosidase and cellulase families identified in the fosmid end-sequences. A further 6 hydrolase families were found at rates greater than 0.5 % of all ORFs (GH16, GH43, GH10, GH30 and GH1, in order from most to least abundant), though not all of these families contain cellulases or β-glucosidases. The third most abundant family (GH16, 0.8 % of predicted fosmid ORFs) is not annotated as containing cellulases or β-glucosidases, but a portion of its characterized members do have activity on glucan polymers with mixed 1,3- and 1,4-linkages. All 30 of the fosmids annotated as containing a GH16 also have either a GH3 or GH5 present. The large number of GH16s recovered is therefore likely due to their association with cellulases or β-glucosidases in clusters of genes that work together to degrade glucans.

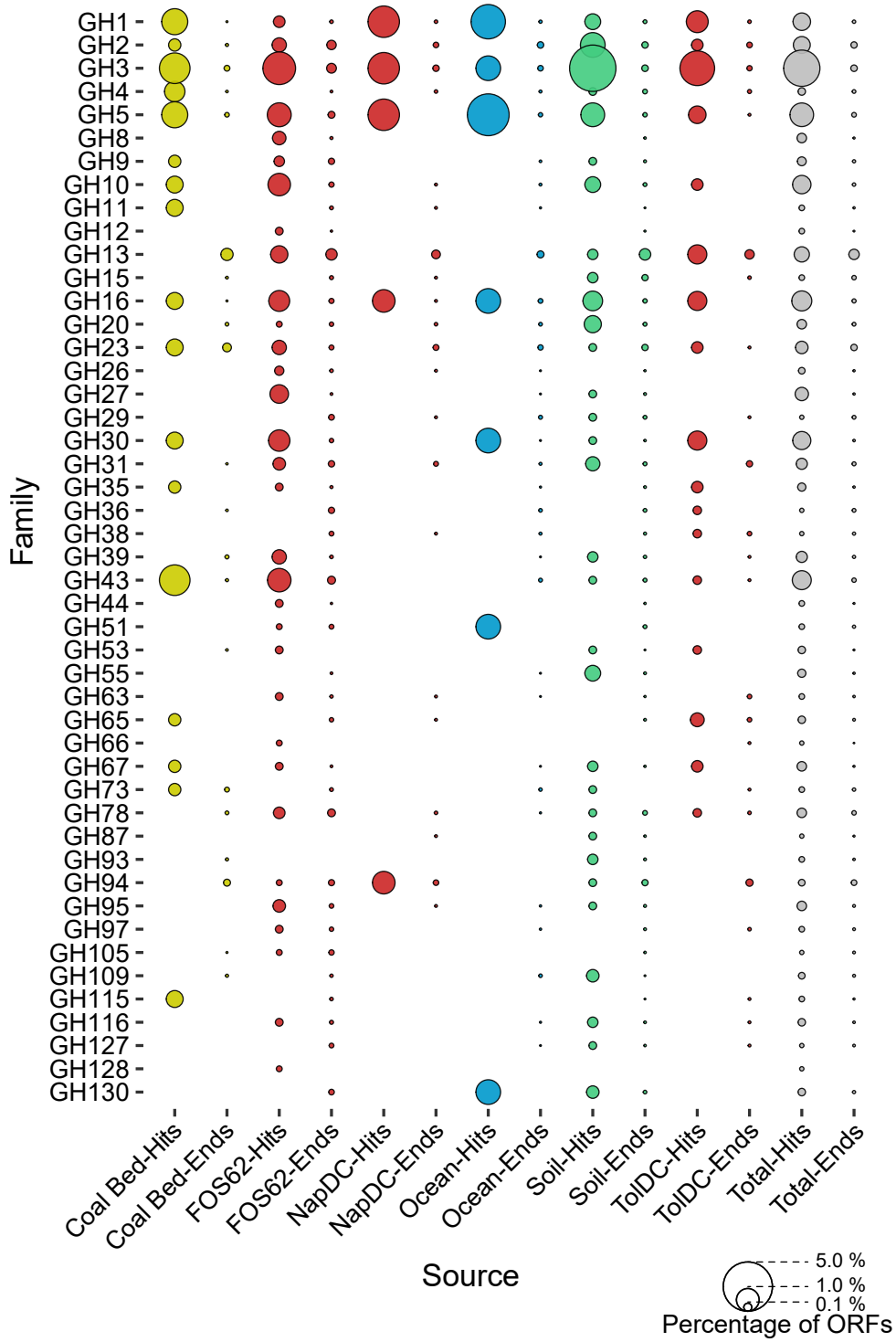

Figure 2.8: *Predicted GH Abundance on Fosmids Hits and on End Sequences* Bubbles show the relative abundances of each GH family recovered from positive fosmid clones for each library source. The bioreactor results are shown for each library. Fosmid encoded GHs are compared to those recovered from end-sequencing of fosmids. Bubbles are coloured by library.

GH43 enzymes, have also not been described as cleaving β-glucans, rather, they are known to act on β-xylosides, α-L-arabinofuranoside, and β-galactans, which are key components of hemicellulose [202] and are often found in hemicellulose- and pectin-degrading loci [215, 257]. As with the GH16 family, the high abundance of GH43 genes can be ascribed to their genomic co-localization with cellulases or β-glucosidases. Furthermore, the abundance of GH43s is likely the cause of the high percentage of hits with arabinosidase and xylosidase activity.

The distribution of predicted fosmid hydrolases was consistent across environments, barring a few exceptions. One abberation was the low number of GH1s predicted from the FOS62 library (5 genes, 0.23 % of predicted ORFs) when compared to all other environments (22 genes, 0.86 % of predicted ORFs). The FOS62 library was also quite diverse in the range of cellulases recovered. Hydrolases belonging to famlies GH8, GH12, GH44 and GH51, were only recovered from the FOS62 library, even though all of these were seen in the end-sequences from the soil libraries. One further surprise was that the GH30 family was predicted on fosmids from coal bed libraries. End-sequencing of these libraries revealed a paucity in the diversity of hydrolase families, with only GH1, GH3, and GH5 families expected to be recovered.

I also sought to investigate how the activities seen in the high-throughput characterization related the genes present on the recovered fosmids. To gain insight into which families are likely responsible for these permiscuous activities the fosmids were divided into sets based on their optimal substrate, Figure 2.9. All fosmids were active on MU-C, thus the presence of GH1s and GH3s and GH5s whithin each of the subsets was unsurprising. However, there were substantial differences between the percentage of certain hydrolases families seen in each subset. The sets with highest activity on monosaccharides all had greater numbers of GH3s, as compared to the set most active on the cellobioside. Fosmids with the highest activity on cellobiosides, usurprisingly, had a higher percentage of ORFs assigned to cellobiohydrolase containing families GH5 and GH8. The set with highest activity on MU-Glc had the highest diversity of GH's seen of any of the sets, which is likely a reflection of the greater number of fosmids within this group. The fosmids which were optimally active on xylose had a higher percentage of GH43s, a family containing β-xylosidases, than any other group of fosmids. This set of fosmids was also highly enriched in GH67 genes, which is surprising as no members of this family have been identified with β-xylosidase activity. The over

abundance of this family, may however be due to its frequent incorporation into glucuronylxylan cleaving gene cassettes, rather than its activity on xylosides.

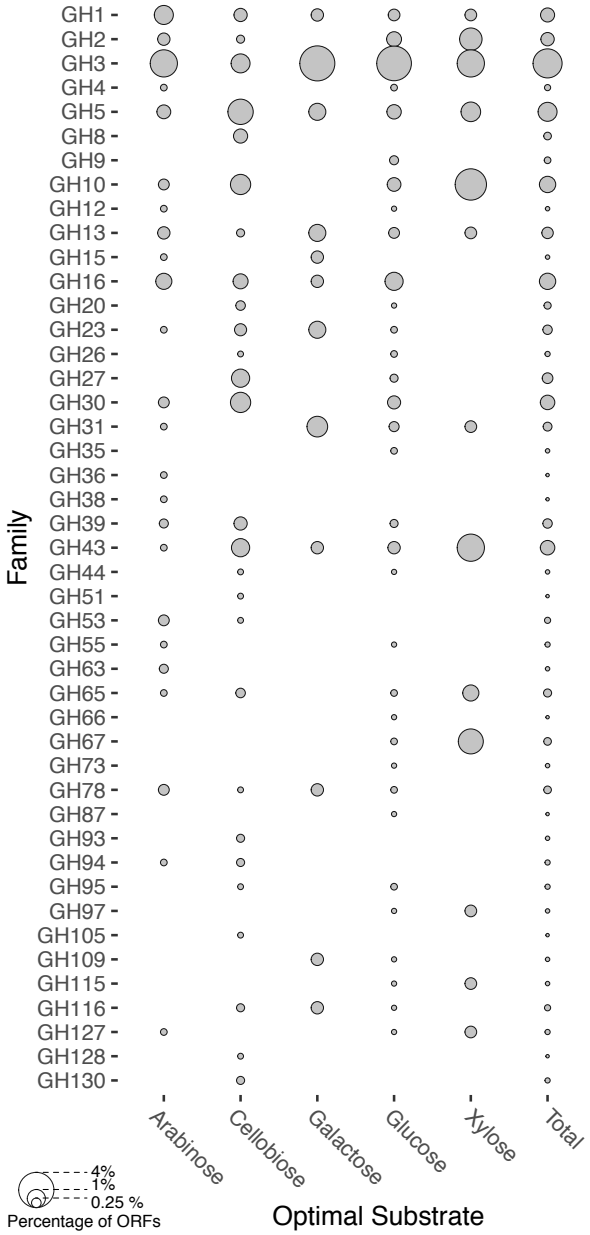

Figure 2.9: *Hydrolase Distribution with Optimal Substrate.* Bubble area is proportional to the percentage of ORFs that belong to each hydrolase family. Fosmids with optimal activity on MU β-D-mannopyranoside, MU lactoside or MU β-D-*N*-acetylglucosaminide are excluded, as there were fewer than 5 fosmids within each of these categories.

To assess the distinctness of recovered glucanases, ORFs belonging to families containing β-

glucosidase or cellulase activity were queried against the National Center for Biotechnology Information (NCBI) non-redundant protein database (accessed April 2017) using BLAST [7]. The proteins recovered were quite distinct, with an average maximum identity of 65.6 ± 12.5 %, Figure 2.10. Of the 331 proteins queried only 15 had a homologous protein with greater than 90 % identity. The lowest % identity uncovered was that of the GH3 FOS62_47_P19-8, which had a maximum identity of 40.5 %. These results highlight the ability of functional metagenomic screening, not only to find functional proteins, but to also reveal a large number of novel proteins.

Phylogenetic Tree Construction

We also sought to gain insight into how the recovered proteins relate to previously characterized members of the same family. Towards this goal we constructed maximum likelihood phylogenetic trees for the glucanase families that were found most frequently on the recovered fosmids (GH1, GH3, GH5, GH8, GH9, GH10 and GH30). Trees were constructed with GH proteins identified through fosmid sequencing and those denoted as characterized in the CAZy database. These two sets of proteins were clustered separately at 95 % to remove duplicates or highly similar sequences. The two sets of proteins were then combined for alignment with COBALT [230], which was trimmed with trimAl [45] and a tree was generated using RAxML [284].

GH1

Many of the recovered GH1s clustered together, even though they were from different libraries, Figure 2.11. The majority of recovered GH1s (15 of 18) clustered within one clade – the group of all proteins originating from a single branch point– which was almost entirely populated with proteins with β-glucosidase activity. Surprisingly, two of the identified GH1s clustered with the phospho-β-glucosidases (TolDC_30_A19-17 and CG23A_01_C20-5), which generally do not display activity against non-phosphorylated glucosides. Additionally, neither of these fosmids contain another hydrolase belonging to a different glucanase family, though the coal bed-derived clone also contained a GH4, a family with phospho-β-glucosidase activity. However, both of these fosmids also contain predicted cellobiose PTS systems (transporters that couple the phosphorylation of sugars to their uptake [198]) in the same reading frame and in close proximity to the identified GH1 genes. This suggests that these clones first phosphorylate MU-C during uptake and before degradation.

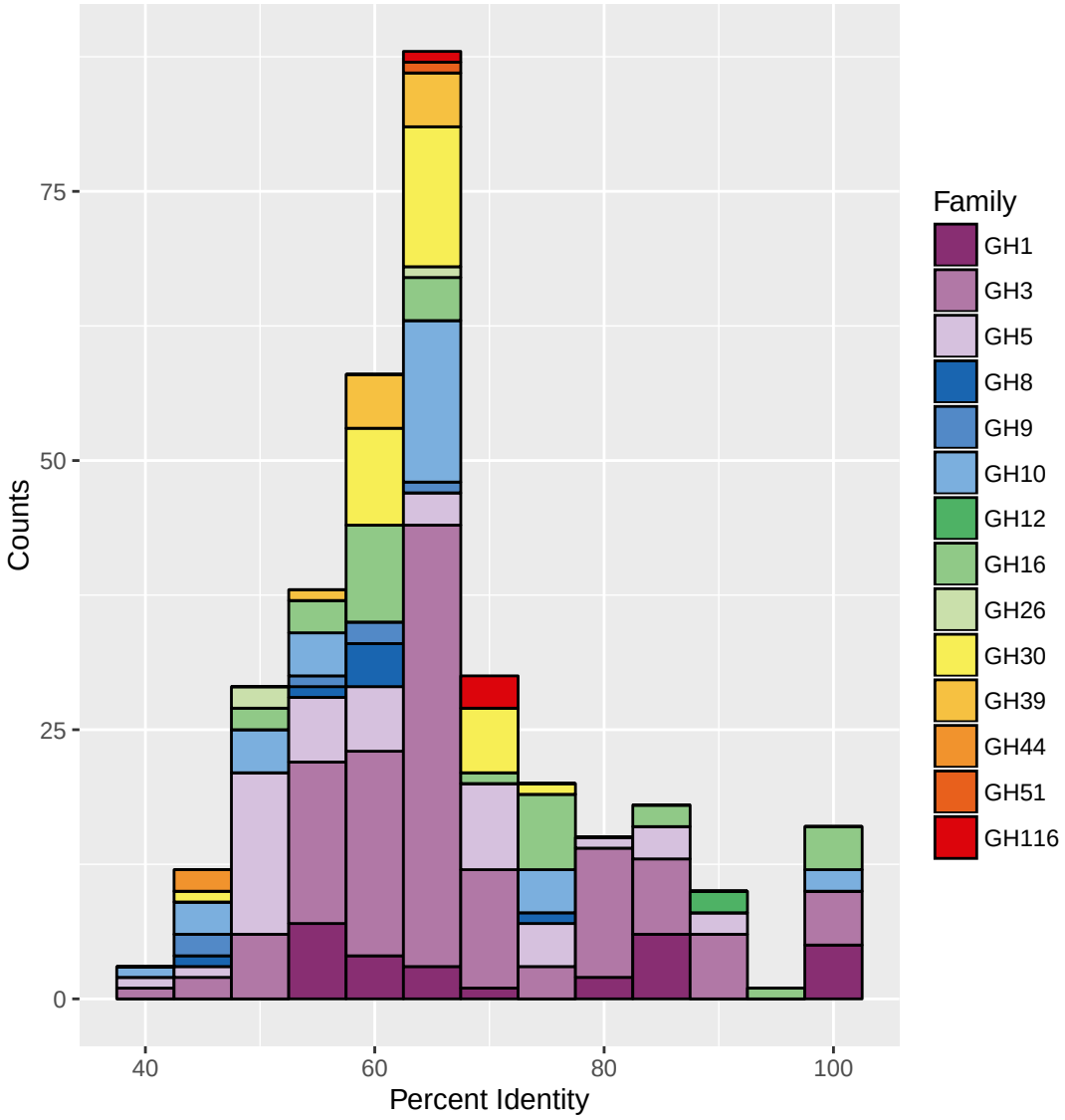

Figure 2.10: *Percent Identities of Best Blast Hits to Putative Hydrolases* All putative hydrolase genes predicted to belong to β-glucosidase or cellulase families were queried against the blast-nr protein database.

GH3

The discovered GH3 proteins appeared more widely distributed as compared to GH1s, Figure 2.12, and tended to be more similar to each other than to previously characterized proteins. Cluster A was deeply branching, with a bootstrap value of 100, and was dominated by GH3s uncovered in this study. The previously characterized proteins within cluster A were the β glucosidase gluA from *Dictyostelium discoideum*, BoGH3B, a β glucosidase present in a xyloglucan degrading PUL from *Bacteroides ovatus* and A1_9 a protein that had been previously discovered when the FOS62 library had been screened with DNP-C. Cluster B, also deeply branching, contained a large number of metagenome-derived genes, however this cluster also had a larger number of previously characterized GH3s than seen in cluster A. The coal bed hits within this group cluster closely with GH3s from the gammaproteobacteria *Cellvibrio japonicus* and *Saccharophagus degradans* which have been characterized [40, 216]. The soil and bioreactor hits within this group however, cluster separately from any previously characterized protein. Cluster C contains 12 GH3s found within this study and a number of proteins from uncultured sources. Two of these proteins (D1_14 and H1_5) were identified in the previous screening of the FOS62 library [201] and three were identified from a termite gut [340]. This cluster also contains the BglX from *E. coli* [330] and Lin1840 from *Listeria innocua*, which has been crystalized [274].

There were a number of fosmid-derived proteins (NA002_01_B04-2, SCR03_04_B15-17, NO001_03_P09-1, FOS62_38_D22-5, NO001_04_B04-0, NO002_11_N21-0, TolDC_31_E21-18, and CO003_01_D22-3) which clustered with enzymes that have activity on xylosides, rather than the expected glucosides (Figure 2.12, cluster D). Of these fosmids, five (NA002_01_B04, SCR03_04_B15, NO001_03_P09, NO001_04_B04 and NO002_11_N21) had more than one predicted GH3, with the additional GH3(s) clustering in a β-glucosidase group. The remaining three fosmids had many GH genes, each with a GH gene from another family likely to be responsible for activity. Finally, cluster E was entirely made up of clones from this study (n= 15), and fell within a clade that has almost entirely β-glucosidase activity.

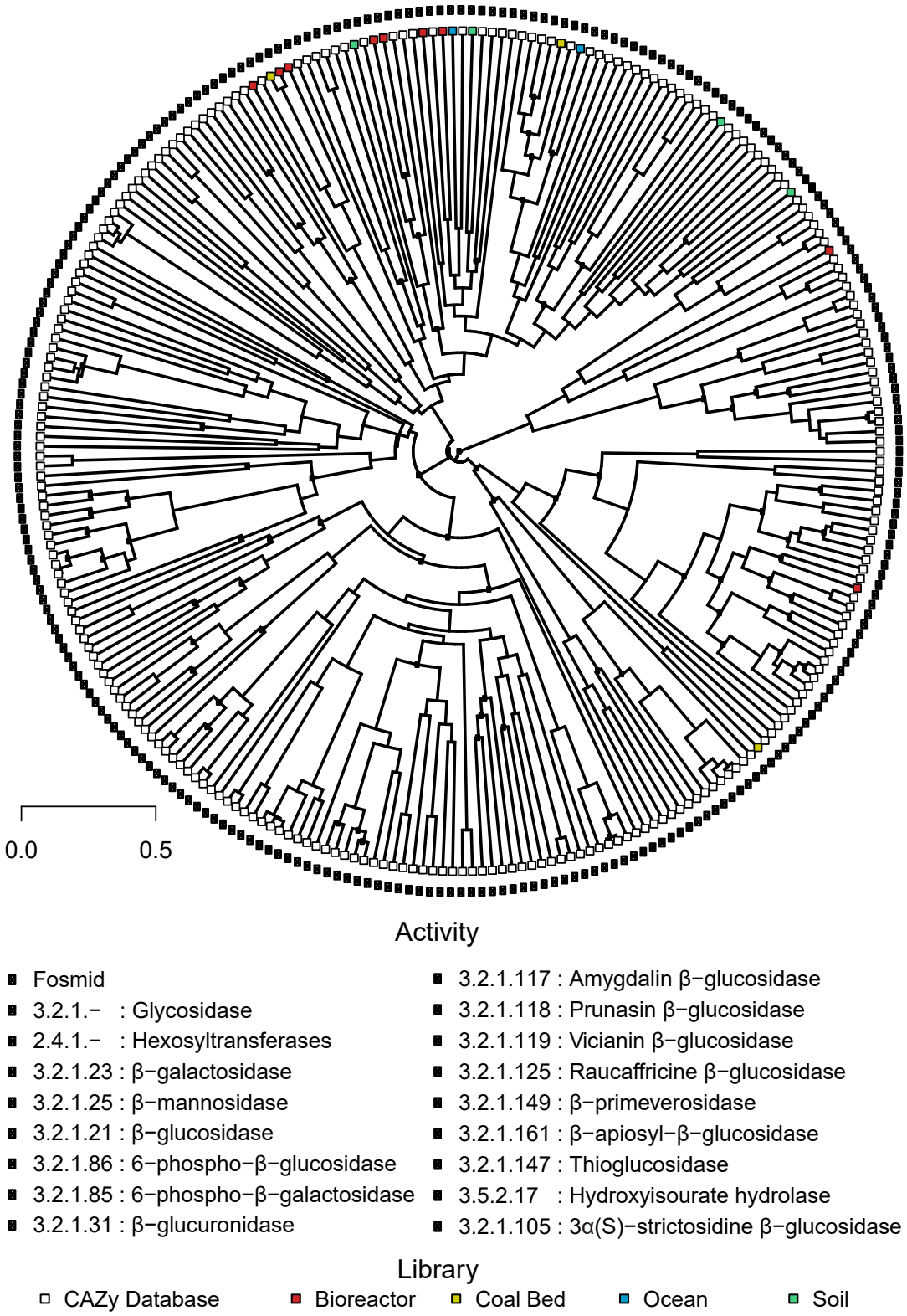

Figure 2.11: *Phylogenetic Tree Containing Discovered GH1s.* The inner ring of squares represents the library from which each protein was retrieved. The outer ring of coloured circles signifies the activity that each characterized protein has annotated in the CAZy database. Branch points with bootstrap values > 70 % are signified with a small black dot.

GH5 The tree generated from GH5 proteins had clustering which was in agreement with both the subfamily designations [15] and the activities of characterized proteins, Figure 2.13. Of the 29 predicted GH5s identified from screening, 11 clustered with previously characterized proteins (2 each in subfamily 4, 26 and 27, and one each in subfamilies 2, 7, 25, 28 and 36). There were 3 clusters of fosmid-derived GH5s that appeared at branches between well clustered subfamilies. The first of these was between subfamilies 12 and 52. This group of soil library GH5s, had the highest similarity to UmCel5F a cellulase within subfamily 39 which was derived from a pulp effluent metagenome [327]. The second cluster of discovered GH5s that did not fit well within a defined subfamily lies between subfamilies 36 and 22. This group contains proteins identified from all four types of libraries screened and one previously characterized protein, SARM_0025. This protein, the sole characterized protein in subfamily 46, was identified from metageomic sequencing of a cow rumen and displayed activity on 1 % Carboxymethyl cellulose agar [120].

The final cluster of discovered GH5s sits between subfamilies 9 and 15, and contains 3 proteins from CO182 and NapDC libraries. The closest characterized gene to this cluster, CW-EG1, is a gene from an uncultured bacterium present in the gut of a cutworm (*Agrotis ipsilon*) [254], which belongs to subfamily 44. However, all three of these GH5s within this cluster have highest sequence identity with subfamily 45 proteins, which currently lacks a characterized representative. Furthermore, two of these three fosmids (CO182_24_J12 and CO182_11_I14) have no other predicted glucanases besides the GH5 gene, making these targets for future characterization.

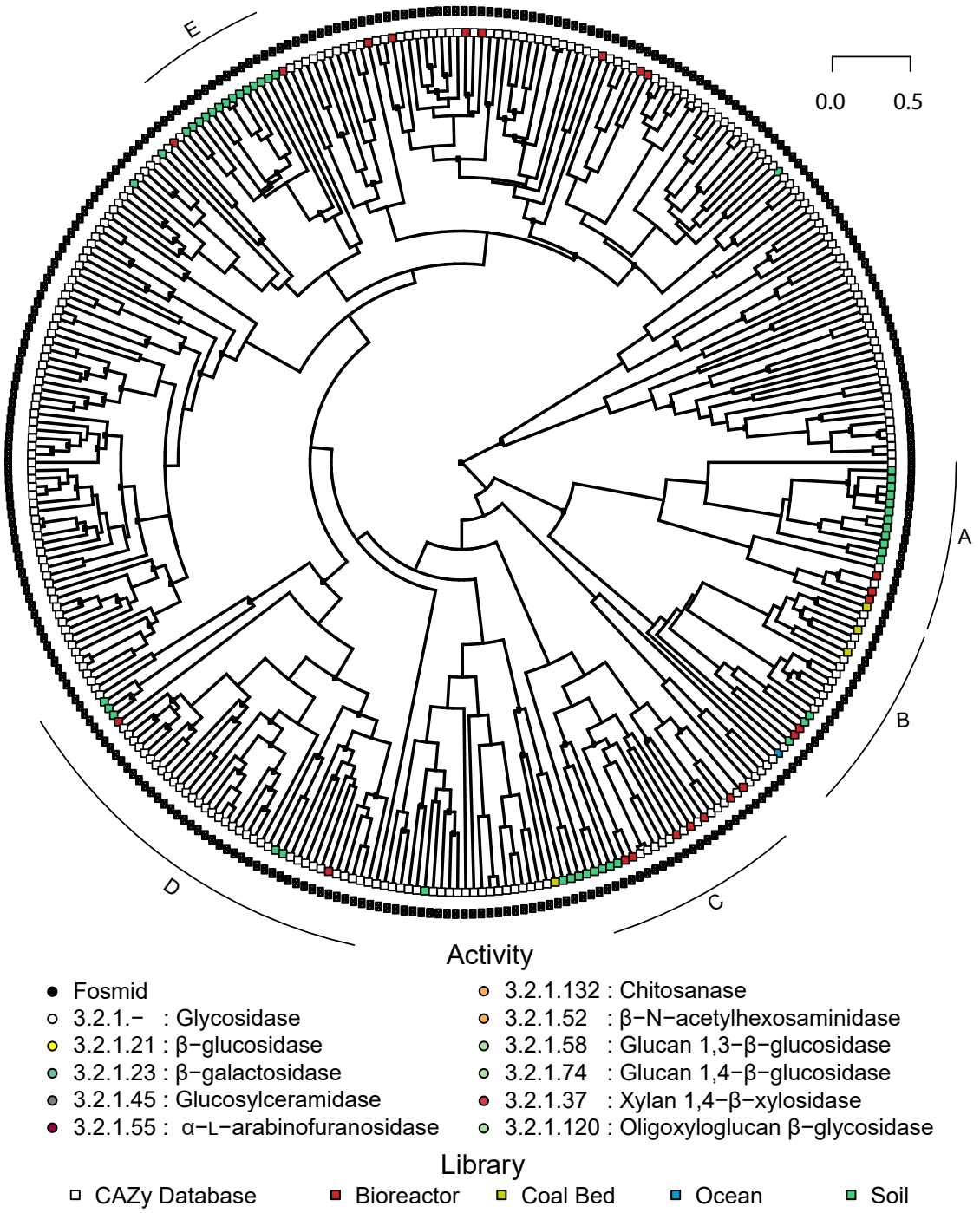

Figure 2.12: *Phylogenetic Tree Containing Discovered GH3s.* The inner ring of squares represents the library from which each protein was retrieved. The outer ring of coloured circles signifies the activity that each characterized protein has annotated in the CAZy database. Arc segments represent clusters of proteins containing metagenome-derived GH3s. Branch points with bootstrap values > 70 % are signified with a small black dot.

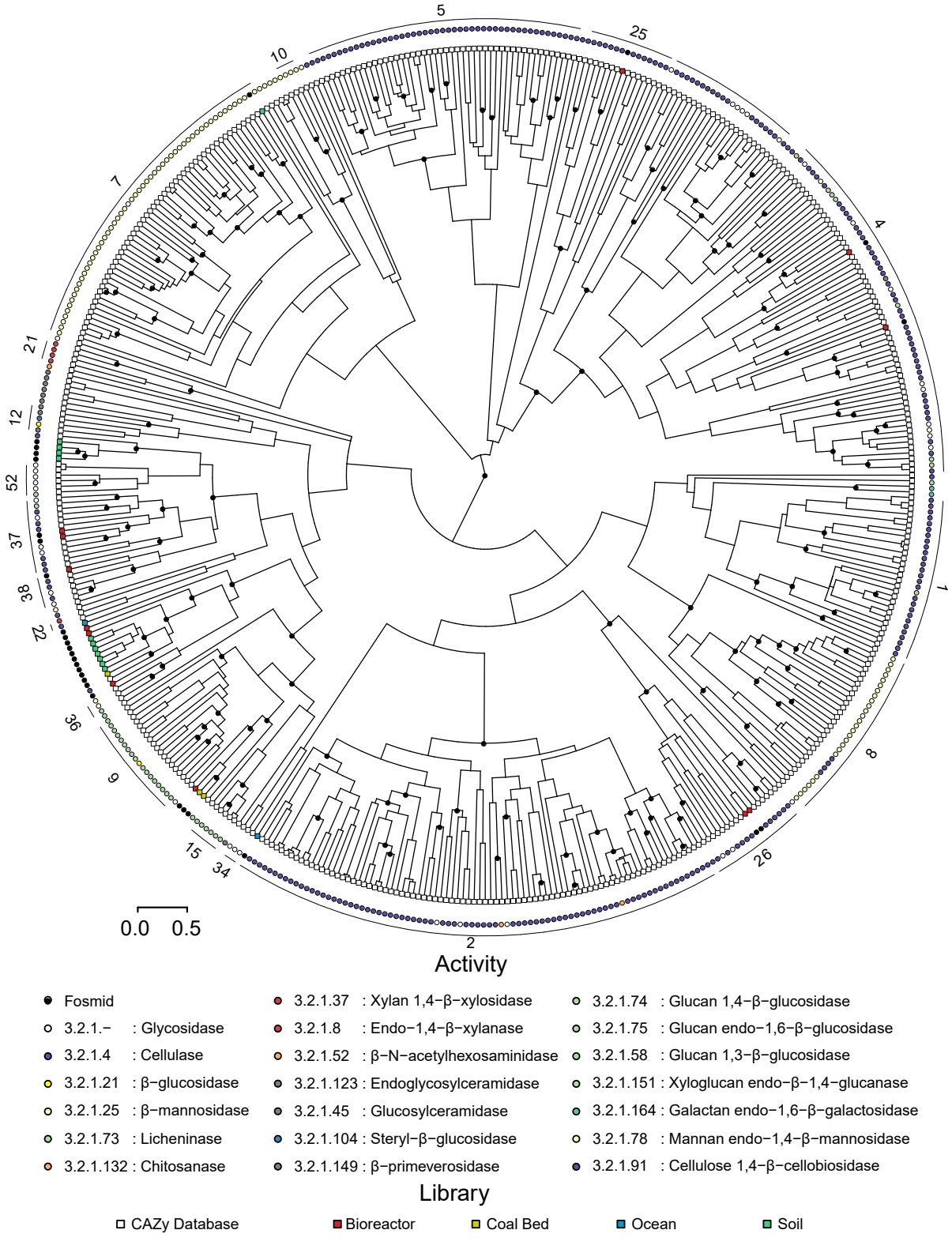

Figure 2.13: *Phylogenetic Tree Containing Discovered GH5s.* The inner ring of squares represents the library from which each protein was retrieved. Coloured circles signifies the activity that each characterized protein has annotated in the CAZy database. Subfamilies are identified with the outermost arc segments [15]. Branch points with bootstrap values > 70 % are signified with a small black dot.

GH8

A total of seven GH8s were identified from the fosmid metagenomic libraries and all of these were found in the FOS62 library. The identified GH8s fell into two groups when clustered at 95 %. These two clusters group separately in the constructed phylogenetic tree, Figure 2.14. The first of these proteins FOS62_40_E07-2 groups closely with Cel8B from *Fibrobacter succinogenes*, an endoglucanase with the highest observed activity on carboxy-methyl cellulose [247]. The second of these genes, FOS62_26_K06-32 is branching from a deeply positioned node (bootstrap value = 98) and belongs to a clade which is dominated by enzymes with cellulase activity.

GH9

The GH9 family is almost exclusively composed of enzymes with cellulase or cellobiohydrolase activiy, Figure 2.15. The four discovered GH9s within this tree fall into two clades, both with bootstrap values greater than 95. The first clade contains two putative GH9s FOS62_30_N01-10 and SCR03_04_B15-0. The bioreactor sourced FOS62_30_N01-10 clusters closest to Cel9B, a multidomain protein from *Ruminococcus champanellensis* that also contains a GH16 domain. This protein is annotated within the CAZy database as a licheninase, however only the GH16 domain was characterized [206], and this annotation cannot be assigned to the GH9 domain. SCR03_04_B15-0 groups with Cel9U from *Clostridium cellulolyticum* and CelD from *Ruminiclostridium thermocellum*, both with activity on insoluble celluloses [102, 249]. The second clade, containing FOS62_25_H06-9 and CO182_36_O01-5, was also almost entirely occupied by bacterial cellulases. The characterized GH9s EgC, Cel9B and SARM_0002 from *Fibrobacter succinogenes* BL2, *Fibrobacter succinogenes* subsp. succinogenes S85 and an uncultured organism from a cow rumen, respectively were most similar to FOS62_25_H06-9. All three of these GH9s were active on a variety of soluble and insoluble celluloses [23, 120, 247]. Lastly, CO182_36_O01-5 was most similar to UmCel9B, from a compost soil metagenome which displayed activity on Carboxymethyl cellulose amongst other glucans [229].

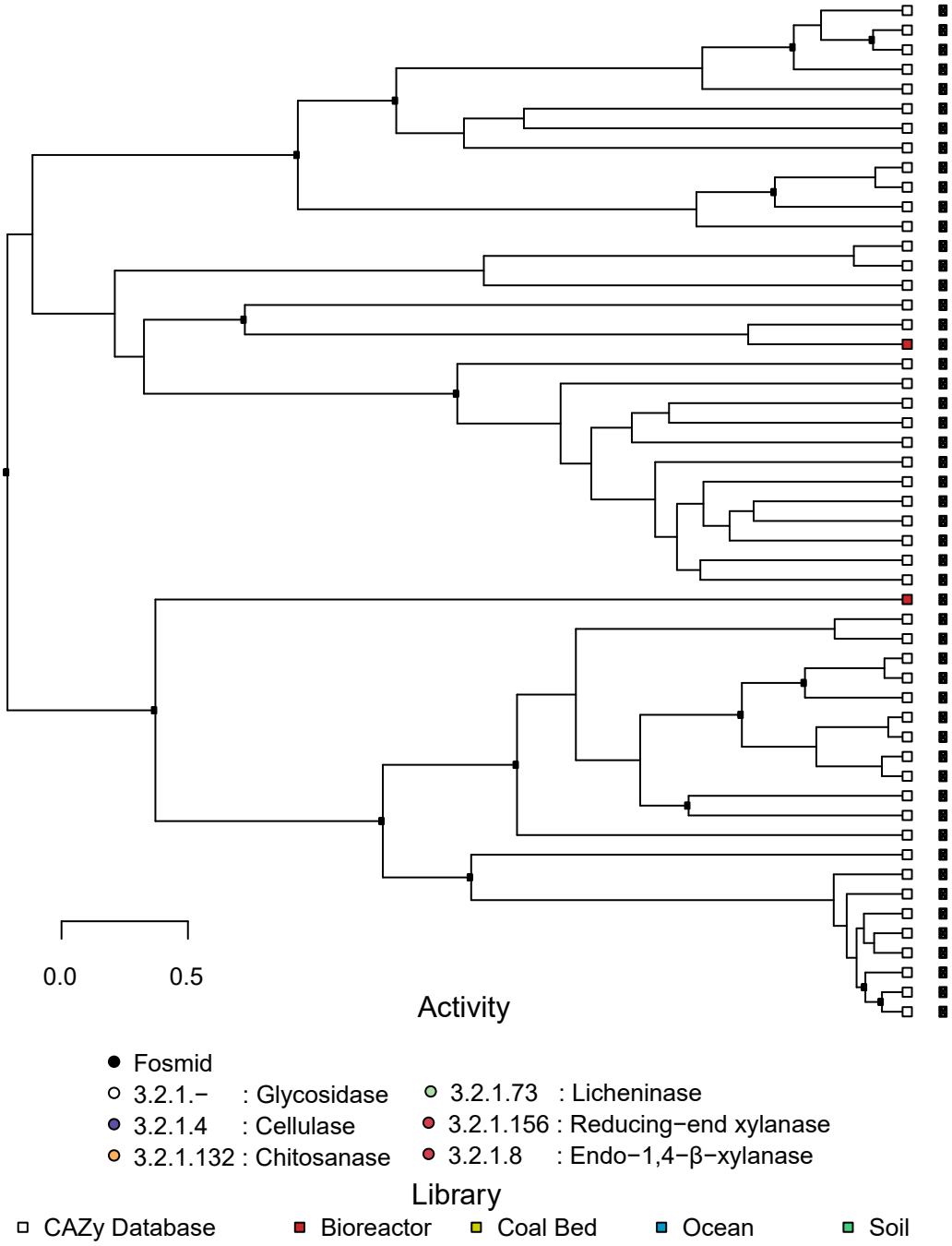

Figure 2.14: *Phylogenetic Tree Containing Discovered GH8s.* The squares placed at branch tips represent the library from which each protein was retrieved. Coloured circles signify the activity that each characterized protein has annotated in the CAZy database. Branch points with bootstrap values > 70 % are signified with a small black dot.

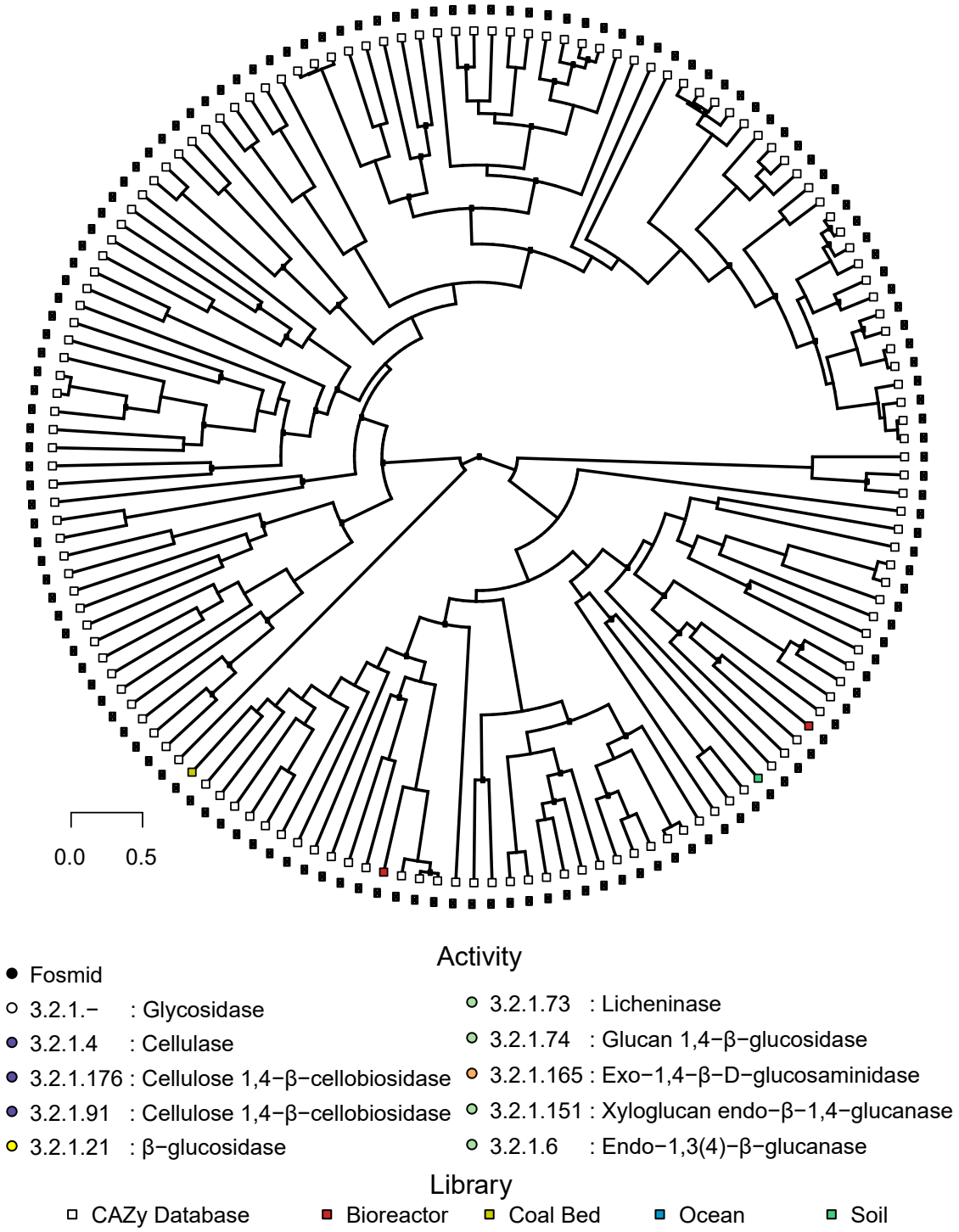

Figure 2.15: *Phylogenetic Tree Containing Discovered GH9s.* The inner ring of squares represents the library from which each protein was retrieved. The outer ring of coloured circles signifies the activity that each characterized protein has annotated in the CAZy database. Branch points with bootstrap values > 70 % are signified with a small black dot

Presence of Gene Clusters

The wealth of fosmids encoding unexpected activities prompted an investigation of the Polysaccharide Utilization Loci (PUL) and gene clusters present on the fosmid hits. PULs have been reported in several cases to synergistically target complex plant polysaccharides, and their composition can give insight into the targeted polymers [71, 145, 165, 215, 257, 294]. A total of 17 fosmids were found to contain PULs, as indicated by the presence of the hallmark SusD/SusC-like protein pairing, Figure 2.16. Several of the identified PULs shared nucleotide identity; FOS62_08_G04, FOS62_08_J18, FOS62_10_O15 were identical at the nucleotide level over the PUL region as were FOS62_29_F15 and FOS62_38_A06. Furthermore, fosmids FOS62_37_N04 and FOS62_38_C16 had 74 % nucleotide identity over 97 % of identified PUL. Synteny was also seen for identified PULs, with the SusC/SusD pair being followed closely by a GH3 in all identified PULs and the frequent inclusion of one or

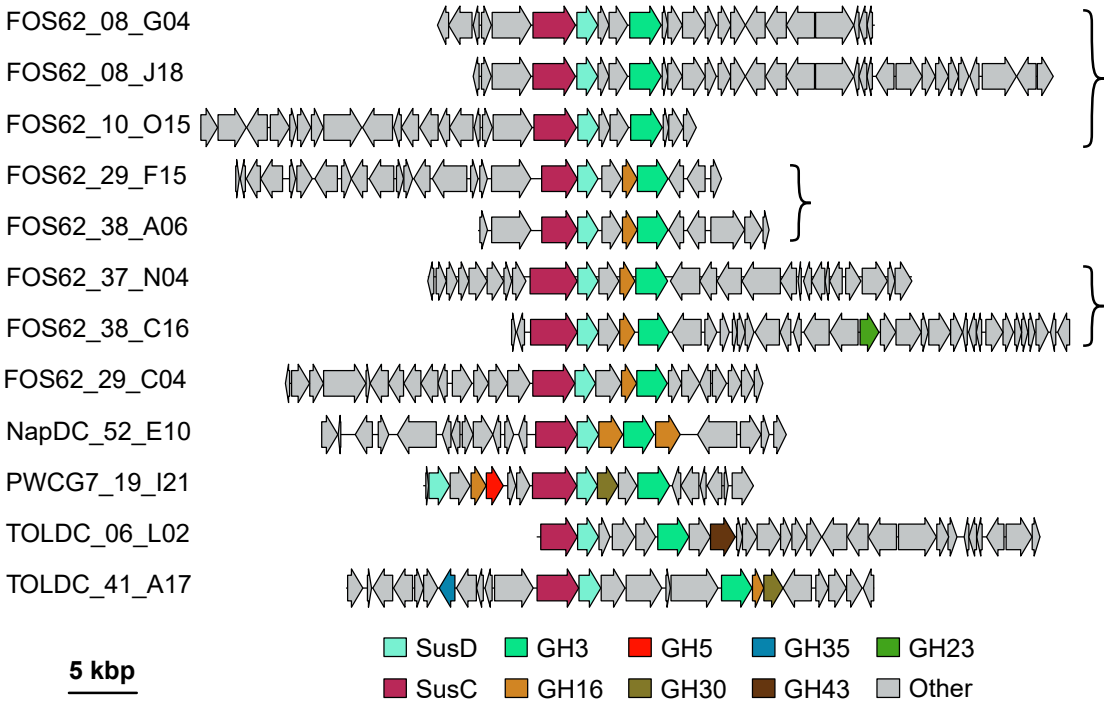

Figure 2.16: *Gene Organisation of SusC/SusD-like Encoding Fosmids.* Putative glycoside hydrolases and SusC/SusD-like proteins are coloured by family. ORFs not annotated as a glycoside hydrolase, SusC-like, or SusD-like, are shown in grey. Fosmids identical to those shown here have been omitted for simplicity. Fosmids have been aligned to highlight synteny. Fosmids pairs or sets within brackets share greater than 95 % identity over their the PUL region.

more GH16s. This motif has been seen previously in the laminarin – a β-glucan containing 1,3- and 1,6-linkages – degrading PUL from the marine bacteria *Gramella forsetii* KT0803 [145] and in the mixed-linkage glucan degrading PUL from *B. ovatus*[287]. This co-occurrence of GH3s and GH16s within the same gene cluster implies that many of these loci target glucans other than cellulose for degradation.

Many fosmids that lack PULs contain clusters of multiple GH genes. Almost 15 % (26 of 178) of the fosmids contained 5 or more GH genes, Figure 2.17. There was one set of two clones (FOS62_37_N12 and FOS62_38_G18) and an additional set of three clones (FOS62_38_D22, FOS62_41_N11 and FOS62_46_E02) with near complete identity over an overlapping region. Additionally, a set of four clones (NO001_07_A13, NO001_01_I19, NA004_04_B18, NR003_09_O07) share between 80 and 90 % identity over a region containing a GH2 and two GH3 genes. Surprisingly, there were a number of clones with gene clusters that appeared to target xylans. These clones contained carbohydrate esterases, GH43s (which have activity towards xylosides and arabinosides), GH67s and GH115s. The last two families (GH67 and GH115) both play a role in the removal of glucuronic acids from glucouronoxylan [196, 210]. Specifically, clones CO182_36_O01, TOLDC_20_J14, FOS62_41_N11, FOS62_46_E02 and NO002_11_N21 all contained either a GH67 or a GH115, while clones FOS62_37_N12, FOS62_38_D22 and FOS62_38_G18 harboured carbohydrate esterases predicted to target acetylations present on xylan. All of these potential xylan targeting fosmids have either a GH3, GH30 or GH10 enzyme present, families which can have members with xylosidase or xylanase activity, and we speculate that these are the proteins which have activity towards MU-C. Future characterization of the PULs and gene clusters has potential to shed light on synergistic mechanisms of degradation occurring within the sampled communities.

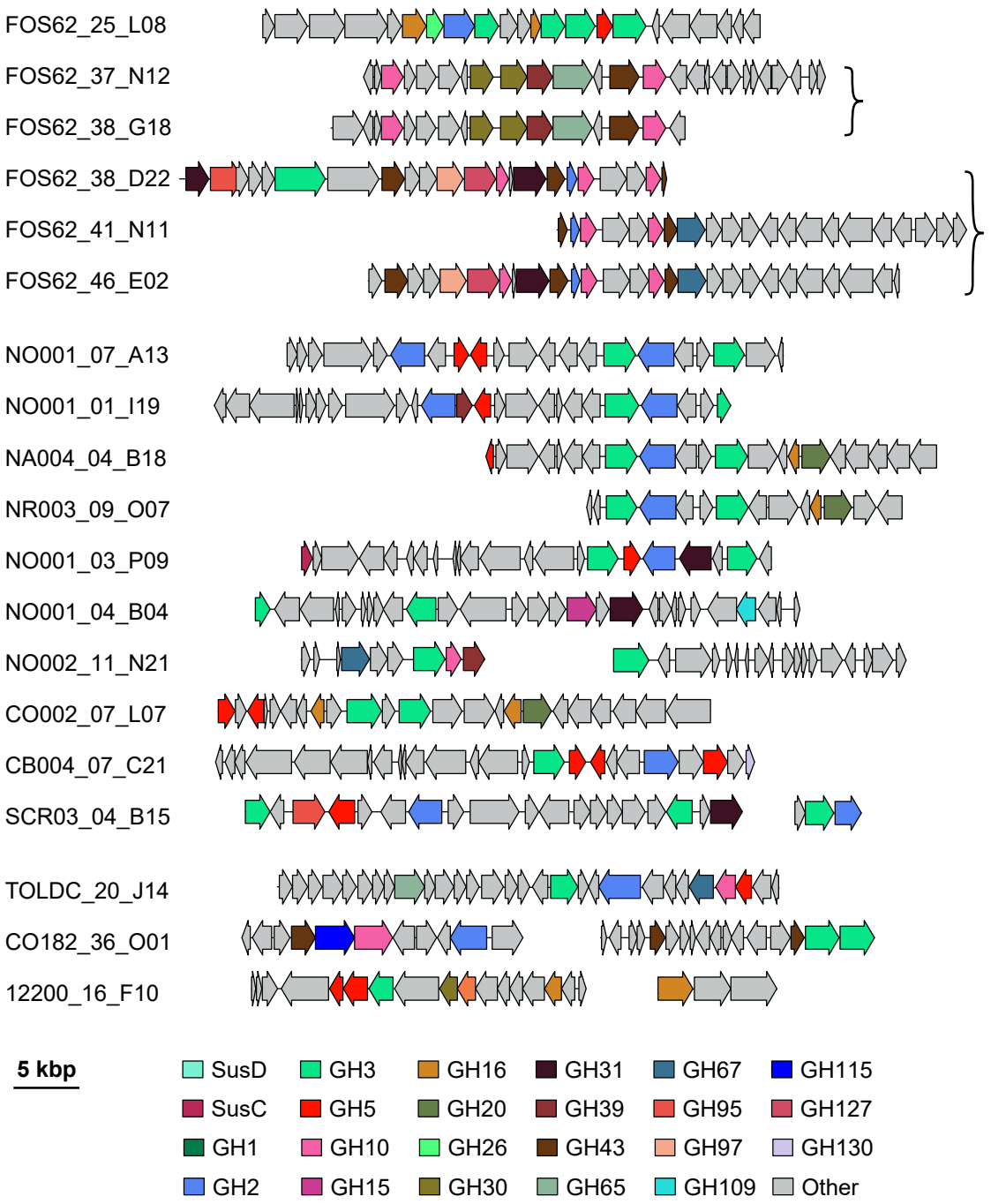

Figure 2.17: *Gene organization of fosmids containing more than 5 GH genes* Putative glycoside hydrolases proteins are coloured by family. ORFs not annotated as a glycoside hydrolase are shown in grey. Fosmids identical to those shown here have been omitted for simplicity. Fosmids have been aligned to highlight synteny. Fosmids pairs or sets within brackets share greater than 95 % identity over their the overlapping region.

2.4 Limitations and Future Directions

The combined use of high-throughput screening and high-throughput characterization of the recovered hits has allowed us rapid access to hundreds of active clones, each with one or more specific activities. This approach allows for the rapid discovery of catalytic function, however as with most high-throughput approaches, compromises have to be made. The first of these lies in the choice of screening host. *E. coli* is a work horse of biotechnology, but it undoubtedly cannot express every protein, especially those requiring extensive post-translational modification. Functional metagenomic screening has been successfully performed using multiple hosts [34], however a commercial system comparable to that used to generate libraries in *E. coli* for other species remains elusive.

The use of a soluble fluorogenic reporter molecule has allowed the rapid screening of hundreds of thousands of clones. However, the choice of reporter molecule certainly introduces bias into the hits recovered. This is evidenced by the comparison of hits recovered from the FOS62 library when screened with DNP-C or MU-C. This has prompted our development of probes with a fluorogenic leaving group with a lower pK_a, the 6-chlorocoumarin containing probes described in Chapter 3. Additionally, the use of glycoside reporters with a reporter functionality at the reducing end may also limit the range of hits discovered. Enzymes that require binding in the +1 subsite for activity may be missed by the use of such substrates. The complementary use of chromogenic hydrogels derived from plant polysaccharides, such as those developed by Willats et. al. [157] may provide further access to such enzymes.

Additional screening conditions would likely reveal a greater number of clones and provide further access to the diversity of degradative enzymes. Assay conditions, such as temperature, pH and concentrations of metal ions and enzyme co-factors undoubtedly affect which hits, and the number of hits, recovered. Re-screening libraries under various conditions would result in a larger number of hits. This being said, it often comes down to a question of resource management; will it be more fruitful to screen the same library with different conditions or a different library with the same conditions? Ultrahigh-throughput technologies such as those utilizing droplet-based microfluidics [63, 127] may enable exploration of larger numbers of screening conditions.

2.5 Conclusions

This study has revealed a diversity of cellobioside-degrading activities from a wide assortment of metagenomes. The coupling of liquid-based high-throughput functional screening, plate-based clone characterization and fosmid sequencing and annotation has allowed us access to the catalytic potential encoded in these metagenomes. This has revealed hundreds of glycoside hydrolases, many of which show low identity to any previously discovered gene. Comparison of these genes sequences to those of characterized glycoside hydrolases, also revealed many genes within clades lacking a characterized representative. The use of large insert libraries also revealed PULs and clusters of multiple GHs, many of which appear to target hemicelluloses. This collection of clones provides a wide range of genes and gene cassettes which may be useful for biomass deconstruction and modification schemes of carbohydrates.

Chapter 3

Functional Screening of the *Castor canadensis* Fecal and Gut Metagenomes

3.1 Summary

Beavers have been described as natures engineers due to their prolific capacity to reshape forest ecosystems into ponds and meadows, raising groundwater levels and creating new habitats for diverse plants and animals. Beyond their capacity to transform landscapes, beavers are extremely efficient consumers of woody biomass relying on bark, shoots, leaves, and other fibers from hardwood deciduous trees as primary nutritional resources. Despite this conspicuous efficiency, the underlying mechanisms enabling beavers to effect woody biomass deconstruction into soluble sugars has remained a mystery. Here we chart the community structure and metabolic problem solving power of the beaver fecal and gut microbiomes using a combination of metagenomic sequencing, functional metagenomics and carbohydrate biochemistry. This has revealed hundreds of functional clones from the beaver fecal and gut microbiomes which are active on model cellulose and xylan substrates. A subset of these fosmids contained GH43 enzymes belonging to previously uncharacterized subfamilies. Cloning and expression revealed the substrate specificity of three previously uncharacterized subfamilies and revealed a mechanism involving two of these family 43 hydrolase domains and an appended family 8 glycoside hydrolase domain which synergistically degrade arabinoxylan oligomers. Fosmid clones recovered in this study also revealed novel assortments of genes clustered into polysaccharide utilization loci (PULs) with the potential to synergistically enhance

biomass deconstruction in support of host nutrition.

3.2 Background

Microbial communities inhabiting the mammalian digestive tract, so-called gut microbiomes, affect host health and mediate essential services including dietary access to recalcitrant glycans or polysaccharides such as starches and fibers [192]. With respect to digestion, the taxonomic composition of these communities correlates with host diet and nutrient acquisition strategies across different mammalian lineages [208]. Multiple studies indicate that mammalian gut microbiomes consist of specialized communities that respond to complex glycans derived from specific dietary sources such as lignocellulosic biomass and release products that can be absorbed into the digestive tract [74, 173, 208, 261, 281]. Consistent with these observations, the digestive tracts of herbivores and wood-feeding (xylotrophic) organisms harbor microbial communities enriched in genes or gene cassettes encoding the corresponding biocatalysts and polysaccharide utilization systems [19, 120, 226, 240, 268, 315]. These communities provide a frame of reference for understanding how lignocellulosic biomass is converted into dietary macronutrients, as well as a deep reservoir of genomic information with potential biotechnological applications [13].

The North American beaver, *Castor canadensis*, provides a useful animal model for the study of xylotrophic microbiomes, as its diet is largely composed of bark, shoots, leaves, and other fibers from hardwood deciduous trees such as poplar, aspen, and cottonwood, which have commercial value in the forestry sector [41, 308]. Hardwoods such as poplar typically have a total polysaccharide content comprising 60-80 % of the dry mass of the wood [319]. Some 35-50 % of this dry mass consists of cellulose followed by hemicellulose (20 % primarily glucuronoxylan) and pectin. Previous studies have shown that beavers are capable of digesting up to 32 % of the available cellulose in consumed hardwoods [70]. However little is known about the utilization of the hemicellulose component by the beaver microbiome, and much less about the enzyme repertoire effecting its deconstruction.

Gruninger and colleagues recently used small subunit ribosomal RNA (SSU rRNA) gene sequencing to profile the microbial community structure of beaver cecum and rectal samples, indicating a typical mammalian hindgut community that is dominated by Bacteroidetes and Firmicutes

[109]. It is also worth noting that a recent paper by Wong and colleagues has also examined the metagenomes of cultures inoculated with beaver droppings and propagated with cellulose or poplar hydrolysate over a three year period[324]. This was done to enrich for species capable of degrading either cellulose or the carbohydrates present in poplar. As the community compositions observed were substantially different from the inoculum (which itself was atypical for a gut sample as it was dominated by Proteobacteria) [323] investigation of beaver microbiome in its native state may reveal further insight.

As a hindgut fermenter, commensal microbes in the lower digestive tract of the beaver are expected to mediate the degradation and fermentation of complex sugars to provide short chain fatty acids that provision host nutrition [217]. Given that the proclivity of beavers to consume wood differentiates them from other hindgut-fermenting herbivores, several questions arise from the initial microbiome study: What is the population structure of the beaver gut microbiome and how does it change throughout the digestive tract? Does the microbiome encode specialized genes or gene cassettes mediating conversion of lignocellulosic biomass? Are some components of the lignocellulose targeted for digestion more than others? Could analysis of these differences reveal new insight into sequential biomass deconstruction of wood-based fibers transferrable to industrial process streams? To begin answering these questions, we used a combination of SSU rRNA gene sequencing, shotgun metagenomics and functional screening to evaluate the community structure and metabolic potential of the beaver microbiome and to recover activities mediating lignocellulosic biomass deconstruction.

3.3 Beaver Fecal Metagenome

3.3.1 16S Ribosomal RNA Profiling

To profile the microbial community composition of the beaver fecal microbiome, I performed 454 pyrotag sequencing of the V6-V8 region of the SSU rRNA gene with three-domain resolution on composite fecal samples from 2 captive beavers. A total of 12,579 rRNA pyrotag sequences, recovered from the fecal sample, were clustered with a 97 % similarity cut-off into 404 operational taxonomic units (OTUs) after singleton removal. Of the OTUs identified, only two could not be

affiliated with described microbial taxa based on comparison to sequences in the Silva database [248]. One of these OTUs showed 99 % identity to the *Castor canadensis* mitochondrial DNA sequence [125], while the other had at most 75 % sequence identity to SSU rRNA sequences from uncultured bacteria. The majority of sequences were affiliated with the bacterial phyla Firmicutes (214 OTUs, 58.4 %) and Bacteroidetes (93 OTUs, 24.4 %), Figure 3.1. Within the Firmicutes, 200 OTUs were affiliated with the class Clostridia (55.9 %), with 143 OTUs (43.6 %) affiliated with the family Lachnospiraceae, which is known to harbor xylanotrophic, butyric acid-producing members [68]. Within the Bacteroidetes, 68 of the 93 OTUs (21.3 %) were affiliated with the class Bacteroidia, with 39 OTUs (15.5 %) affiliated with the uncultured S24-7 group. In a recent culture-dependent study, S24-7 comprised approximately 4 % of the beaver fecal microbiome prior to methanogenic enrichment on different lignocellulosic biomass substrates [323]. Overall, these results are consistent with the observations of Gruninger and colleagues, although the proportions of Firmicutes and Bacteroidetes did vary between the studies [109]. As a high proportion of identified OTUs lacked cultured representatives (354 of 370 bacterial OTUs) specific metabolic roles could not be inferred with confidence. To this end we used shotgun metagenome sequencing to predict metabolic functions encoded in the beaver microbiome.

3.3.2 Metagenome Sequencing

Shotgun metagenomic sequencing was conducted on the 454 platform using DNA from the same preparations used in SSU rRNA gene pyrotag analysis. This resulted in the production of 469.2 million base pairs (Mbp) of total sequence information (616,811 reads with average length 761 bp). Raw sequences were trimmed to Q30 quality score using prinseq lite+ [265] and assembled using MIRA [54] by Dr. Keith Mewis. This resulted in 75,523 contigs with an N50 of 1,787 bp and 130.5 Mbp of consensus sequence. To explore potential bias in community structure based on pyrotag analysis, we examined SSU rRNA gene sequences recovered from the metagenome by comparing unassembled reads to the Silva SSU database using MetaPathways [155]. Of the 616,811 unassembled reads 1,812 were annotated as having SSU rRNA genes. The majority of these sequences were affiliated with either Firmicutes (890) or Bacteriodetes (438), consistent with pyrotag results (Figure 3.1). A notable exception was the relative abundance of Tenericutes within

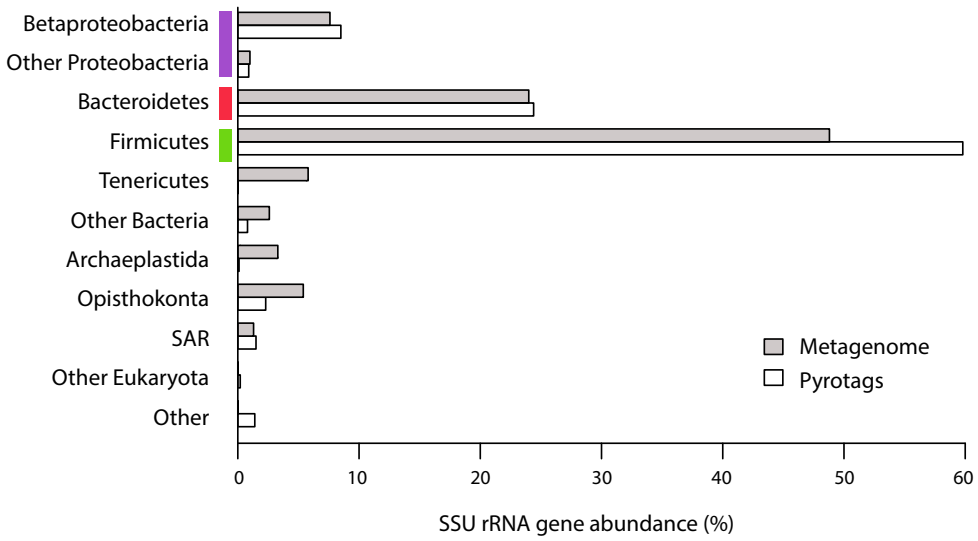

Figure 3.1: *Beaver Fecal Community Composition.* The relative abundance of 16s rDNA genes found in the metagenome are compared to those identified by pyrotags. Both methods reveal a metagenome dominated by Firmicutes (green), Bacteroidetes (red) and Proteobacteria (purple) phyla.

the metagenome, which was greater than that seen in the pyrotag data (percentages of 0.02 % and 5.8 % respectively). This may be due to amplification bias, as has previously been observed for Mycoplasma, the dominant Tenericutes genus identified in the beaver fecal metagenome [159]. In addition to bacteria, we detected SSU rRNA gene sequences affiliated with Archaeplastida (predominant class Liliopsida) and Opisthokonta (predominant class Insecta) at low abundance. The presence of these eukaryotic taxa in the beaver microbiome could reflect captive dietary intake or colonization post defecation.

To investigate the abundance of CAZymes within the fecal metagenome, the unassembled reads were queried against the CAZy database [181] using LAST [150] implemented in MetaPathways [155]. This revealed 28,107 ORFs (3.85 %) annotated as belonging to a CAZy family. GH genes were the most numerous CAZyme category identified, comprising 2.14 % of all annotated ORFs. The five most abundant hydrolase families identified were GH13, GH2, GH3, GH43 and GH31, which were present at 0.35, 0.20, 0.17, 0.12 and 0.07 % of all ORFs, respectively (Figure 3.4, panel B). All of these top families, GH13 excluded, are known to be involved in the degradation of various plant polysaccharides. GH13, the most abundant GH found within the beaver gut is often

associated with α-amylase and α-glucosidase activity. Although starch is present in most plants, as a form of energy storage, the abundance of this family may also be due to the pervasive use of glycogen by bacteria for energy storage [17]. There were a few GH families that were conspicuous in their absence; neither GH6 nor GH12, which are involved in the degradation of cellulose, were found within the fecal metagenome. The absence of these families, however, has been noted in other herbivorous mammals [241]. Futhermore, hierarchical clustering analysis done by Dr. Keith Mewis of CAZyme profiles of mammalian gut microbiomes revealed that the beaver sample clusters well with other herbivores (Figure, Appendix B.1).

3.3.3 Functional Screening

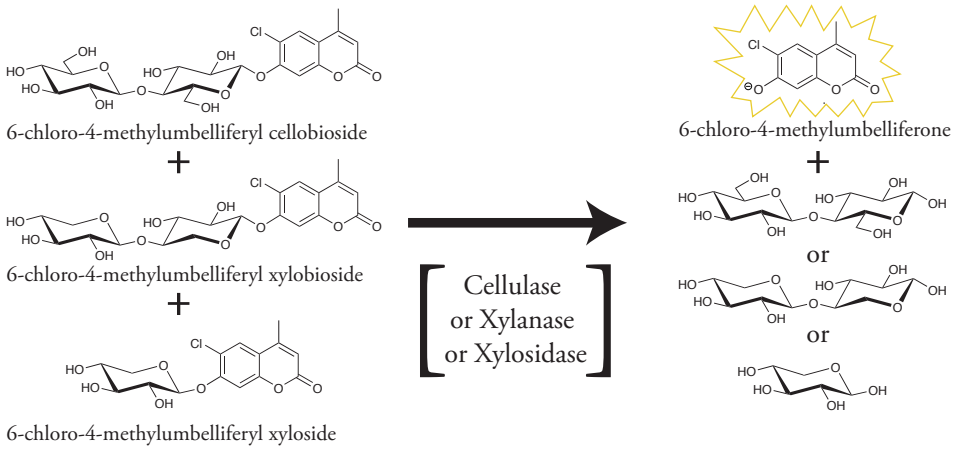

Figure 3.2: *Substrates Used in Multiplex Screening.* 6-Chloro-4-methylumbelliferyl cellobioside (CMU-C), 6-chloro-4-methylumbelliferyl xyloside (CMU-X) and 6-chloro-4-methylumbelliferyl xylobioside (CMU-X2) were used for functional screening. Hydrolysis of these substrates liberates 6-chloro-4-methylumbelliferone which can be detected through fluorescence spectroscopy at a neutral pH.

A fosmid library containing over 4,500 clones was constructed from the same DNA used in shotgun metagenome sequencing using the pCC1 copy control system expressed in *E. coli* EPI300 (Epicentre). Activity assays were based on methods described by Mewis and colleagues [200, 201] but instead of chromogenic substrates we utilised cellobioside, xyloside, and xylobioside fluorogenic substrates bearing the 6-chloro-4-methylumbelliferyl aglycone, resulting in greater sensitivity with improved signal to noise ratio [52]. The incorporation of a chlorine atom lowers the pK_a of the

released aglycone to 5.9 ± 0.1, increasing substrate reactivity when compared to the parent 4-methylumbelliferone ($pK_a = 7.8$), and allowing direct sensitive detection at neutral pH values [52]. This is a further improvement on the screening conditions employed in Chapter 2. We combined the three substrates in a multiplex format to reduce both the time required and materials costs (Figure 3.2). Multiplex screening identified 51 validated fosmids that hydrolyzed at least one of the three fluorogenic substrates (z-score >3); a hit rate of 1.1 % (Figure 3.3).

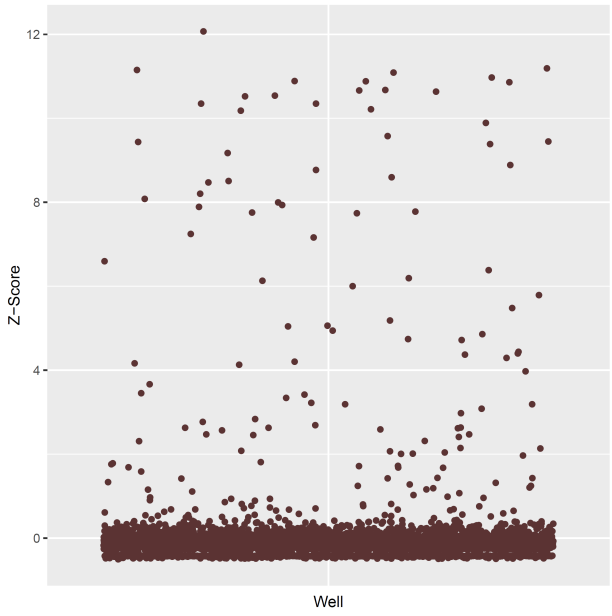

Figure 3.3: *Functional Screening of Beaver Fecal Library.* Z-score values for fluorescence were calculated for each plate. Clones above the z-score threshold of 3 were chosen for further validation.

To further characterize active clones recovered in multiplex screening, initial rates of hydrolysis were assessed against a panel of nine separate fluorogenic substrates by Dr. Feng Liu. A majority of clones were most active against either β-glucosides or β-xylosides. However, six clones displayed higher activities against alternative substrates including arabinosides (06_E19, 09_O03, 09_O15), galactosides (10_J12), lactosides (09_I18), and mannosides (05_B01). This suggests that either the active enzymes encoded on these fosmids possess broad substrate specificities, or that multiple functions are encoded and expressed from individual clones consistent with gene cassettes e.g. cellulosomes or polysaccharide utilization loci (PULs) involved in the extracellular deconstruction of insoluble biomass [91] and the utilization of soluble carbohydrates within the cell respectively

[191].

3.3.4 Fosmid Sequencing and Gene Annotation

To identify individual genes or gene cassettes mediating substrate conversion we fully sequenced the 51 active clones. Reads were assembled using ABySS [275] and ORFs were predicted and annotated using the MetaPathways pipeline [275]. Comparison between sequences identified 38 non-redundant clones based on a threshold of >95 % similarity across >90 % of insert length (Figure 3.4). Additional queries against the CAZy database identified 135 GH genes from 28 GH families encompassing 11.13 % of the annotated ORFs (Figure 3.4). Unexpectedly, active fosmids harbored only 5 GH genes from families with annotated cellobiohydrolases or endoglucanases, one each from GH families 5 and 51, and three from GH8 with none from the cellulolytic GH families 6, 7, 9, 12, 26, 44, 45, 48, 74 or 124. As the metagenome encodes for a total of 1,085 cellulases from 9 of 14 cellulase families (0.15 % of all predicted metagenome ORFs) it may be that we have only captured the most abundant taxa within the fosmid library and that cellulose is being degraded by rarer taxa within the community.

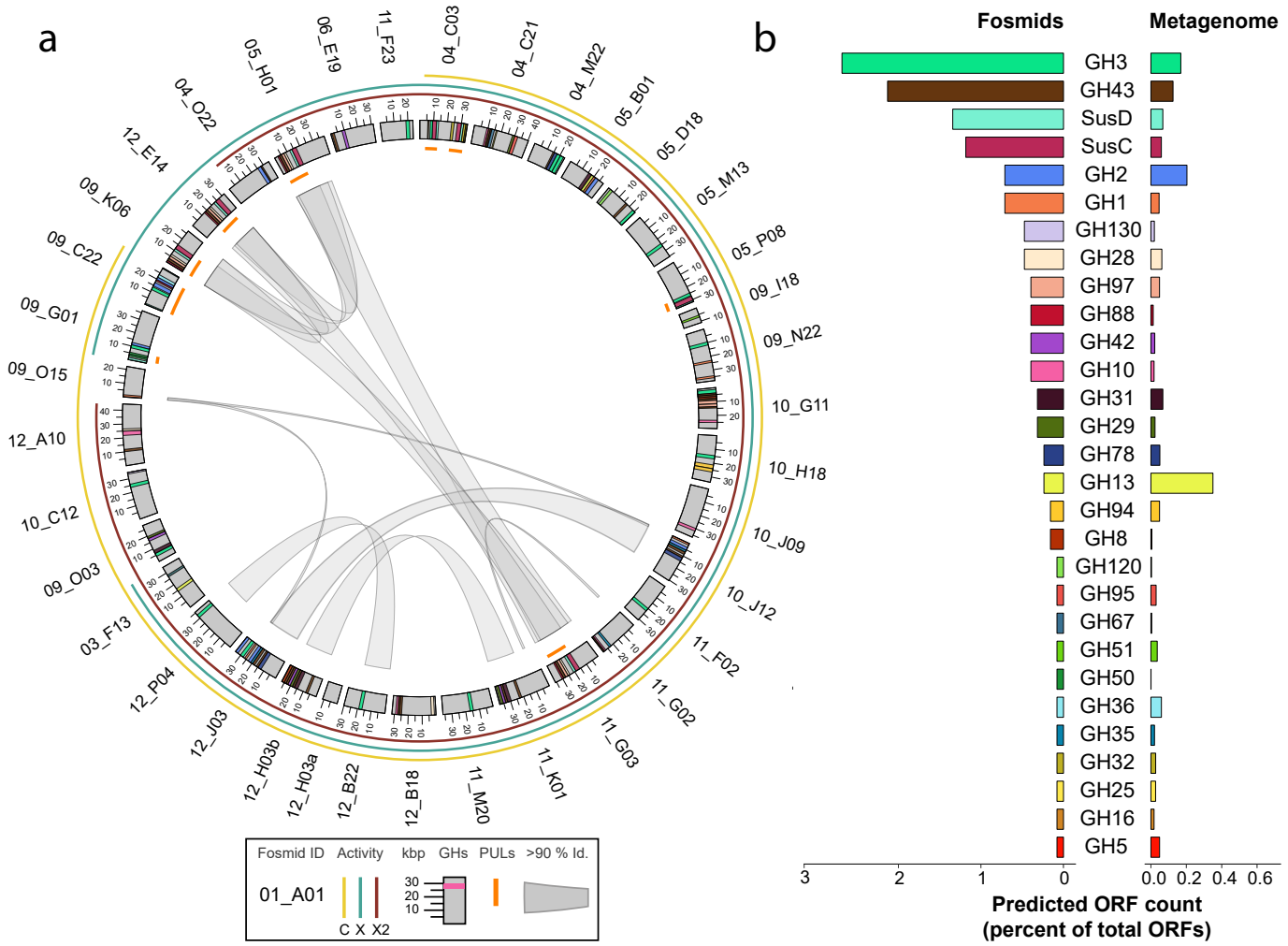

Figure 3.4: *Fosmids Identified from High-Throughput Screening of Fecal Library.* **a** Schematic representing fosmid hits, gene presence and similarity. Grey bars represent each fosmid and are proportional to their length. Fosmids sharing 100% identity with another fosmid were removed. Connections in the center represent areas of 90% or greater nucleotide identity between fosmids. Inner track represents the locations of identified PULs. Outer coloured track represents activities identified for each fosmid from functional screening (C: CMU-cellobiose, X: CMU-xylose, X2: CMU-xylobiose). Coloured bars within each fosmid represent GH domains as predicted by BLASTP against the CAZy database. **b** Histogram displays colour encoding of GH gene families and relative abundance of each family in the complete fosmid dataset compared to the abundance of the same gene families in the unassembled metagenomic dataset. Figure generated by Dr. Keith Mewis.

In contrast, there was an abundance of genes encoding hemicellulose-converting enzymes, particularly xylan. The most abundant GH family recovered was GH3, which contains both β-xylosidases and β-glucosidases. This was followed in abundance by GH43, a family containing both β-xylosidases and xylanases [202]. Within the unassembled metagenome, 912 GH43 genes (0.12 % of all metagenomic genes) from 23 subfamilies were identified. A total of 27 GH43 genes (1.8 % of all fosmid genes) from 10 subfamilies were identified on active clones (Table 3.1). Five GH43 genes identified on fosmids belonged to subfamilies containing no previously characterized members (subfamilies 2, 7, and 28, see Table 3.1) thus were of unknown specificities. Genes encoding xylan side-chain removing enzymes (α-glucuronidase from GH67) were also present. Interestingly, we identified a number of genes encoding multiple GH domains (Figure 3.5). Four of these encoded predicted endo-acting and exo-acting domains. These include 04_C21-10 and 12_B18-19 which contain both GH43 and GH10 domains and are likely involved in xylan degradation, as well as 10_G11-03 and 12_H03-13, which both contain GH43 and GH8 domains consistent with a role in xyloglucan or xylan degradation. The presence of both endo- and exo-glycosidase domains within the same protein can lead to synergism in efficient deconstruction of these xylan substrates, as has been observed previously for other polysaccharides [36].

Table 3.1: GH43 Subfamilies Identified on Functionally Active Fosmids.

GH43 Subfamily	ORFs
1	04_C21-11, 11_G02-25
2	10_G11-3, 12_H03-13
7	12_H03-12
10	05_H01-3, 09_K06-1, 11_G03-16, 12_E14-11
11	05_D18-17, 06_E19-1, 12_B18-18
12	04_C21-10, 11_K01-12, 12_A10-9, 12_B18-19, 12_H03-3
19	04_O22-25
24	04_M22-11, 10_J12-7, 12_J03-15
28	10_J12-4, 12_J03-18
29	10_G11-4, 10_G11-8, 11_K01-19, 12_H03-10

3.3.5 Gene Characterization

To better understand the substrate specificities and activities of the enzymes present, we focused our attention on the uncharacterized GH43 subfamilies identified in functional screening. To this end

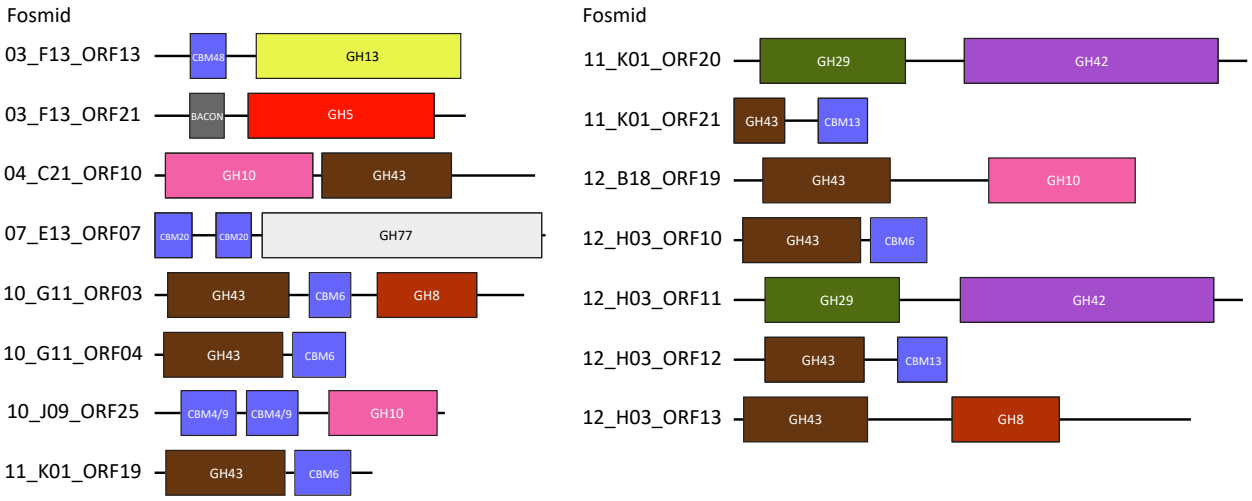

Figure 3.5: *Gene Organization of Multi-Domain Proteins Identified on Functional Fosmids.* Proteins containing more than one domain with a CAZy annotation are shown. The colouring of domains is consistent with Figure 3.4

we generated constructs that were used to overexpress and purify recombinant proteins (12_H03-13 from subfamily 2, 12_H03-12 from subfamily 7, and 12_J03-18 from subfamily 28). Since 12_H03-13, also contained a GH8 domain, we created two additional constructs in which the GH8 and GH43 domains were inactivated independently by mutation of the catalytic acid residue (GH8 domain variant H03-13_E507A and GH43 domain variant H03-13_E209A). Of the three wild-type enzymes, two had detectable cleavage activity on aryl-monosaccharides (Table 3.3); both 12_H03-13 and 12_J03-18 cleaved CMU-xyloside. The specificity constant of the 12_H03-13 wild-type enzyme was the same, within error, as that of the H03_E507A variant in which the GH8 activity was abated (Table 3.2), indicating that the GH43 domain is responsible for the hydrolysis of CMU-X. Surprisingly, none of the enzymes cleaved any of the other aryl glycosides tested (Table 3.3), reinforcing the utility of the inherently more reactive chlorocoumarin glycosides for detection of previously unknown activities.

Table 3.2: Kinetic Rates Determined for Purified GH43 Enzymes with CMU-X.

Enzyme	K_M (mM)	$k_{cat}(s^{-1})$	k_{cat}/K_M $(s^{-1}mM^{-1})$
12_J03-18	0.48 ± 0.06	0.22 ± 0.02	0.45 ± 0.07
12_H03-13 WT	0.19 ± 0.03	0.80 ± 0.06	4.2 ± 0.7
12_H03-13_E507A	0.14 ± 0.01	0.73 ± 0.02	5.2 ± 0.4

Table 3.3: Activity of Purified GH43 Enzymes on Aryl-glycosides.

Substrate	12_J03-18	12_H03-12	12_H03-13 WT	12_H03-13 E507A	12_H03-13 E209A
pNP-X	ND	ND	ND	ND	ND
MU-X	ND	ND	ND	ND	ND
CMU-X	Yes	ND	Yes	Yes	ND
pNP-Ara	ND	ND	ND	ND	ND
MU-Ara	ND	ND	ND	ND	ND

ND : None Detected

The activities of enzymes 12_H03-12, 12_H03-13 and its variants were also tested on a set of arabinoxylan oligosaccharides. This revealed a synergistic degradation mechanism in which the GH43 domain of 12_H03-13 (subfamily 2) releases undecorated xylose from the non-reducing end of the oligosaccharides while the GH8 domain of 12_H03-13 (a reducing end xylose-releasing exo-oligoxylanase [Rex]) releases xylose from the reducing end of decorated oligosaccharides (Figure 3.6). The activity displayed by 12_H03-13 is further complemented by GH43 12_H03-12 (subfamily 7) which cleaves arabinose decorations from arabinoxylans, releasing arabinose and xylobiose, an activity which is only observed in the presence of 12_H03-13. This establishes the intriguing possibility that 12_H03-12 is activated by 12_H03-13 which should be the subject of future work. The xylobiose generated by these two enzymes appears to be resistant to further degradation. As GH8 Rex genes typically require at least a trimer for activity this domain is not expected to hydrolyse xylobiose [124, 163]. The GH43 domain of 12_H03-13 was expected to further degrade xylobiose, yet this is not the case, suggesting that the presence of an arabinose sidechain may be important for the xylosidase activity of this domain. This represents, to my knowledge, the first multi-domain protein containing both a GH43 and GH8 domain to be characterized and the first description of how these two domains function synergistically on arabinoxylan oligosaccharides converting them into arabinose and xylobiose. Collectively these results illuminate the substrate specificity and activity of GH43 subfamilies 2, 7 and 28 within the context of the beaver fecal microbiome with direct relevance to lignocellulosic biomass conversion and host nutrition.

3.3.6 Presence of Hemicellulose Targeting Loci

In addition to providing a route toward functional validation of predicted GH genes, active clone sequences contained information about the structural organization of GH gene cassettes. This has

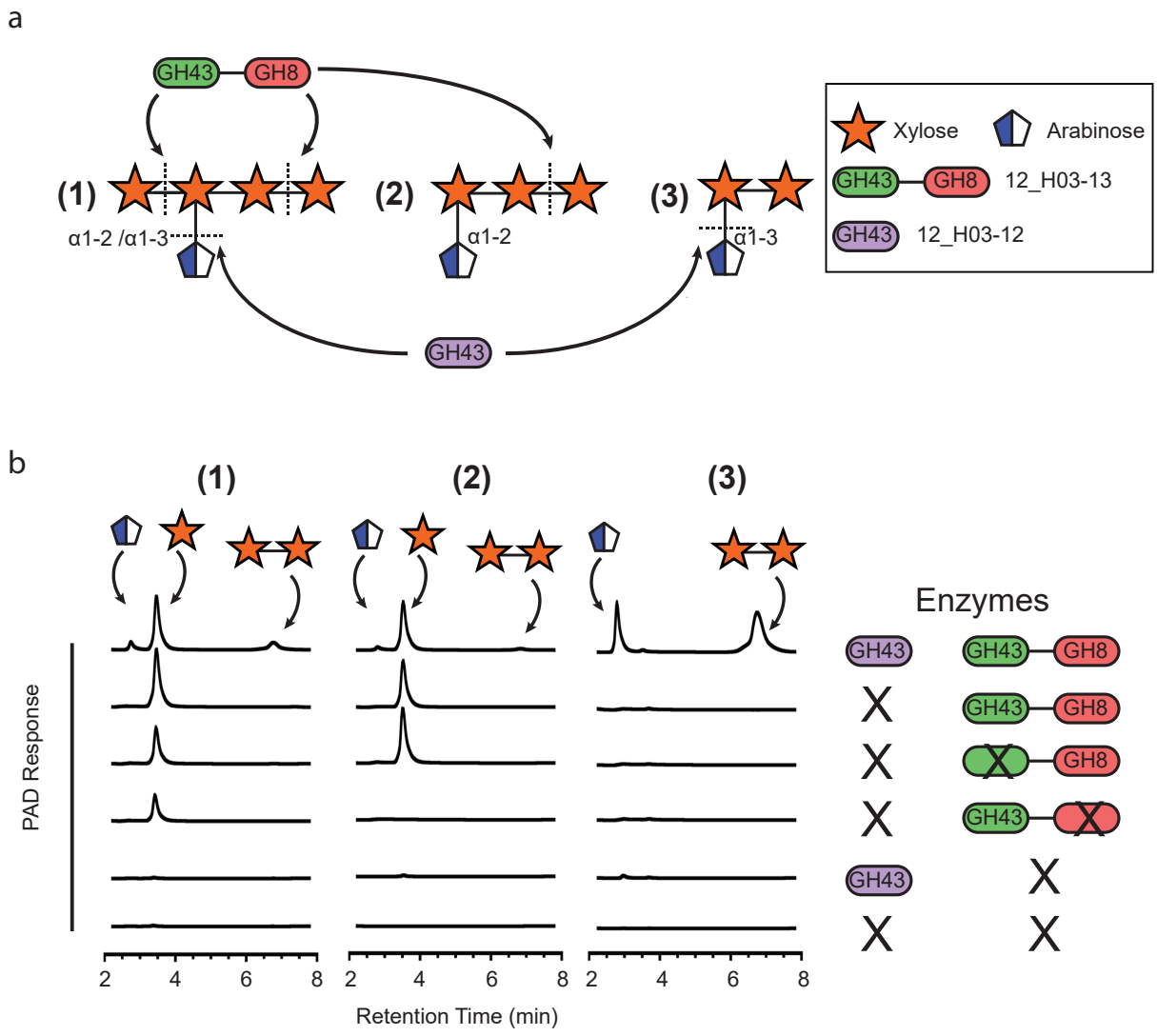

Figure 3.6: *Synergistic Degradation of Arabinoxylooligosaccharides by H03-13 GH43 Enzymes.* **A.** Schematic of the activities of the individual domains of 12_H03-13 and 12_H03-12 on arabinoxylans. These two enzymes were tested for activity on mixture of 2^3-α-L-arabinofuranosyl-xylotetraose and 3^3-α-L-arabinofuranosyl-xylotetraose **(1)**, 2^3-α-L-arabinofuranosyl-xylotriose **(2)** and 3^2-α-L-arabinofuranosyl-xylobiose **(3)**. The GH8 domain of 12_H03-13 releases xylose from the reducing end of **(1)** and **(2)**. The GH43 domain of 12_H03-13 releases xylose from the non-reducing end of **(1)**. 12-H03-12, a GH43 belonging to subfamily 7 is able to release arabinose from the oligomers containing an arabinose α-1,3-linkage. **(B.)** High performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) analysis of the degradation of **(1)**, **(2)** and **(3)**) catalyzed by H03-12, H03-13, H03-13_E209A (GH43 domain mutant, denoted with an X on the GH43 domain) or H03-13_E507A (GH8 domain mutant, denoted with an X on the GH8 domain) and their combinations.

revealed several GH gene clusters that appear to target plant cell wall hemicelluloses (Figure 3.7a). The fosmid 04_C03 contains a motif (GH16 and GH3 adjacent to SusC and SusD-like proteins) which has synteny with a PUL recently shown to be active against mixed-linkage glucans [287]. Several fosmids (04_C21, 10_G11, 11_G02, 12_H03 and 11_K01) also appear to target xylans. The four fosmids 04_C21, 11_G02, 12_H03 and 11_K01 all harbor GH10 genes, which often act as endo-xylanases, and GH43s which may have exo-xylanase activity. Furthermore, Fosmid 04_C21 contains a motif (GH10-GH43 [subfamily 12] protein followed by an additional GH43 [subfamily 1] and a GH67) which has synteny with a gene cluster identified in *Bacteroides intestinalis* [311]. The GH10-GH43 homolog from *B.intestinalis* has endo-xylanase and arabinofuranosidase activity, which is able to release xylose, xylosoligosaccharides and arabinose from arabinoxylans [311]. Although Fosmid 10_G11 lacks a GH10 it does possess a two domain GH43-GH8 gene, which we speculate may have similar activity to target arabinoxylanoligomers as H03-13. The presence of a GH29 (a family with α-L-fucosidases), GH42 (a family with β-galactosidases) and GH31 (a family with α-xylosidases) on the three fosmids 09_O03, 12_H03 and 11_K01 leads us to speculate that these fosmids may target fucogalactoxyloglucan which is present in most dicots and gymnosperms [130, 233].

Moreover, we identified 15 fosmids spanning 5 identity groups containing Sus-like genes (SusC or SusD), leading indicators for the identification of PULs using an automated PUL prediction tool [295] (Figure 3.7b). Eight of these clones exhibited near complete nucleotide identity (05_O06, 05_O07d, 05_O08, 05_P05, 05_P06, 05_P07, 05_P08, and 05_P12) and 4 clones shared near complete nucleotide identity specifically in the PUL interval (12_E14, 11_G03, 05_H01, and 09_K06). The remaining 4 clones (04_C03, 09_C22 and 09_G01) contained unique PUL intervals. Representative PULs from each identity group were compared to the RefSeq database to see if they are also found in sequenced microbial genomes. PULs from 09_C22 and 09_G01 exhibited 99 % and 98 % nucleotide identity respectively to distinct regions of the *Alistipes senegalensis* JC50 genome whereas the most common PUL represented by 05_P08 exhibited 99 % nucleotide identity to the genome of *Alistipes finegoldii* DSM17242. The remaining identity groups exhibited less than 7 % nucleotide identity to reference genomes, indicating previously unrecognized architectures. In addition to the fosmids which appear to target hemicelluloses mentioned above, 05_H01 and homologs appear to target pectic polymers as they contain a GH28 (a family containing polygalacturonases) and a

GH88 which may target the unsaturated reducing ends generated by pectate lyases. The substrates targeted by the PULs present on fosmids 05_P08, 09_G01 and 09_C22 are not immediately apparent, and further biochemical characterization will be needed to reveal their activity.

Taken together, the PULs and gene clusters identified on fosmids appear to target many of the hemicellulosic components of plant cell wall, including glucuronylxylan, xyloglucan and pectins, which would be present in hardwoods. Some of the polymers which these gene cassettes likely act on, however, are not present in hardwoods, such as mixed linkage glucans, which are mainly found in grasses, and arabinoxylans, which are present in grasses and softwoods. The ability to degrade these polymers may provision for host nutrition when a preferential food source is scarce. Future characterization of these PULs has the potential to shed light on combinatorial biomass deconstruction within the beaver microbiome.

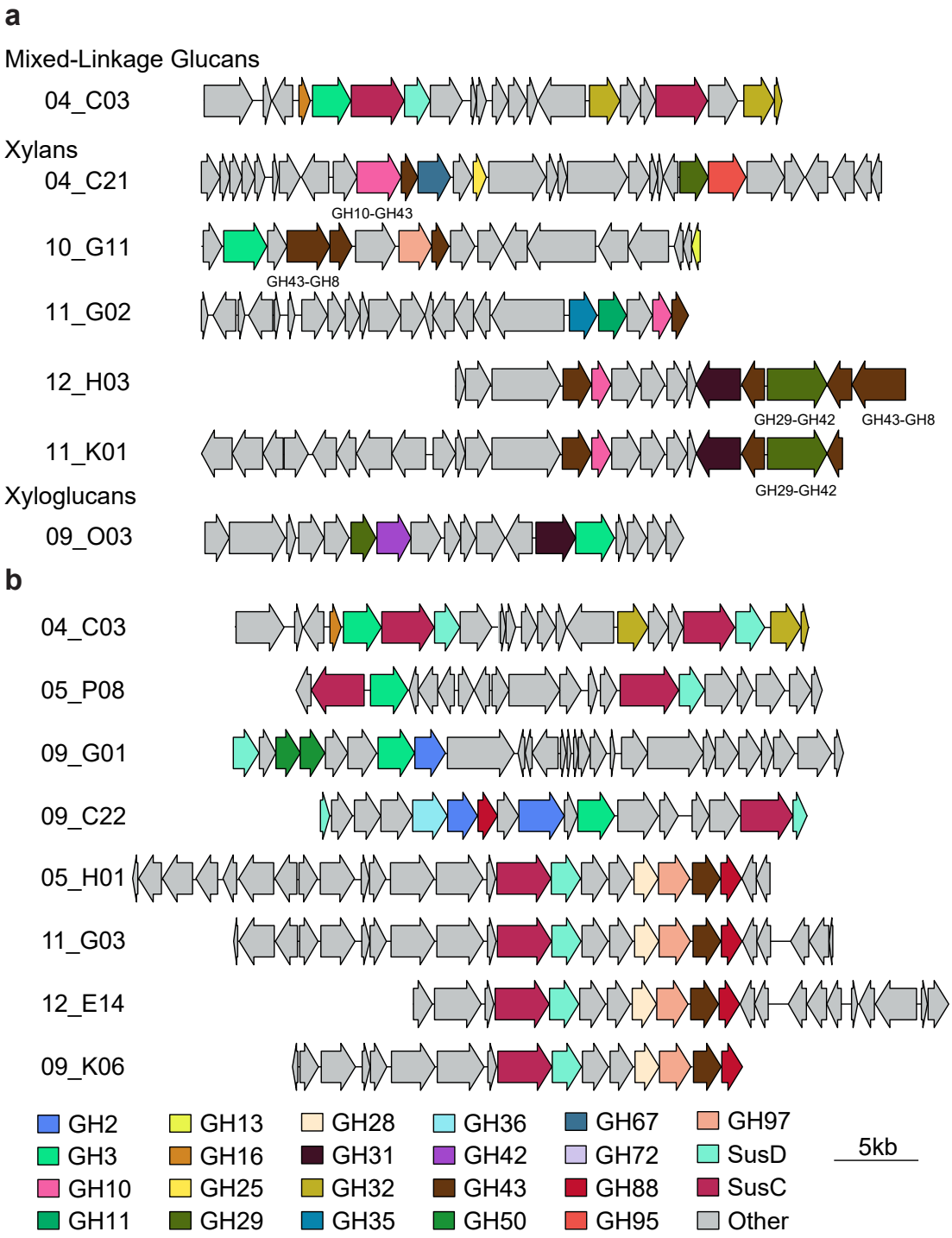

Figure 3.7: *Gene Organization of Putative Hemicellulose Targeting Fosmids and SusC/SusD-like Encoding Fosmids.* **A** Fosmids with gene clusters that may target the hemicellulosic portion of plant biomass within the beaver diet. **B** SusC/SusD-like encoding fosmids. Putative glycoside hydrolases and SusC/SusD-like proteins are coloured with the same scheme as Figure 3.4. ORFs not annotated as a glycoside hydrolase, SusC-like, or SusD-like, are shown in grey. Fosmids Identical to 5_P08 have been omitted for simplicity.

3.4 Beaver Gut Metagenome

Investigation of the beaver fecal metagenome left several unanswered questions that we intended to address by applying similar methods to samples taken along the beaver digestive tract. We hoped to determine whether the microbes and genes found within the fecal sample were reflective of those found within the internal digestive compartments of the beaver. We also aimed to gain insight into the variability of the gut microbiome, both along the gut transect and between individual beaver. To approach these goals, six beaver were dissected and chyme (partially digested food matter) and feces were collected from five sites within the digestive tract (Figure 3.8). These samples were then subjected to the same interrogative methods used for the feces: 16s rRNA sequencing shotgun metagenome sequencing, and high-throughput functional screening of fosmid libraries.

3.4.1 16S Ribosomal RNA Profiling

To ascertain the microbial community structure throughout the beaver gut the V6-V8 region of the SSU rRNA gene was subjected to 454 pyrotag sequencing. Primers with the same sequence (excluding the bar-coding region) as those used for the beaver fecal DNA were employed to facilitate comparison. This revealed 142,453 rRNA pyrotag sequences, after quality control and singleton removal, which were clustered with a 97 % similarity cut-off into 1,115 OTUs, see Table 3.4. Of the OTUs identified from the gut sequences, 1009 were bacterial, 12 were archael and 93 were eukaryotic. The majority of eukaryotic sequences were either mammalian, likely host associated (29,639 of 142,453, 20.8%), or from likely food sources (6,199 of 142,453, 4.4 %), though there were also sequences belonging to the digestive parasite class Trematoda, also known as flukes, found in the small intestine of beaver 3 and cecum of beaver 3, 4 and 5 (335 of 142,453, 0.2 %). The stomach and small intestine sequences are particularly dominated by eukaryotic sequences with the stomach averaging 64.3 $\pm$ 38 % and the small intestine averaging 65.5 $\pm$ 35 % assigned to an eukaryotic OTU, reflecting both the decreased concentration of bacterial cells and the increase in partially digested plant matter, Figure 3.9. There was a surprisingly large variability in the ratio of eukaryotic:bacterial OTUs within the stomach and small intestine, which I speculate could be a result of fecal matter present in the stomach due to coprophagy, which is exhibited by beaver [38].

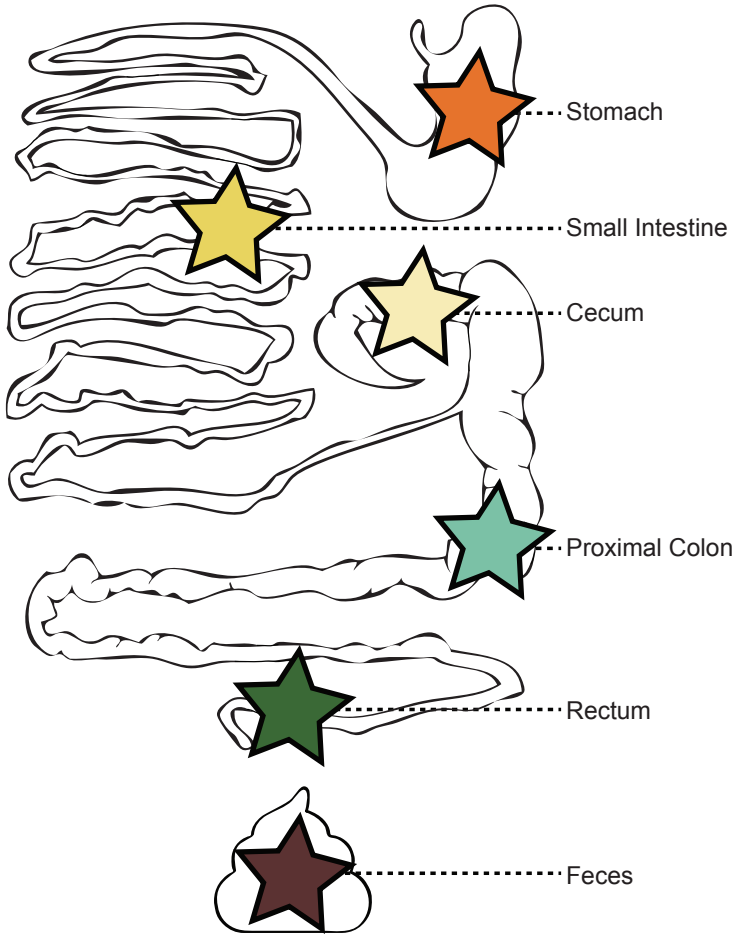

Figure 3.8: *Beaver Gut Sampling Sites.* Stars denote the location of sampling sites throughout the digestive tract. Figure adapted from Vispo and Hume [308].

The large intestine, which is composed of the cecum, proximal colon and rectum, is dominated by Firmicutes and Bacteroidetes, Figure 3.9. These two phyla account for 88 ± 13 % of all phyla within these compartments. One outlier from this was the beaver 3 proximal colon sample, which had substantial fusobacterial counts. The most abundant Firmicutes family was, as seen for the feces, Lachnospiraceae which comprised 30 ± 11 % of all counts within the cecum, proximal colon and rectum samples. This proportion of Lachnospiraceae is somewhat decreased relative to the fecal sample (43.6 %), however it is consistent with the relative proportions seen by Gruninger et. al. [109] (cecal samples : 25.4 %, rectal samples : 28.3 %). The most abundant Bacteroidetes families within the cecum, proximal colon and rectum were Bacteroidaceae and S24-7 (20.4 ± 8.3

% and 12.8 ± 4.4 % of all counts, respectively). While the S24-7 family was seen at similar levels in the fecal sample, the relative number of counts for the Bacteroidaceae family was greatly increased in almost all of the gut cecum, proximal colon and rectum samples (20.4 ± 8.3 %) when compared to the fecal sample (3.9 %).

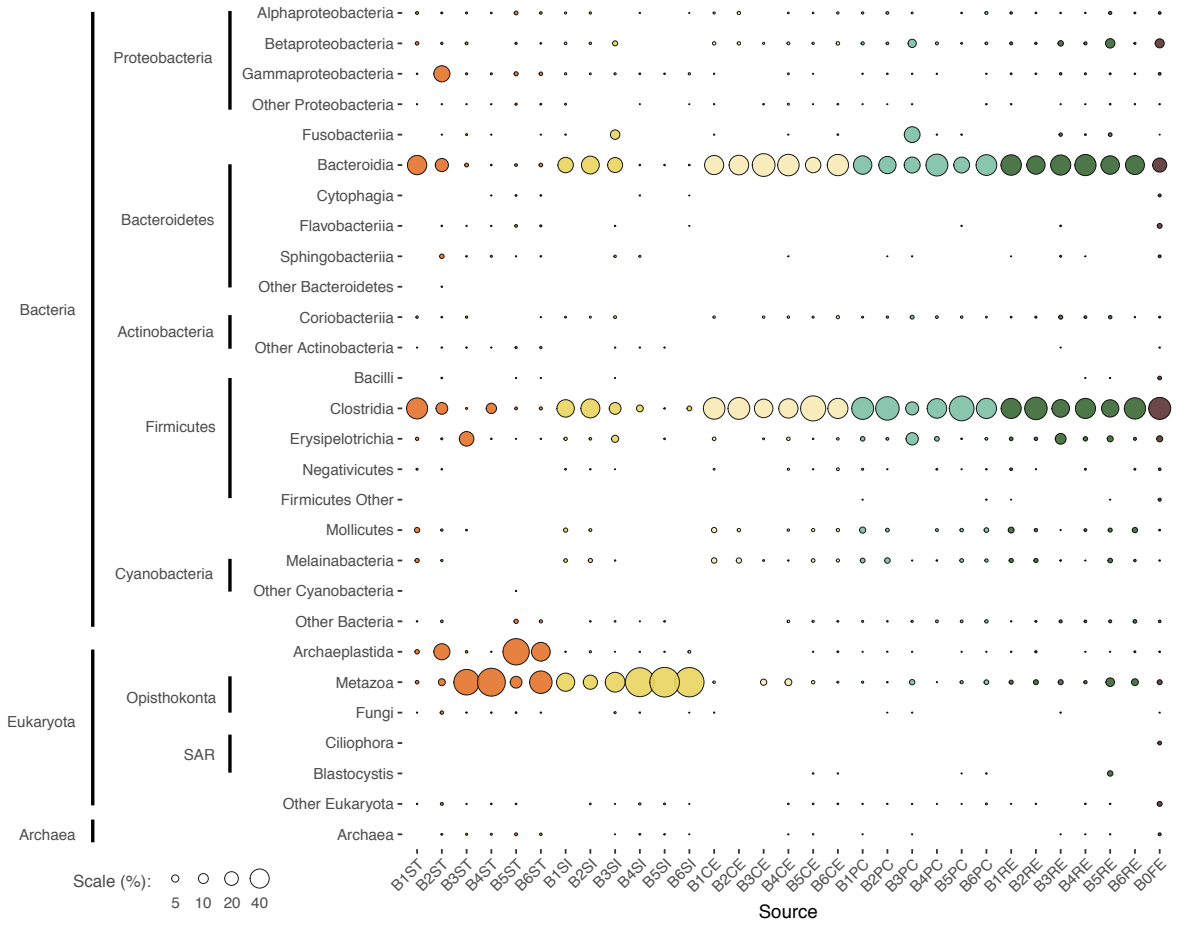

Figure 3.9: *Bubble Plot of Beaver Gut Pyrotags.* Phyla representing greater than 0.5 % in at least one sample; all other taxa are binned into a higher taxonomic group or other categories. 100 % of the total pyrotags clustered at 97 % in OTUs are represented in this plot. Sample names are abbreviated as follows: ST: Stomach, SI: Small Intestine, CE: Cecum, PC: Proximal Colon, RE: Rectum, FE: Feces. Bubbles are coloured by sample source.

To investigate the similarity of samples, hierarchical clustering of the pyrotag counts was performed with all OTU counts, Figure 3.10. As the Beaver 2 cecum sample had a much lower count number than any other sample (n = 241) this sample was excluded from the analysis. Clustering was performed using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) [278]

with dissimilarity between samples calculated using the Bray-Curtis statistic [22]. Clustering revealed, firstly, that the fecal sample is quite distinct, and clusters distinctly from all other samples. Furthermore, the majority of small intestine and stomach samples form a cluster, while the sites forming the large intestine (cecum, proximal colon and rectum) form another. Within the large intestine cluster, the samples also cluster more frequently by organism, rather than by chamber. This is not entirely surprising when taken in the context of other microbiome studies, which have seen large interanimal/personal variation in the species-level abundance of phylotypes in the gut microbiota [21, 67].

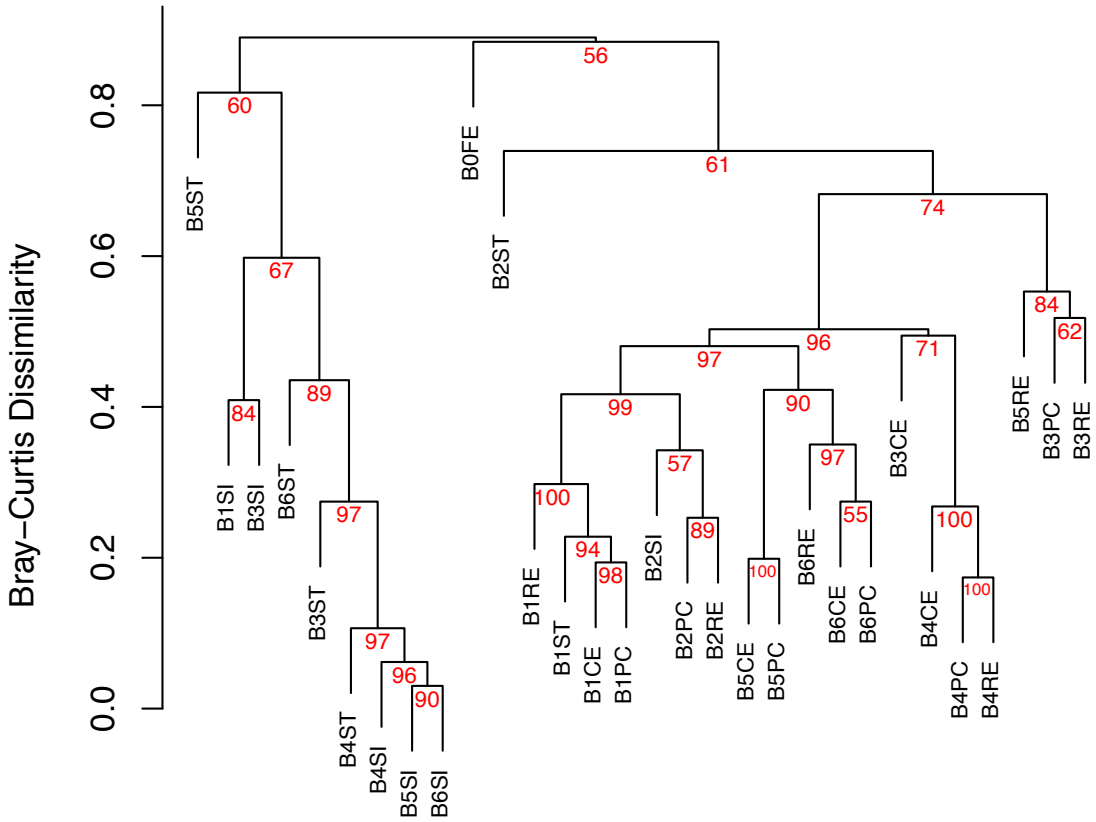

Figure 3.10: *Hierarchical Clustering Analysis of Gut Pyrotags.* Hierarchical cluster analysis of pyrotags (V6-V8 SSU rRNA) from table of OTU counts. A Bray-Curtis distance matrix was used to determine dissimilarity. Approximately unbiased p-values for each branch in the dendrogram were determined through bootstrap resampling (n= 10,000). Sample names are abbreviated as follows: ST: Stomach, SI: Small Intestine, CE: Cecum, PC: Proximal Colon, RE: Rectum, FE: Feces. Beaver 2 Cecum sample is excluded due to low total counts.

At least one third of the bacterial OTUs present in the fecal sample were found in at least one of the other samples, however, the fecal sample was a clear outlier. The fecal sample had the highest number of unique OTUs (n = 204) and this sample was distinct in the clustering analysis. The post-defecation colonization of the fecal sample may be, in part, to responsible for this distinctness. However, there are many other variables which are likely to have an effect on the microbial community present in the fecal sample. All of the gut samples were taken from wild beaver, for which food choice is dictated by self-grazing on local species. This is contrasted with the beaver fecal sample which was taken from beaver who were being supplied by humans with plant material, rather than by foraging. The discrepancy between the plant species chosen by beaver and by humans, may account for some of the gut community variability. Additionally, the close proximity of other mammalian species, including raccoons, deer, and otters, may have had an effect on the community composition, possibly through cross-species microbial exchange, such as that seen by Song et. al. [280]. Moreover, captivity has been observed to alter the microbial community of several mammalian species [59, 77, 197] and may have had an effect in this case also.

3.4.2 Metagenome Sequencing

To gain insight into the genetic potential encoded within the beaver gut we sought to sequence extracted DNA from all five compartments from three beaver. Extracted DNA, the same as used for pyrotag analysis, was sequenced by Dr. Keith Mewis on the Miseq platform (Illumina) using individual barcodes for each sample. This resulted in 4.1 GB of raw sequence data, which was trimmed to Q30 quality score using prinseq lite+ [265] by Dr. Keith Mewis. Unassembled sequence data was used for further analysis as attempts to assemble the sequences proved ultimately unsuccessful. ORF prediction was performed as for the fecal sample and resulted in a total 4,910,871 ORFs, an average of 350,777 $\pm$ 213,140 per sample.

We next sought to investigate the genetic capacity of the beaver gut microbiome to degrade the complex polysaccharides. To this end, the predicted ORFs were annotated as for the fecal sample using the Metapathways pipeline [155]. Of the 4,910,871 predicted ORFs, 156,933 (3.2%) were annotated as CAZymes, Table 3.5. The relative abundance of GHs, CEs and PLs were all significantly increased (p-value = 0.006, 0.031 and 0.011 respectively) in the large intestine samples

Table 3.4: OTU Counts from Beaver Fecal and Gut Samples

Site	Beaver	All OTU Counts	All OTUs	Bacterial OTU Counts	Bacterial OTUs	Unique OTUs	Unique Bacterial OTUs
Stomach	1	3659	431	3560	421	4	3
Stomach	2	3594	288	2350	265	23	18
Stomach	3	3580	53	931	41	8	6
Stomach	4	3564	36	424	23	7	5
Stomach	5	4482	120	305	95	85	62
Stomach	6	3927	104	195	82	27	21
Small Intestine	1	2115	299	1349	295	3	3
Small Intestine	2	5178	357	4038	349	4	3
Small Intestine	3	7718	272	4378	264	6	6
Small Intestine	4	4339	27	218	14	4	4
Small Intestine	5	2620	21	13	10	6	5
Small Intestine	6	8066	31	198	24	3	2
Cecum	1	2593	370	2582	368	3	3
Cecum	2	241	105	241	105	0	0
Cecum	3	1264	178	1218	176	2	2
Cecum	4	4818	281	4596	276	3	0
Cecum	5	5050	392	4990	379	6	5
Cecum	6	2028	276	2018	271	0	0
Proximal Colon	1	5222	478	5205	473	7	7
Proximal Colon	2	16367	512	16345	505	13	12
Proximal Colon	3	3163	165	3078	159	3	3
Proximal Colon	4	4300	280	4298	278	3	2
Proximal Colon	5	4290	372	4255	367	4	4
Proximal Colon	6	4156	370	4063	362	5	4
Rectum	1	15015	588	14799	582	24	23
Rectum	2	3690	344	3598	340	3	2
Rectum	3	4033	271	3940	266	3	3
Rectum	4	4130	294	4092	289	3	3
Rectum	5	5412	373	4855	366	7	7
Rectum	6	3839	319	3657	314	4	4
Feces	0	11575	355	10930	318	259	204
Total Counts	-	154028	1374	116713	1213	532	426
Mean	-	4969	332	3871	326	17	14

when compared to the stomach and small intestine samples.

To further compare the presence of CAZymes within the different compartments of the beaver gut hierarchical clustering of gene abundance was performed. The relative family abundance for the fecal sample was also included in this analysis as a point of comparison. As we were focused on the presence of genes responsible for the degradation of polysaccharides, only the CE, PL and GH

Table 3.5: CAZyme Relative Abundance (% of All ORFs) in Beaver Gut Samples

Site	Beaver	AAs	GHs	GTs	CBMs	CEs	PLs	Total
Stomach	1	0.01	1.61	0.95	0.09	0.15	0.07	2.87
	2	0.32	1.36	1.84	0.11	0.20	0.05	3.88
	3	0.00	0.31	0.19	0.01	0.03	0.01	0.54
Small Intestine	1	0.00	0.87	0.58	0.04	0.11	0.04	1.64
	2	0.00	0.28	0.20	0.01	0.03	0.02	0.55
	3	0.00	0.08	0.10	0.01	0.00	0.00	0.19
Cecum	1	0.00	2.16	1.22	0.09	0.20	0.09	3.76
	2	0.00	2.99	1.49	0.10	0.28	0.14	5.01
	3	0.00	2.10	1.01	0.09	0.20	0.10	3.50
Proximal Colon	1	0.00	2.17	1.29	0.09	0.19	0.08	3.83
	2	0.00	2.74	1.35	0.10	0.27	0.13	4.59
	3	0.00	0.62	0.49	0.03	0.07	0.02	1.23
Rectum	1	0.00	1.95	1.14	0.09	0.18	0.07	3.44
	2	0.00	1.32	0.71	0.06	0.13	0.06	2.28
	3	0.00	1.33	0.78	0.06	0.13	0.04	2.34

families were used for clustering. The results of hierearchichal clustering (Figure 3.11), show, as for the OTU clustering, a split between the stomach and small intestine samples and the samples taken from the large intestine. Furthermore, the fecal sample clustered with the large intestine group (cecum, proximal colon and rectum). This suggests that although the community composition of the fecal sample was substantially different from the gut samples, the carbohydrate degradative capabilities encoded by these communities is similar. The Beaver 3 proximal colon sample was a clear outlier from the observed separation of the large intestine samples from the stomach and small intestine samples. This sample was also distinct in the pyrotag analysis, as it contained a higher proportion of Fusobacteria and Erysipelotrichia counts and was reduced in the phyla Clostridia and Bacteroidia. This intimates that the presence of Fusobacteria, Erysipelotrichia has shifted the polysaccharide degrading capacity of the sample.

To further investigate the degradative potential of the fecal samples we examined the specific families known to be responsible for plant polysaccharide degradation, Figure 3.12. The most abundant families across all samples were GH43, GH2, GH3, and GH5 (0.097, 0.084, 0.087, 0.054 % of all ORFs, respectively). These families are all able to catalyse the degradation of a number of different components of holocellulose, Figure 3.12, suggesting that their proliferation is a result of the multiple members within a family, each with distinct polymer specificities. The majority of plant

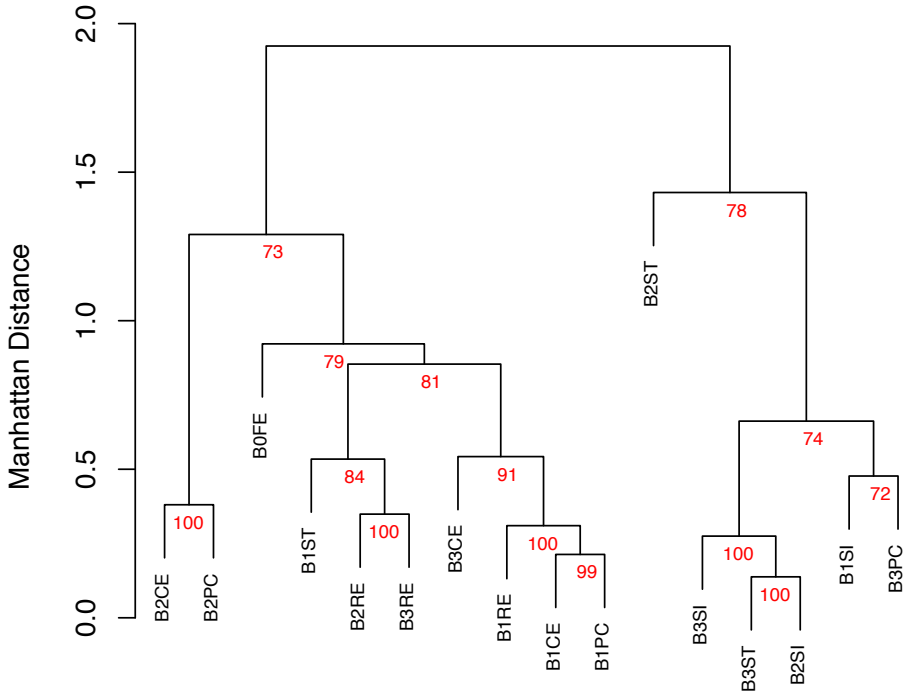

Figure 3.11: *Hierarchical Cluster Analysis of Beaver Gut and Feces CAZyme Abundance.* Hierarchical cluster analysis of the relative abundance of ORFs annotated as GHs, CEs, and PLs. A Manhattan distance matrix was used to determine distance. Approximately unbiased p-values for each branch in the dendrogram were determined through bootstrap resampling. Sample names are abbreviated as follows: ST: Stomach, SI: Small Intestine, CE: Cecum, PC: Proximal Colon, RE: Rectum, FE: Feces.

biomass-degrading CAZy families were more abundant in the large intestine compartments, with the average fold increase being 3.0 ± 2.2. The beaver fecal sample also displayed a similar profile of biomass targeting enzymes (Figure 3.12), mirroring what was seen with hierarchical clustering.

3.4.3 Functional Screening

DNA from the intestinal samples was also used to construct fosmid libraries for functional screening. To this end, DNA from all five sites collected from beaver 2 were used in an attempt to create fosmid libraries. Unfortunately, libraries could not be created for the stomach and small intestine. The DNA from these samples was much more fragmented, than that from the cecum, colon and rectum samples, making library creation lower yielding. The DNA from the small intestine and

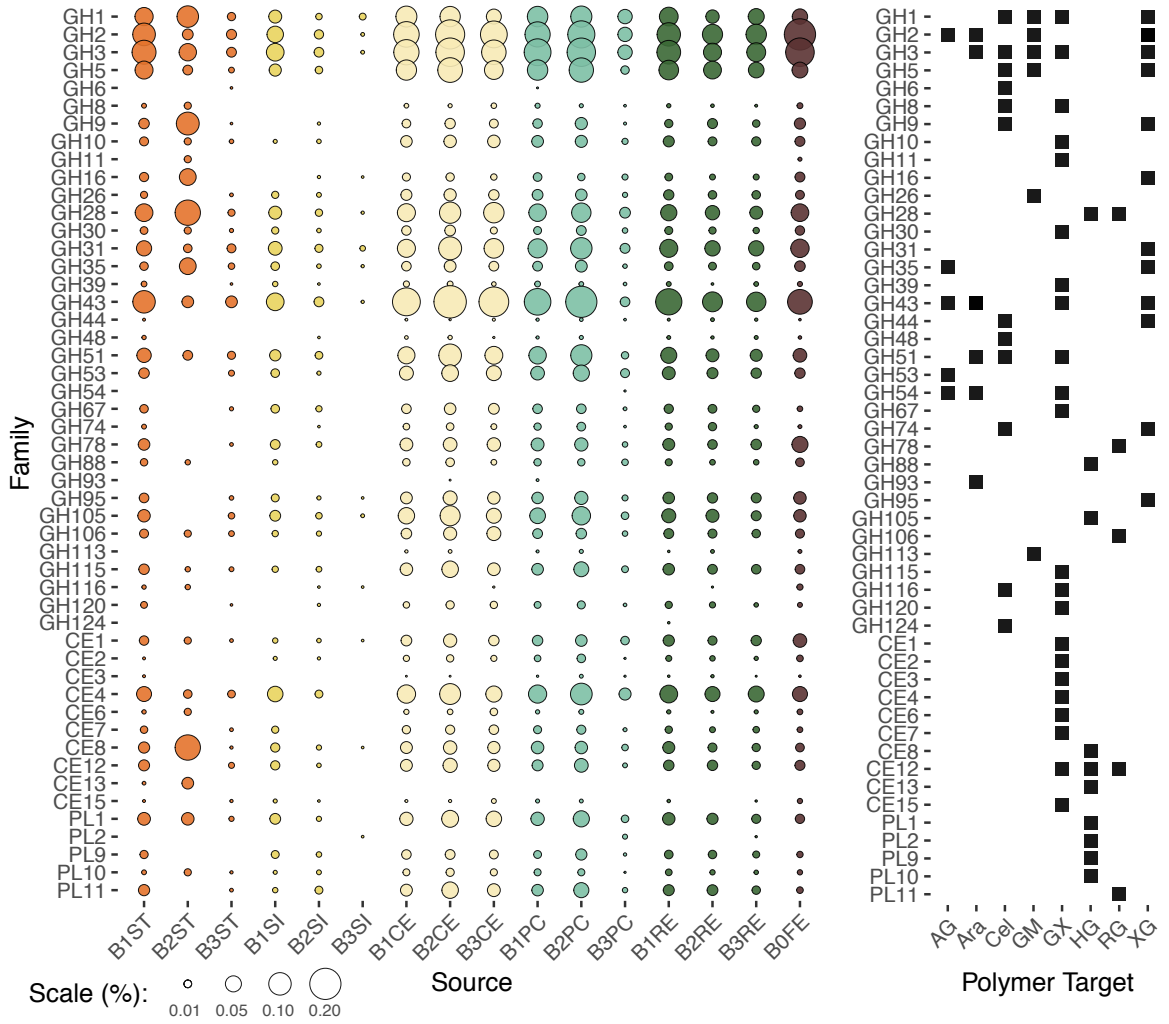

Figure 3.12: *Abundance of Plant Polysaccharide Degrading Cazymes in Beaver Gut Metagenomes.* CAZy families with annotated activities against plant polysaccharides are shown. Only families present in at least one sample are displayed. Sample names are abbreviated as follows: ST: Stomach, SI: Small Intestine, CE: Cecum, PC: Proximal Colon, RE: Rectum, FE: Feces. Bubbles are coloured by sample source. The polymer targets of each family are indicated by the presence of a black box. Polymers are abbreviated as follows: AG: Arabinogalactan, Ara: Arabinan, Cel: Cellulose, GM: Glucomannan, GX: Glucuronoxylan. HG: Homogalacturonan, RG: Rhamnogalacturonan backbone, XG: Xyloglucan. Bubbles are coloured by sample source.

stomach certainly have fewer bacterial colonies than further down the digestive tract and higher concentrations of DNA from food sources, which is likely to be partially degraded. Other hindgut-fermenting mammals, humans for example, have approximately 10^7 fold fewer bacterial cells in the stomach and small intestine than they do in the large intestine [270].

The three libraries made from beaver 2 intestinal DNA contained a total of 43,776 clones. Of these clones, 6,528 were derived from the cecum, 14,976 from the proximal colon and 22,272 from the rectum sample. Together these libraries contain nearly 10 times the number of clones created from the fecal DNA. The generated fosmid libraries were then subjected to functional metagenomic screening as described for the fecal library in order to identify active clones. The only alteration to the screening protocol was the use of a mixture of CMU-X2, CMU-C and CMU-Man as screening substrates instead of the mixture of CMU-X2, CMU-C and CMU-X, which was used for the fecal library. This was done for two reasons: we were able to achieve better signal to noise ratios when the mannoside was used instead of the xyloside, and almost all clones active on the xyloside were also active on the xylobioside. Functional screening identified a total of 374 clones that had plate-based

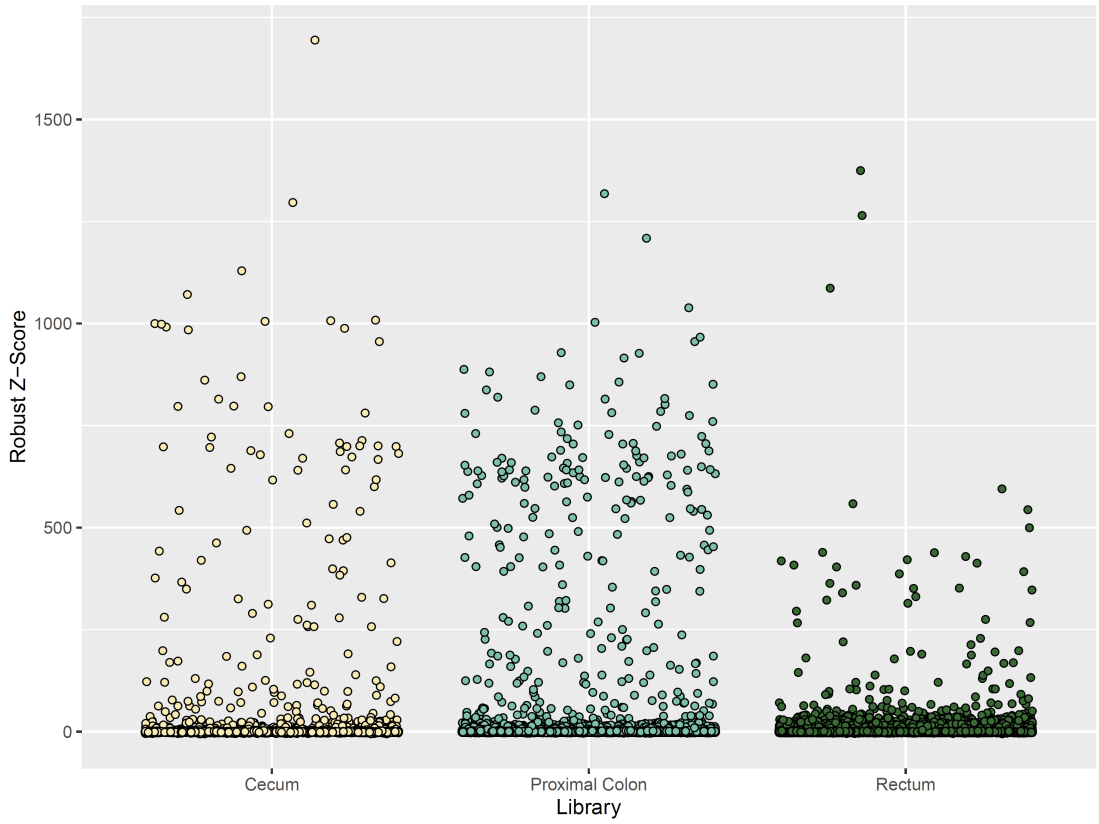

Figure 3.13: *Functional Screening of Beaver Gut Libraries.* Robust z-score values for fluorescence were calculated for each plate. Clones above the robust z-score threshold of 40 were chosen for further validation.

robust z-scores of 40, Figure 3.13. Validation and deconvolution of these hits revealed 196 active

clones (z-score > 10), of which 138 were active on the xylobioside, 104 active on the cellobioside and 5 active on the mannoside (Table 3.6). Quite a few clones (n = 48) were active towards both the xylobioside and the cellobioside suggesting the presence of gene clusters, as seen in the beaver fecal fosmids. The number of hits varied greatly between libraries, with the cecum derived library yielding the highest percentage of hits (1.29%), while the rectum (0.32 %), and proximal colon (0.27 %) libraries had much lower hit rates.

Table 3.6: Beaver Gut Hits

			Active Clones (Z-score >10)		
Library	Clones	Verified Hits	CMU-Cellobiose	CMU-Xylobiose	CMU-Mannose
Cecum	6,528	84 (1.29%)	40 (0.61%)	61 (0.93%)	2 (0.03%)
Proximal Colon	14,976	40 (0.27%)	23 (0.15%)	26 (0.17%)	1 (0.01%)
Rectum	22,272	72 (0.32%)	41 (0.18%)	51 (0.23%)	2 (0.01%)
Total	43,776	196 (0.45%)	104 (0.24%)	138 (0.32%)	5 (0.01%)

3.4.4 Fosmid Sequencing and Gene Annotation

To reveal the gene(s) responsible for activity, the DNA from fosmid hits was sequenced, assembled, and annotated. Of the 196 fosmids identified and validated 168 have been sequenced, all with greater than 20 kbp of sequence, Figure 3.14. The average insert length on these fosmids was 35,880 ± 7,064, for a total of 5.57 Mbp of DNA sequence data. ORFs were predicted as for the fecal fosmids, revealing an average of 21 ± 5.8 ORFs per fosmid. Out of the 168 sequenced fosmids, 119 were non-redundant. The threshold for redundancy was set as being > 95 % identity over 90 % of the sequenced fosmid. Several fosmids had redundancy with fosmids obtained from other compartments, for example: B2Rectum_12_G19 was redundant with three cecal fosmids (B2Cecum_01_M24, B2Cecum_02_L22, B2Cecum_06_M06) a proximal colon fosmid (B2PC_42_E09) and three rectal fosmids (B2Rectum_02_O10, B2Rectum_13_C08, and B2Rectum_19_N19). This cross-compartmental redundancy indicates that at least a subset of the active members within the metagenome can be found throughout the cecum to rectum transect.

To assess the presence of carbohydrate-degrading enzymes, the sequenced fosmids were annotated, as for the fecal fosmids, using a LAST [150] comparison to the CAZy database [181]. This revealed a total of 430 GHs spanning 29 different families. GH genes accounted for 11.08 % of all

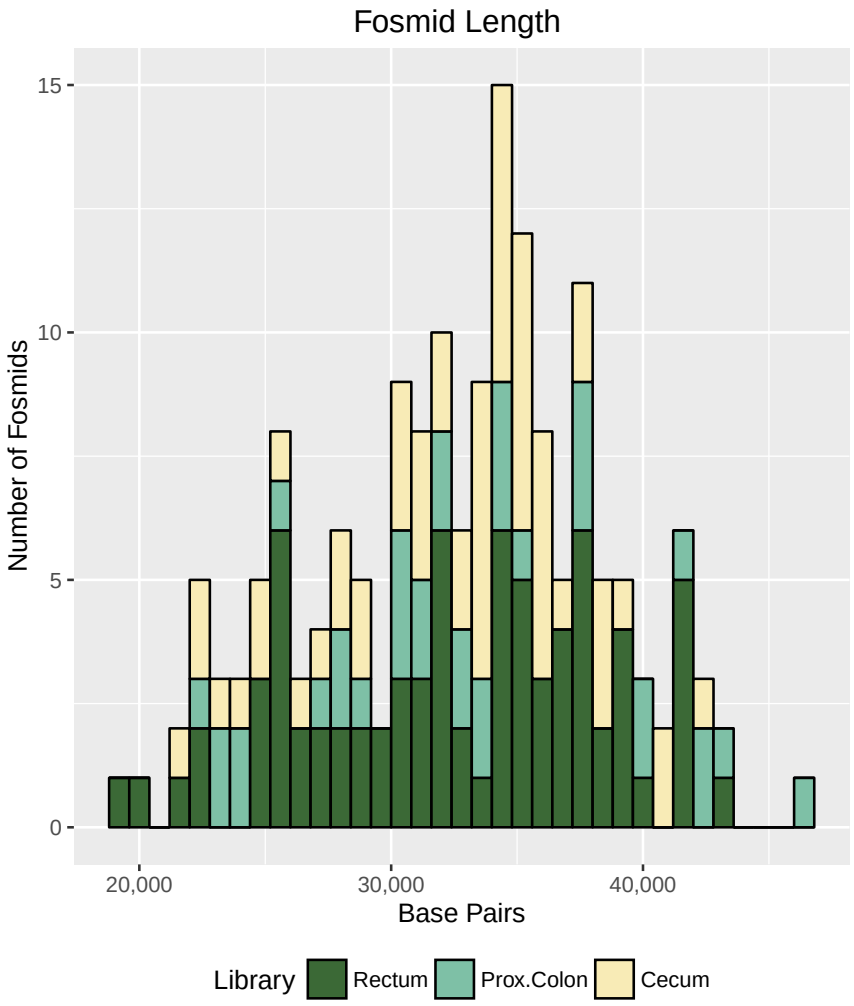

Figure 3.14: *Distribution of Beaver Gut Fosmid Insert Length.* Histogram showing the number of sequenced fosmids with a specified length, bars are coloured by the library source.

predicted fosmid ORFs, a relative frequency nearly identical to that seen for the fecal fosmids (11.13 %), Figure 3.15. The most abundant GHs found on the gut fosmids were GH43, GH3, GH2, GH67 and GH10 (1.86%, 1.78 %, 1.48 %, 1.39 % and 1.29 % of all ORFs respectively). This is somewhat different from the proportions of GHs found on the fecal fosmids, where the five most abundant families were GH3, GH43, GH1, GH2, and GH130 (2.68 %, 2.13 %, 0.71 %, 0.71%, and 0.47 % of all ORFs respectively). The most apparent difference between these two sets of fosmids was the substantial increase in xylan α-1,2-glucuronidases, from GH67 and GH115 in the gut fosmids. These two families made up 1.38 %, for GH67, and 0.46 %, for GH115, of all ORFs within the

gut fosmids, whereas only one GH67 (0.08 % of all ORFs) and no GH115s were identified on the fecal fosmids. Neither of these families have previously shown activity on xylosides, mannosides, or cellobiosides, thus their presence is likely due to frequent incorporation into glucuronylxylan cleaving loci. The stark contrast between the presence of α-1,2-glucuronidases may be partially explained by the differing frequencies of these two GH families within the different metagenomes screened. Both families are found more predominantly (3.9 fold and 2.3 fold greater for GH67 and GH115, respectively) in the beaver 2 average gut metagenome than in the fecal metagenome.

Another difference between the beaver fecal fosmid genes and gut fosmid genes was the decreased recovery of GH1 genes within the gut libraries. Although the beaver 2 gut had on average a greater percentage of ORFs assigned to the GH1 family (0.046 % for feces and 0.069 % for the average of gut compartments) the beaver feces fosmids had a far greater proportion of GH1 genes (0.71 % of ORFs for feces fosmids, 0.05 % for gut fosmids). This 14 fold increase over the gut libraries is likely due to the substrates used for screening. The GH1 family, which contains members with β-xylosidase activity, is likely more enriched in the fecal fosmids as this library was screened with the CMU-X, which was absent from the assay mix used to screen the gut libraries.

An additional aberration from expectations was the absence of any GH8 from the recovered fosmids. The GH43-GH8 gene recovered from the fecal fosmids showed an intriguing synergistic mechanism tuned for the degradation of arabinoxylans. I would have expected to find similar genes within the fosmids recovered from the beaver gut microbiomes, or GH8 genes present within xylan degrading gene clusters. One contributing factor to the absence of recovered GH8s may be that the relative abundance of GH8 genes was substantially lower in beaver 2 large intestine metagenomes (2.8 fold decrease from the fecal metagenome).

Analysis of the sequences revealed that two of the fecal fosmids, displayed homology to fosmids recovered from screening of the beaver gut. The first of these feces sourced fosmids, 04_C21, had greater than 95 % identity to B2Cecum_08_E22 over a 25 kb region containing five GH genes (GH43, GH67, GH95, GH29, GH25). The fecal fosmid 04_C03 also had $>$ 95 % similarity to B2Rectum_42_O23 over a total of 22305 bp, containing a GH3, GH16 and a GH32. The repeated discovery of fosmids containing nearly identical functional genetic regions from distinct samples differing in multiple aspects (digestive compartment, sampling procedure and host organism) alludes

to the presence of a core functional membership within the beaver microbiome.

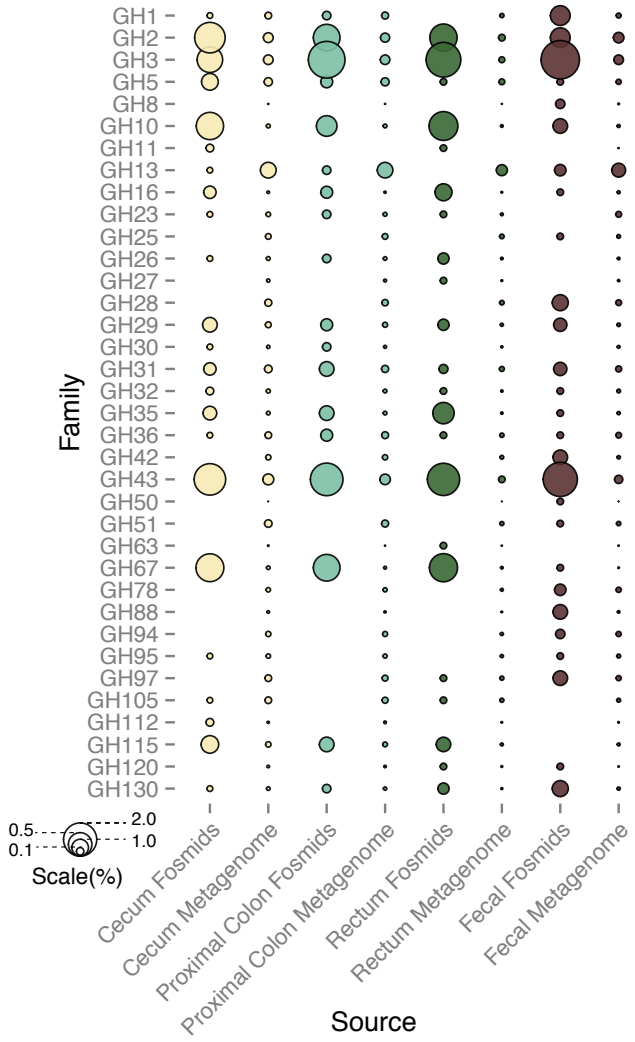

Figure 3.15: *Relative Abundance of Glycoside Hydrolases in Sequenced Fosmids and Metagenomes.* Bubbles plot shows the relative abundances of each GH family recovered from positive fosmid clones for each library source, including the beaver fecal library. Bubble area is proportional to the relative abundance. Bubbles are coloured by library source. Only the GH families found within at least one sequenced fosmid are shown.

3.4.5 Presence of Polysaccharide Utilization Loci

As the fosmids recovered from the fecal library contained a multitude of PULs, including those with novel organizations, I sought to identify PUL-containing fosmids recovered from the gut fosmids.

As previously, PUL-containing fosmids were identified by their signature tandem SusC-like/SusD-like pairing, which is a hallmark of the presence of PULs [295]. In total 69 of the 168 fosmids, 41 % of all sequenced beaver gut fosmids, contained PULs. Within this set of PUL-containing fosmids, 50 were non-redundant (< 95 % similarity over 90 % of the fosmid).

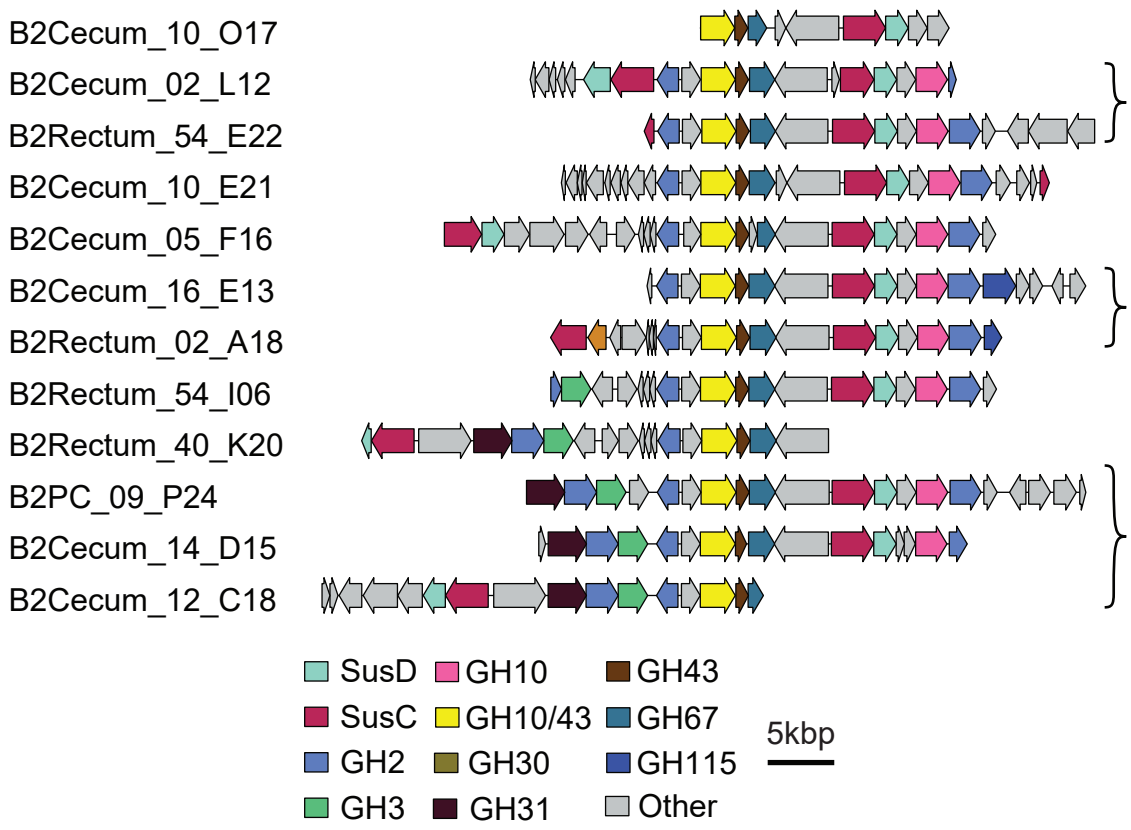

Figure 3.16: *Gene Organisation of Beaver Gut Fosmids Containing SusC/SusD-like Proteins and a Two Domain GH10-GH43.* Putative glycoside hydrolases and SusC/SusD-like proteins are coloured with the same scheme as Figure 3.4, with the exception of the two domain GH10-GH43. ORFs not annotated as a glycoside hydrolase, SusC-like, or SusD-like, are shown in grey. Fosmids pairs or sets within brackets share greater than 95 % identity over their overlapping region. Fosmids have been aligned to highlight synteny.

Inspection of this set of all beaver gut, PUL-containing fosmids revealed multiple syntenic groups, see Figures 3.16, 3.17 and 3.18. The first of these groups is recognized by a conserved motif of a dual domain GH10-GH43 (GH43 subfamily 12) protein followed by an additional GH43 (subfamily 1) and a GH67. All of these GH10-GH43 containing PUL fosmids had optimal activity on

xylobiose. This motif is present in 12 non-redundant fosmids, of which 3 sets have nucleotide identity greater than 98 % over their overlapping regions. This motif, including the two domain protein, is also present on the fecal fosmid 04_C21. This cluster of genes has synteny with the xylan degrading cluster of genes identified within *Bacteroides intestinalis* [311]. Specifically, BACINT_04202, from Bacteroides intestinalis, which contains a GH10 fused to a GH43 subfamily 12 domain, has 55 -57 % amino acid identity with the GH10-GH43 proteins identified from the beaver gut. The BACINT_04202 protein has endo-xylanase and arabinofuranosidase activity, which is able to release xylose, xylosoligosaccharides and arabinose from cereal arabinoxylans [311].

Several of the fosmids identified with the GH10-GH43 gene cluster also have additional hydrolase genes. A set of 9 fosmids also code for a GH10 and GH2 downstream of the GH10-GH43 cluster, two with an additional GH115. Another set of four GH10-GH43 containing fosmids had a GH31, GH2, GH3 motif upstream, which resembles a portion of the characterized xyloglucan degrading XyGUL-PUL from *B.ovatus* [165]. This GH31, GH2, GH3 motif is also seen on two other fosmids, B2Cecum_01_K09 and B2Cecum_13_L24, which lack a the GH10-GH43, GH43, GH67 genes. Furthermore, all fosmids containing this XyGUL like motif also had activity against CMU-C, suggesting that this region is responsible for the observed activity.

A second subset of fosmids, was active on CMU-X2, yet lacked the GH10-GH43 region, see Figure 3.17. Many of these fosmids contained GH genes that were present in the downstream region of the first cluster; GH10, GH2, GH115 and GH43s were commonly seen to be present on these fosmids. In fact, a GH10 protein was seen in all 17 fosmids within this group, highlighting the importance of this family, which is well know for the degradation of xylans [105]. The high abundance of α-glucuronidases within the beaver gut fosmids is in part due to their presence in these loci, as over half of the fosmid GH67 and GH115 genes were found on PUL-containing fosmids.

The final set of 11 PUL-containing fosmids was most active on CMU-C, and showed limited activity against CMU-X, Figure 3.18. This set of fosmids bore a resemblance to the PUL-containing fosmids in Chapter 2, which were identified with the substrate MU-C. All of these fosmids had GH3 present, likely the active protein, and many of these contained an additional GH16 domain within the PUL. As noted earlier, two of these fosmids, B2Cecum_01_K09 and B2Cecum_13_L24, had synteny with GH10-GH43 containing fosmids and likely target xyloglucans.

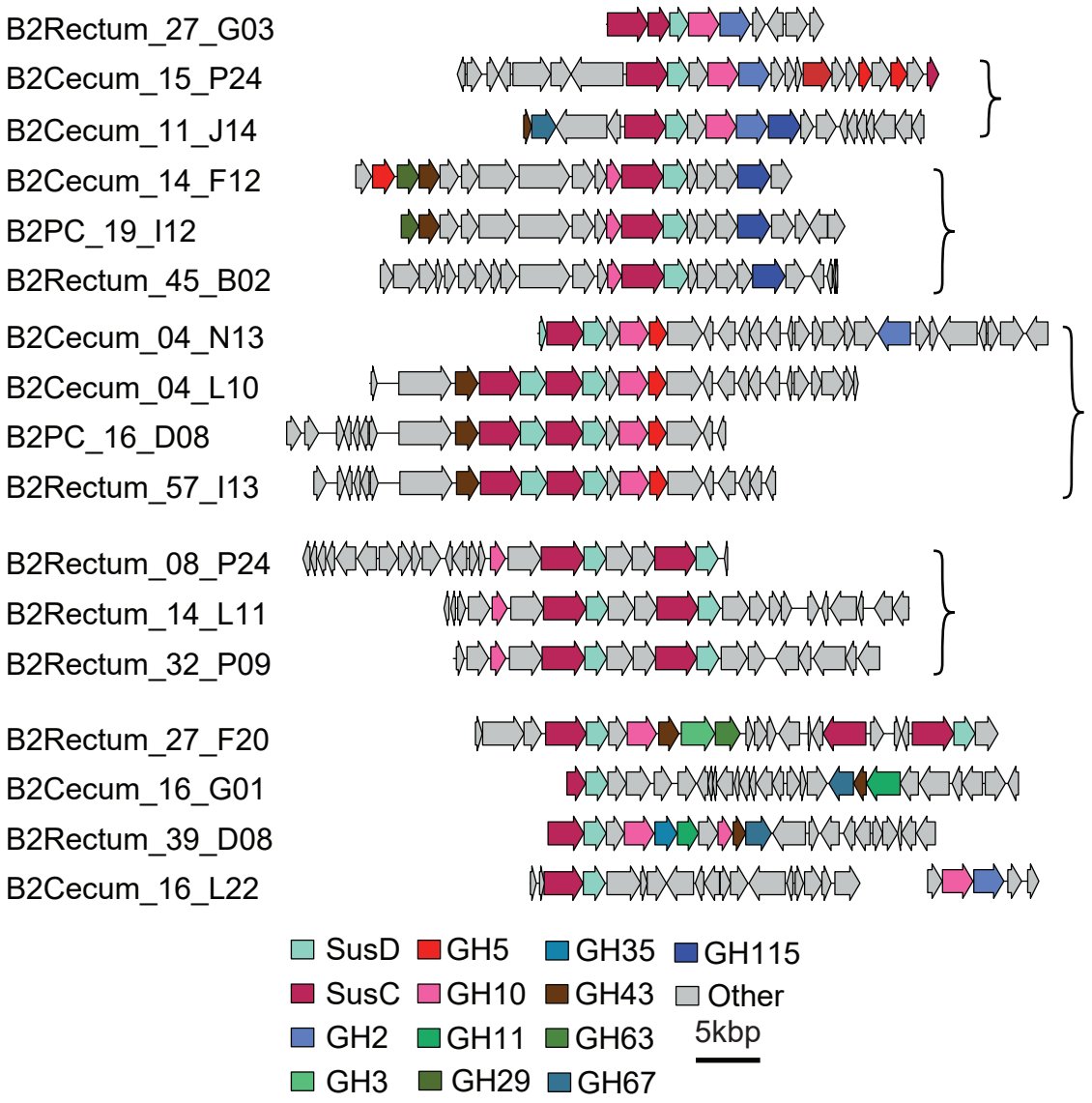

Figure 3.17: *Gene Organisation of Beaver Gut Fosmids Containing SusC/SusD-like Proteins With Highest Activity on CMU-X2.* Putative glycoside hydrolases and SusC/SusD-like proteins are coloured with the same scheme as Figure 3.4. ORFs not annotated as a glycoside hydrolase, SusC-like, or SusD-like, are shown in grey. Fosmids pairs or sets within brackets share greater than 95 % identity over their overlapping region. Fosmids have been aligned to highlight synteny.

3.5 Limitations and Future Directions

This multifaceted analysis of beaver fecal and gut microbial communities has revealed both the community members present within these microbiomes and the molecular mechanisms they im-

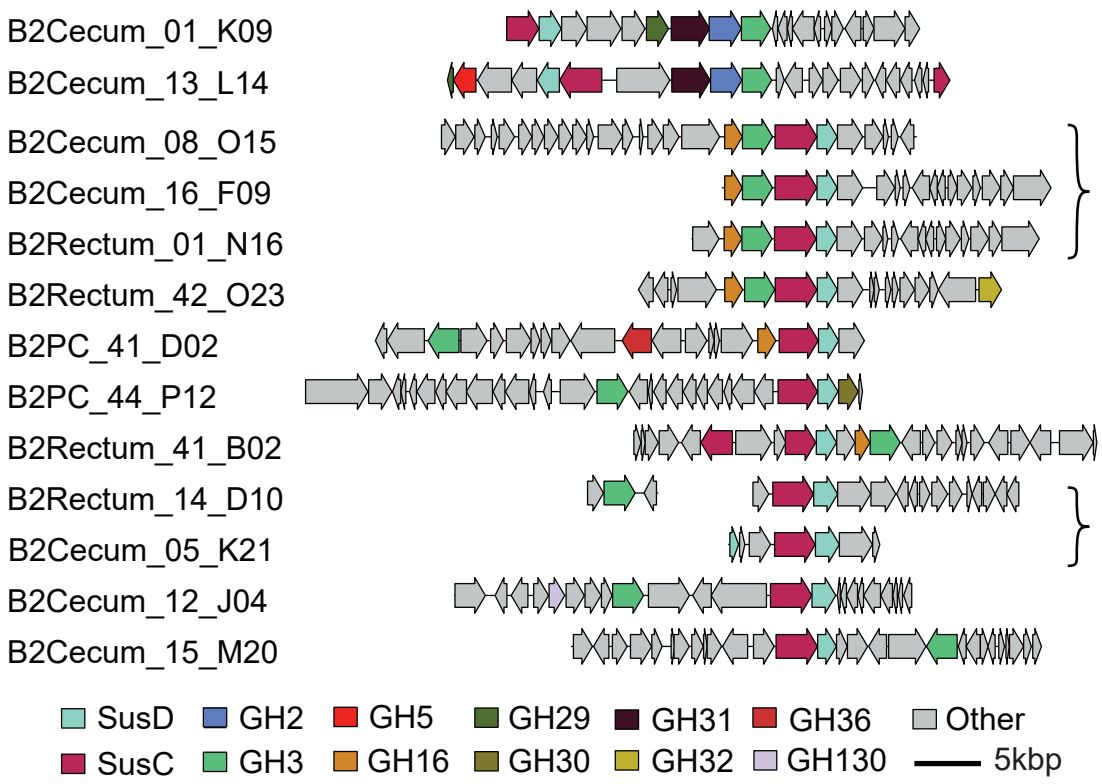

Figure 3.18: *Gene Organisation of Beaver Gut Fosmids Containing SusC/SusD-like Proteins With Highest Activity on CMU-C.* Putative glycoside hydrolases and SusC/SusD-like proteins are coloured with the same scheme as Figure 3.4. ORFs not annotated as a glycoside hydrolase, SusC-like, or SusD-like, are shown in grey. Fosmids pairs or sets within brackets share greater than 95 % identity over their overlapping region. Fosmids have been aligned to highlight synteny.

plement to degrade plant polysaccharides. There are however some constraints which limit our interpretation of the data presented. The first of these limitations is the under-sampling of clone libraries. The number of clones needed to ensure a library is representative can be estimated using the Carbon and Clarke formula [58], equation 3.1.

$$N = ln(1 - P)/ln(1 - f) \tag{3.1}$$

Where N is the number of clones needed, P is the probability that a given sequence is present in the library, and f is the fractional size of each insert of the total genome. Assuming that there are 1,000 species of bacteria present within the beaver gut, the average fosmid insert is 40 kbp

and the average bacterial genome is 5 Mbp in length, you would need a library of over 370,000 clones of to have a 95 % chance of finding a certain genetic locus. This number of clones becomes significantly higher when the low abundance of rare taxa is taken into account. The largest library generated within this study contained 22,272 clones which is certainly an under-sampling of the environment. As the number of clones needed to be screened for diverse environment nears the technical limitations for plate-based screening, other screening technologies may offer the ability to more fully screen metagenomic libraries. Recent advances in using droplet-based microfluidics for functional metagenomics are particularly alluring [211] and should allow for a more complete sampling of microbial communities.

The beaver diet, like that of many mammals, is seasonal. During the spring and summer months herbaceous aquatic vegetation constitutes a higher proportion of their diet than in the winter months, when they become more dependent on tree species [5, 35]. As such, one might expect that the microbial communities shift throughout the seasons. Indeed, this seasonal variability of gut community composition has been observed for several other mammalian species, including wild mice [194], reindeer [92], bison [190], cattle [276] and humans [277]. This study is limited in the assessment of seasonal variation as a majority of the samples were obtained in the spring, and no samples were collected in the summer or fall. Seasonal bias in sampling is difficult to avoid, as the British Columbia Ministry of Environment currently restricts beaver trapping to the period between October 15th and April 30th in the lower mainland region. Further examination of the beaver gut at different times throughout the year could reveal a shift in community which coincides with a shift in diet from herbaceous material to woody plant matter.

Another source of variation in the beaver diet originates from the wide range of habitats in which it can be found. North American beaver have been found to inhabit Arctic Tundra and range as far south as Northern Mexico and are an invasive species in Tierra del Fuego, Argentina [56], and Finland [231]. As the plant species within these ecozones varies widely one would expect the diet of the various beaver sub-populations to also vary. As the glycan component of plant matter is also variable between plant species [42], it stands to reason that different beaver populations would be exposed to plants with variable carbohydrate compositions and substitution patterns. Investigation of the gut community composition as it varies with diet could lead to a better understanding of

the microbial degradation of specific plant polymers from defined species.

This thesis chapter has examined the genetic potential encoded within the fecal and gut microbiomes. However, the genes present within the community are surely expressed at differently levels. Metatranscriptomic analysis, such as that performed for Arctic ground squirrels [115], cattle [64] or humans [95, 238], would reveal which degradative enzymes within the beaver gut are actively being transcribed. Moreover, this could illuminate transcriptional differences as digesta transit through the gut. Integration of metaproteomic data could further refine the mechanistic details involved in the degradation of plant matter within the beaver gut. Metaproteomic studies have looked at plant polysaccharide digestion in the termite [315] and shipworm [226]. The application of this method to the beaver gut could reveal not just which proteins are being expressed, but also what subset are being secreted into the gut environment to degrade plant polymers.

3.6 Conclusions

This chapter opens a functional metagenomic window on the capacity of the beaver fecal and gut microbiomes to deconstruct woody plants into soluble sugars supporting host nutrition. Although beavers are considered xylotrophic organisms, their microbiome composition is most similar to those of hindgut-fermenting mammals dominated by Bacteroidetes and Fimicutes. Multiplexed high-throughput functional metagenomic screening was applied to recover active glycoside hydrolase (GH) enzymes from these metagenomes. The functional screening of fosmid libraries sourced from the beaver fecal and gut microbiomes with model cellulose and xylan substrates revealed 247 fosmids with an array of carbohydrate degradation profiles. A subset of these fosmids contained GH43 enzymes belonging to previously uncharacterized subfamilies. Cloning and expression revealed the substrate specificity of three previously uncharacterized subfamilies and revealed a mechanism involving two of these family 43 hydrolase domains and an appended family 8 glycoside hydrolase domain which synergistically degrade arabinoxylan oligomers. Fosmid clones recovered in this study also revealed novel assortments of genes clustered into polysaccharide utilization loci (PULs) with the potential to synergistically enhance biomass deconstruction in support of host nutrition. Interestingly we did not identify an abundance of fosmid genes encoding cellulases indicating the

potential for overexpression by specialized strains.

The panoply of genes encoding enzymes with hemicellulose degrading activities is presumably required to tackle the complex structures of hardwood biomass in which glucan and xylan backbones are extensively decorated with appended sugars. The presence of two-domain enzymes endows the beaver fecal and gut microbiomes with synergistic capacity to efficiently degrade the specific linkages present within the hemicellulose. Not only does this liberate monosaccharides, but the degradation of hemicellulose scaffolds also exposes the underlying cellulose fibers for digestion by the cellulase repertoire, enhancing conversion efficiency. Taken together these results provide a unique perspective on the modular domain architecture and functional specialization driving combinatorial biomass deconstruction in the beaver fecal and gut microbiomes.

Chapter 4

Harnessing Natural Diversity to Profile Promiscuity and Create New Glycosynthases

4.1 Summary

The use of chemical probes bearing unnatural functional groups has enabled the discovery and characterization of enzyme activity. However we do not generally know how well specific modifications on the sugar ring are tolerated by glycoside hydrolases. Functional screening performed in Chapters 2 and 3 produced a library of functional clones containing a diverse set of glycoside hydrolase genes. Access to both this fosmid hit library and a synthetic GH1 gene library has enabled us to address the issue of tolerance of azido-, amino- and methoxy- functional groups by glycoside hydrolases. Additionally, it would be valuable to exploit any promiscuous activities identified to create variant hydrolases with synthetic capacity. To this end, I performed high-throughput plate-based screening of two separate libraries, followed by kinetic analysis to identify clones and their expressed enzymes with promiscuous hydrolase activity. I then characterized the acceptor specificity of these enzymes and used site-directed mutagenesis to create active glycosynthases. This revealed which modification positions and functional groups are best tolerated. Eight new glycosynthases, with a variety of activities and ability to incorporate modified glucosides and galactosides, were created from the selected promiscuous enzymes.

4.2 Background

Enzymes are often thought to be highly specific, yet many enzymes have minor activities on substrates for which they were never specialized [149, 220]. Termed promiscuous activities, these secondary functions are often starting points for the evolution of new activities, or hold-overs from previous ancestral functions [149, 224]. Furthermore, promiscuous enzyme activities can be exploited to create new biocatalysts [33]. These latent secondary activities, however are often difficult to predict and are often overlooked when enzymes are characterized.

Investigation of promiscuous activities has particular relevance to chemical glycobiology, as a number of studies have relied on promiscuous enzymatic activity towards chemical probes containing unnatural functional groups, to reveal functions in both *in-vitro* and *in-vivo* experiments. Several studies have used azido sugars to enable the metabolic labelling and visualization of glycans on cell surfaces and in organisms [51, 167, 310, 341]. Alkynyl sugars, modified at several different positions, have similarly been used to label cancer cell glycans [131, 140], animal glycans [50] and plant glycans [343]. Examples of both alkynyl and azido sugars are given in Figure 4.1. Despite their apparent utility in interrogating biochemical activities, we don't know how well specific modifications are tolerated by CAZymes, including the glycoside hydrolases.

Assessment of hydrolase promiscuity can also help us to address another long-standing problem in the biological sciences: ready access to synthetic oligosaccharides. The development of facile methods for the synthesis of peptides and oligonucleotides revolutionised the biological sciences. However, despite excellent progress in automated glycan synthesis [111, 214] we still lack a universal and reliable method for glycan assembly, largely because of the enormously greater regio- and stereochemical challenges posed. The alternative to chemical synthesis involves use of enzymes, most likely either the natural nucleotide phosphosugar-using glycosyltransferases or synthetically useful variant forms of glycoside hydrolases – glycosynthases [12]. These latter enzymes are created by mutating the catalytic nucleophile residue of a retaining glycosidase (Glu or Asp in almost all cases) typically to either Ala, Gly or Ser. Such variants are hydrolytically incompetent with natural substrates, but when presented with a glycosyl fluoride donor sugar of inverted anomeric configuration, will typically transfer that sugar to an appropriate acceptor, often in near stoichiometric

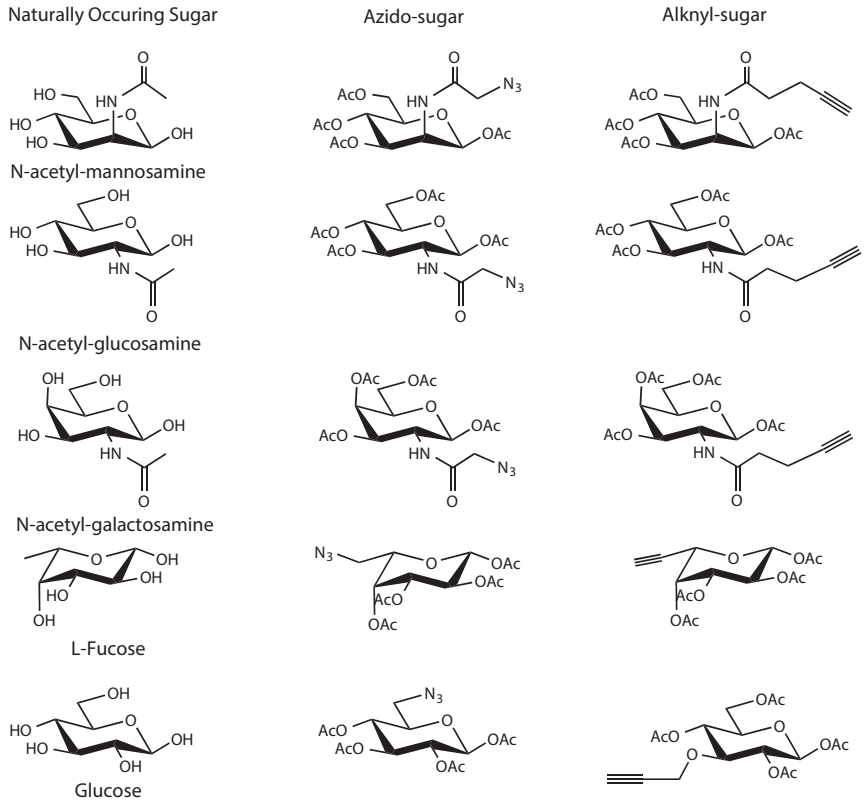

Figure 4.1: *Modified Sugars Used for Metabolic Labelling.* Examples of per-acetylated azido- and alkynyl-glycosides used in metabolic labelling experiments and their unmodified parent.

yield, without subsequent hydrolysis [12, 72]. The first glycosynthase generated was from the GH1 β-glucosidase from *Agrobacterium sp.* [188] and many have been developed since and are used to create a variety of glycans including: glycolipids [253], glycosidase inhibitors [106] and defined glycoproteins [61, 72, 103, 110, 160].

This has been accomplished through the creation of glycosynthases from new families [110], and through directed evolution [153]. Our intent in this work was to use libraries of fosmids and large synthetic gene libraries as a means of identifying useful catalysts for the formation of specific glycosides that could not be assembled with currently known enzymes. Of particular interest was the ability to assemble oligosaccharides containing amino- or azido substituents at the 3, 4 or 6 positions. These would be useful not only in the degradation/assembly of aminosugar-containing glycans, such as antibiotics or bacterial surface polysaccharides, but also as a way of incorporating

modifiable glycans into biomolecules under mild conditions for subsequent tagging. More broadly this served as a test system for generating a pipeline for the discovery and generation of custom biocatalysts for glycan assembly.

In this chapter I aimed to harness the diverse set of enzymes encoded on the fosmid hits and within a synthetic gene library to evaluate their capacity to hydrolyse modified synthetic glycosides. The research performed in Chapters 2 & 3 provided us with a panel of hydrolases with which we could begin to explore the distribution of promiscuous hydrolase activity. Additionally, we have obtained a synthetic gene library, produced by the JGI, which contains a diverse set of 175 GH1 family genes [117]. By similarly interrogating this synthetic gene library we hoped to reveal the promiscuity within this specific family. By assaying these two libraries with azido-, amino- and methoxy-glycosides we hoped to gain insight into which promiscuous activities are more prevalent among glycoside hydrolases.

I have furthermore exploited the promiscuous enzymes identified to generate new biocatalysts. A set of 15 glycoside hydrolases, from both the metagenomic and GH1 libraries, were selected to be further investigated and transformed into glycosynthases. The acceptor (+1 site) specificities of all selected enzymes were then explored against a panel of acceptors in a second, plate-based screen, identifying optimal candidates to utilize as acceptors in subsequent glycosynthase reactions (Figure 4.2). The alanine, serine and glycine variants of the nucleophilic glutamate residue were created for each of the fifteen GHs and the activities of the corresponding 45 variants were evaluated. Variant forms of one of the seven fosmid derived genes and seven of the eight chosen synthetic GH1s acted as competent glycosynthases yielding the desired glycans containing azido or amino substituents. Finally, the utility of these glycosynthases in the assembly of taggable activity-based probes was demonstrated.

4.3 Fosmid Hit Libraries

4.3.1 Screening with Modified Glycosides

In the search to identify glycoside hydrolases with activity on modified glycosides we harnessed the functional clone collections generated in Chapters 2 and 3. Active clones that were identified with

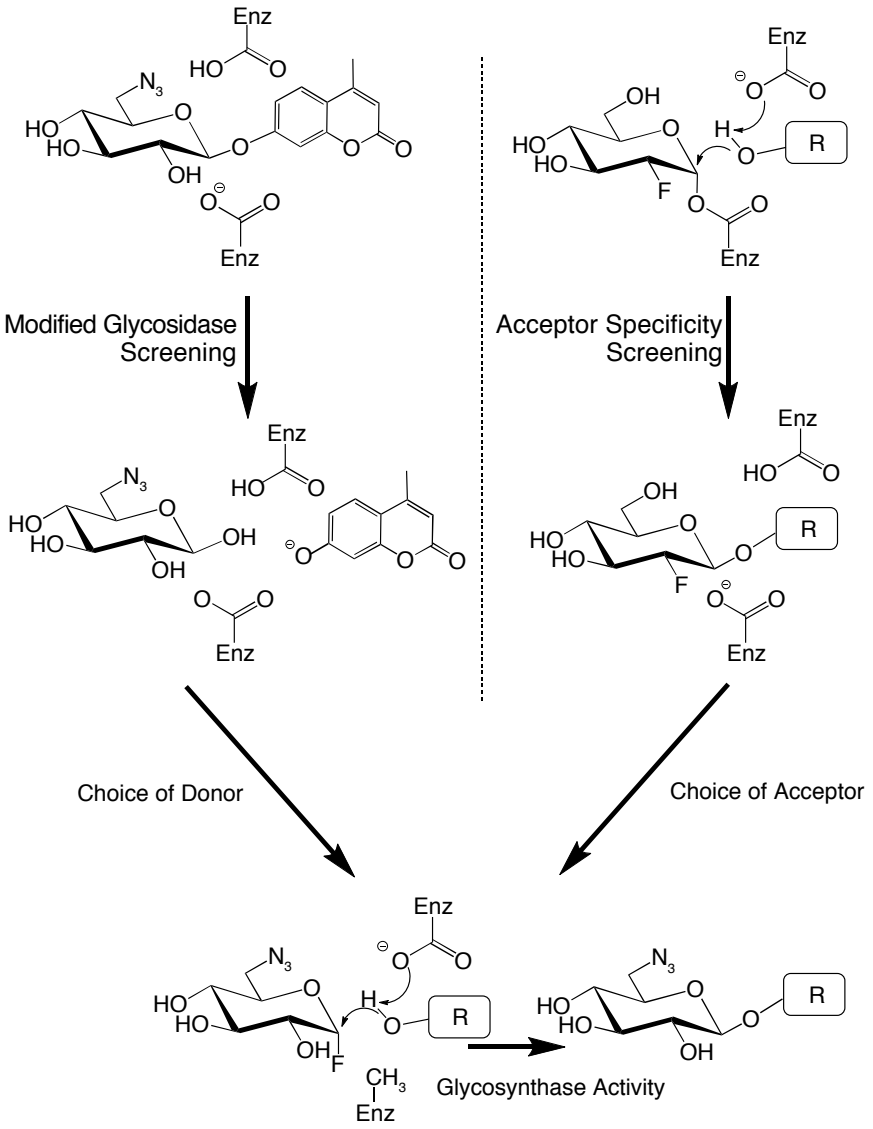

Figure 4.2: *Screening Methodology.* The -1 subsite specificity is evaluated through glycosidase screening with fluorogenic substrates. Acceptor specificity is based on stimulation of reactivation of a trapped 2-fluoroglycosyl enzyme through transglycosylation to a competent acceptor. The extent of reactivation is assessed with a chromogenic substrate, giving an indication of +1 subsite specificity. The corresponding glycosynthase can then be employed with cognate acceptor and donor substrates.

DNP-C or CMU-C from the libraries described in Chapter 2 with were also screened. In total 653 active clones, from soil, ocean, bioreactor, coal bed and beaver fecal libraries were screened with 10 different modified glucosides and galactosides, see Figure 4.3. The fluorogenic probes used in the

screens contained either an amino group at the 3-, 4-, or 6-position, an azido group at that 3-, 4-, or 6-position or a methoxy group at the 3- or 4-position, Figure 4.3. Screening was performed in the same manner used to originally identify the clones.

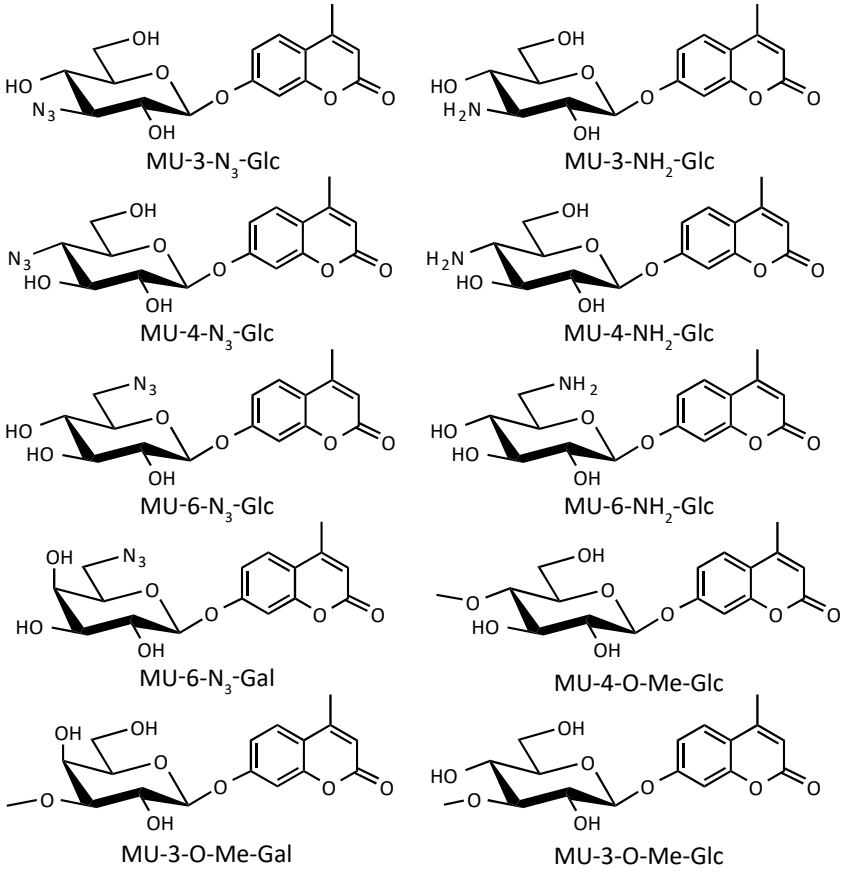

Figure 4.3: *Modified Glucosides and Galactosides Used for Screening.* The fluorogenic substrates used were: 4-methylumbelliferyl 3-amino-3-deoxy-β-D-glucopyranoside (MU-3-NH_2-Glc), 4-methylumbelliferyl 4-amino-4-deoxy-β-D-glucopyranoside (MU-4-NH_2-Glc), 4-methylumbelliferyl 6-amino-6-deoxy-β-D-glucopyranoside (MU-6-NH_2-Glc) an azido group at that 3-,4-, or 6-position (4-methylumbelliferyl 3-azido-3-deoxy-β-D-glucopyranoside (MU-3-N_3-Glc), 4-methylumbelliferyl 4-azido-4-deoxy-β-D-glucopyranoside (MU-4-N_3-Glc), 4-methylumbelliferyl 6-azido-6-deoxy-β-D-glucopyranoside (MU-6-N_3-Glc), 4-methylumbelliferyl 6-azido-6-deoxy-β-D-galactopyranoside (6-N_3-Gal MU) or a methoxy group at the 3-position (3-methoxy-β-D-galactopyranoside (MU-3-O-Me-Gal) and 3-methoxy-β-D-glucopyranoside (MU-3-O-Me-Glc)).

Screening revealed a total of 264 clones that hydrolysed at least one of ten compounds tested, as determined by a robust z-score greater than 10, see Figure 4.4 and Table 4.1. The most frequently accepted modification was the incorporation of an amino group, with the 6-, 4- and 3-amino glu-

cosides being hydrolysed by 87, 164 and 17 clones respectively. The azido modification was also accepted by a number of clones, though substitution at the 6-position was substantially better tolerated than the 4- or 3-position for the glucosides. Few clones were able to hydrolyse the methoxy modified glycosides (only 13 hits were identified for MU-3-O-Me-Gal and none for MU-3-O-Me-Glc and MU-4-O-Me-Glc), though substrates with methoxy groups at the 6-position were not tested.

The presence of multiple GH families within this library and fosmids with multiple GHs complicates the rationalization of the observed substrate preferences. However, one initial point of comparison is the GH3 family, as this was the most abundant GH family found on fosmids from Chapter 2 and Chapter 3. Investigation of GH3 β-glucosidase structures both containing bound substrates (PDB:1IEW, 2X41) reveals that the 3- and 4-hydroxyl groups both have a greater number of amino acid residues positioned within hydrogen bonding distance than does the 6-hydroxyl [128, 242]. Additionally, the 6-position appears to be pointing out of the active site, whereas the 3- and 4-hydroxyls are facing the enzyme interior. This corresponds well with our hit rates, as the amino group (which is small and able to maintain hydrogen bonding) has a higher hit rate for the 3- and 4-positions than the azido- or methoxy- modified sugars, while modifications at the 6-position seem to be accommodated regardless of their size. It is difficult to expand these justifications to the modified galactosides, as the GH3 families do not typically exhibit β-galactosidase activity.

Modification	Parent	3-Position	4-Position	6-Position
Azido	Glucose	4	8	158
Amino	Glucose	17	164	87
Methoxy	Glucose	0	0	N/A
Azido	Galactose	N/A	N/A	79
Methoxy	Galactose	13	N/A	N/A

N/A: Not tested

Table 4.1: Number of Fosmid Hits for Each Modified Substrate (Robust Z-Score >10). A total of 653 active clones were screened of which 203 cleaved MU-Glc.

4.3.2 Kinetic Characterization of Hydrolases

Sequenced hits with the highest fluorescence were selected for further characterization, Table 4.2. In total seven fosmid hits were selected, which together cleaved eight of the ten modified glycosides. Several of the selected hits cleaved more than one modified glycoside, see Table 4.2. Additionally,

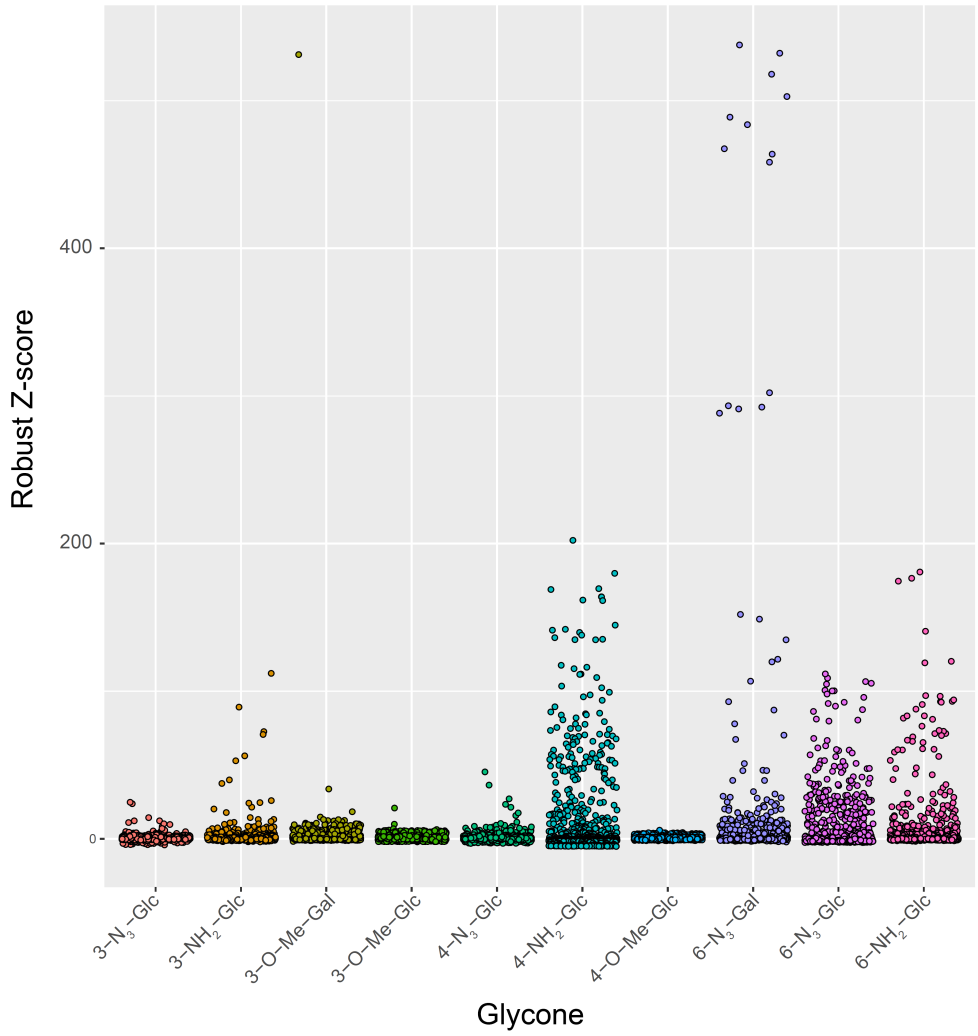

Figure 4.4: *Functional Screening of Hit Libraries with Modified Glycosides.* Robust z-score values for fluorescence were calculated on a per plate basis.

six of the seven selected hits had more then one glycoside hydrolase gene encoded on the fosmids, TolDC_15_C08 was the exception to this as it only encoded one GH from family 3, see Figure 4.5. To limit the number of genes for down stream analysis, I selected only those genes from families with known activities that corresponded with the parent compound, i.e. β-glucosidase families for modified glucosides and β-galactosidase families for modified galactosides. These genes were further limited to the set that had a retaining mechanism, in the hopes that these could eventually be transformed into glycosynthases. A total of 10 genes were selected for sub-cloning and expression tests, see Table 4.2 and Figure 4.5 . Only one fosmid, CA233_02_C24 had multiple genes which

Table 4.2: Selected Fosmids with Activity on Modified Glycosides and the Genes Selected for Sub-Cloning and Expression

Clone	Substrate (MU-)	Fluorescence (RFU)	Robust Z-score	Proteins Selected	GH Family	Expected Activity
CG23A_23_I01	3-N_3-Glc	91	13.9	I01_GH1	GH1	6-phospho-β-glucosidase
	3-NH_2-Glc	3,518	55.8			
TolDC_15_C08	3-NH_2-Glc	1,718	21.1	C08_GH3	GH3	β-glucosidase
Beaver_09_O03	3-O-Me-Gal	2,574	465	O03_GH42	GH42	β-galactosidase
NapDC_14_D08	4-N_3-Glc	379	44.9	D08_GH3	GH3	β-glucosidase
	4-NH_2-Glc	24,832	163.5			
CA233_02_C24	4-N_3-Glc	317	36	C24_GH3-1,	GH3	β-glucosidase
	4-NH_2-Glc	24,505	168.5	C24_GH3-2,	GH3	β-glucosidase
	6-N_3-Glc	2892	111.3	C24_GH3-3,	GH3	β-glucosidase
	6-NH_2-Glc	13,237	180.3	C24_GH3-4	GH3	β-glucosidase
FOS62_41_C11	6-N_3-Gal	4793	467	C11_GH1	GH1	β-glucosidase
FOS62_40_O22	6-N_3-Glc	2734	104.9	O22_GH3	GH3	β-glucosidase
	6-NH_2-Glc	9,117	118.8			

fit the aforementioned criteria. This fosmid contained four GH3 enzymes, any of which may have been responsible for the detected activities.

The gene selected from CG23A_23_I01, belongs to the GH1 family which contains many members with β-glucosidase activity. However, inspection of the gene sequence revealed a SKY motif, identified by Heins et al. [117] as indicative of 6-phospho-β-glucosidase activity. Furthermore, this fosmid also codes for genes annotated as PTS IIA, IIB and IIC, essential components of the phosphotransferase system, which concomitantly imports and phosphorylates sugars [198]. Therefore it may be the case that the observed activity is a result of the hydrolysis of the screening compounds once they have been phosphorylated.

All ten selected genes were sub-cloned into a pET28 vector backbone with a C-terminal hexahistine tag. As all four CA233_02_C24 GH3 genes had N-terminal signal peptides (as determined by the SignalP server [218]) N-terminal truncated genes were used. All ten sub-cloned genes were then transformed into *E. coli* BL21(DE3) for expression. Nine of the ten genes were expressed at sufficient levels for purification, with the exception being C24-3-GH3. Kinetic characterization of the wild-type proteins was performed to confirm the activities observed from fosmid clones, see Table 4.3.

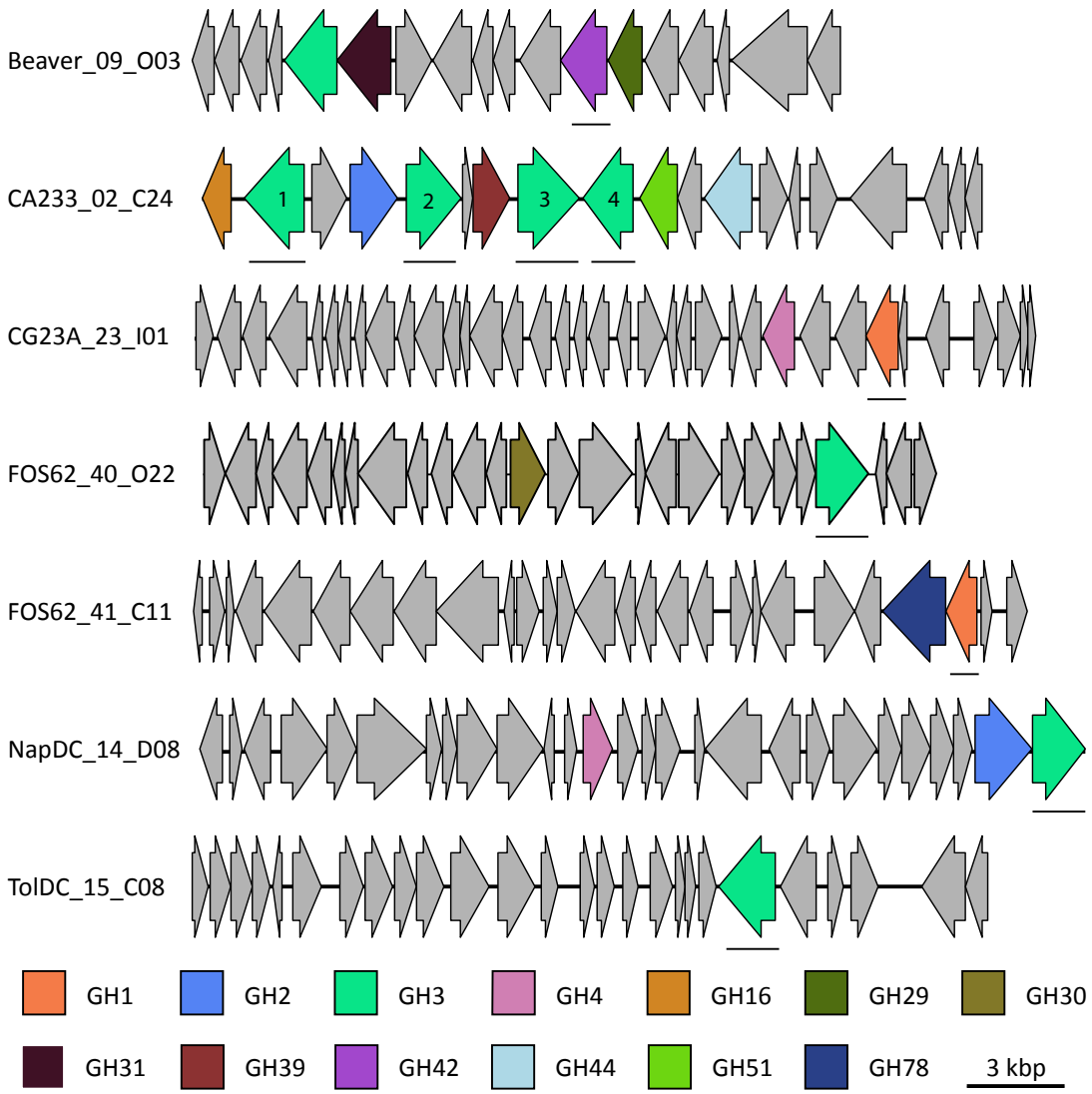

Figure 4.5: *Gene Organisation of Selected Fosmids With Activity on Modified Glycosides.* Putative glycoside hydrolases are coloured with the same scheme as Figure 3.4. ORFs not annotated as a glycoside hydrolase are shown in grey. ORFs Selected for sub-cloning and further characterization are underlined. As there were multiple GH3s selected for sub-cloning from fosmid CA233_02_C24 these were numbered 1-4.

Table 4.3: Kinetic Constants for Fosmid Sourced Hydrolases.

Fosmid	Enzyme	Substrate	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M ($mM^{-1}s^{-1}$)
FOS62_41_C11	C11-GH1	MU-Glc	63±5	0.20±0.03	320±50
		MU-6-N_3-Glc	14±1	0.05±0.01	280±60
		MU-Gal	0.9±0.2	0.05±0.02	20±9
		MU-6-N_3-Gal	3.9±0.5	0.10±0.02	40±10
NapDC_14_D08	D08-GH3	MU-Glc	0.07±0.005	0.064±0.009	1.1±0.2
		MU-4-N_3-Glc	-	-	0.013±0.002
		MU-4-NH_2-Glc	-	-	32.1±0.4
Beaver_09_O03	O03-GH42	MU-Gal	-	-	95±3
		MU-6-N_3-Gal	-	-	3.2±0.1
		MU-3-O-Me-Gal	-	-	4.4±0.2
TolDC_15_C08	C08-GH3	MU-Glc	2.7±0.2	0.039±0.005	70±10
		MU-3-NH_2-Glc	0.017±0.005	0.5±0.2	0.03±0.01
FOS62-40-O22	O22-GH3	MU-Glc	1.10±0.05	0.050±0.005	20±2
		MU-6-N_3-Glc	1.4±0.2	0.24±0.05	5±1
		MU-6-NH_2-Glc	-	-	No Activity
CA233_02_C24	C24-GH3-1	MU-Glc	0.025±0.001	0.0012±0.0003	20±5
		MU-6-N_3-Glc	0.37±0.03	0.06±0.01	6±1
		MU-6-NH_2-Glc	0.00161±0.00004	0.0029±0.0006	0.6±0.1
		MU-4-N_3-Glc	-	-	0.01±0.0007
		MU-4-NH_2-Glc	2.5±0.2	0.32±0.04	8±1
CA233_02_C24	C24-GH3-2	MU-Glc	0.00068±0.00006	0.004±0.002	0.15±0.08
		6-N_3-Glc	0.012±0.004	0.13±0.07	0.09±0.06
		MU-4-NH_2-Glc	-	-	No Activity
		MU-4-N_3-Glc	-	-	No Activity
CA233_02_C24	C24-GH3-4	MU-Glc	-	-	0.83±0.02
		MU-6-N_3-Gal	-	-	0.0018±0.0003
		MU-4-NH_2-Glc	-	-	0.0048±0.0001
		MU-4-N_3-Glc	-	-	No Activity
CG23A_23_I01	I01-GH1	pNP 6-PO_4-Glc	16±2	0.04±0.01	500±100
		pNP-Glc	-	-	No Activity
		MU-Glc	-	-	No Activity
		MU-3-N_3-Glc	-	-	No Activity
		MU-3-NH_2-Glc	-	-	No Activity

A majority of the activities detected from the initial screening were confirmed to be a result of the sub-cloned and expressed proteins. The GH42 from Beaver_09_O03 was confirmed to cleave MU-3-O-Me-Gal, however this activity was an order of magnitude less than that seen for the galactoside. The GH1 from FOS62_41_C11 had the suspected activity against 6-N_3-Gal MU, with a specificity constant (k_{cat}/K_M) on the same order of magnitude as the unmodified glycoside. Additionally, C11-GH1 cleaved MU-6-N_3-Glc and the unmodified glucoside, with specificity constants an order of magnitude greater than for the corresponding galactosides. D08-GH3 had a detectable, but low activity on the 4-azido glucoside, however the specificity constant for the 4-amino glucoside was, surprisingly, an order of magnitude greater than that of the unmodified glucoside. The GH3 expressed from TolDC_15_C08 was able to hydrolyse the 3-amino glucoside, however the K_M value for this hydrolysis was much larger than that seen for the unmodified glucoside.

As CA233_02_C24 contained several genes which may have been responsible for the observed activity, each of the purified proteins was assayed against all four of the modified glucosides that the fosmid clone cleaved. C24-GH3-2 and C24-GH3-4 both cleaved MU-6-N_3-Glc, however, specificity constants were substantially lower than those observed for C24-GH3-1. The C24-GH3-2 enzyme cleaved none of the other modified glycosides with which it was interrogated. C24-GH3-4 cleaved the 4-amino glucoside, however this activity again paled in comparison to the activity displayed by C24-GH3-1. Furthermore, C24-GH3-1 also cleaved the 4-azido glycoside, albeit slowly, as well as the 6-amino glucoside, indicating that this enzyme alone is sufficient to explain the activity seen in the initial screen.

Two of the expressed proteins did not hydrolyse the expected modified glycosides. The first of these, the GH3 from FOS62-40-O22, was expected to hydrolyse both the 6-amino and 6-azido glucosides. However, O22-GH3 only cleaved the azido-substituted glucoside and not the amino-substituted glucoside. This fosmid also contains a GH30, a family known to contain β-glucosidases, which may have been responsible for the activity on the 6-amino glycoside. The GH1 from CG23A_23_I01 was expected to cleave both the 3-amino and 3-azido glucosides. However, neither of the 3- modified glycosides nor MU-Glc were hydrolysed by I01-GH1, even after a prolonged (18 hour) incubation. As mentioned earlier, this fosmid also contains PTS IIA, IIB and IIC genes directly upstream and on the same DNA strand as the I01, suggesting that these genes are co-

expressed. We then decided to test I01-GH1 with pNP 6-phospho-β-D-glucoside. Indeed, I01-GH1 was able to catalyze hydrolysis of the 6-phospho-β-D-glucoside. It is therefore likely that the amino- and azido-glucosides must first be phosphorylated before hydrolysis. Attempts were made to phosphorylate MU-3-N_3-Glc and MU-3-NH_2-Glc with an ATP dependent β-glucoside kinase (BglK) [296] from *Klebsiella pneumonia*, however, neither were phosphorylated by this enzyme. The presence of a GH4 on the CG23A_23_I01 fosmid also obfuscates the observed activity. The GH4 family contains members with activity on 6-phospho-β-glucosides, however this family employs an unusual mechanism involving reduction and elimination steps, which is initiated by oxidation of the 3-hydroxyl [333]. Since this is not possible for the 3-azido-glucoside it is difficult to ascribe the activity on the 3-azido-glucoside to the GH4.

4.3.3 Acceptor Specificity

As we hoped to use the selected enzymes for synthesis, through the generation of glycosynthases, it was pertinent to identify the range of possible acceptors. To probe the enzyme acceptor specificity we followed the method developed by Blanchard et. al. [29]. This method is based upon screening of relative rates of reactivation, through transglycosylation, of a stabilised, but catalytically competent, glycosyl enzyme intermediate. This method employs a mechanism-based inactivator, a 2-deoxy-2-fluoro-glucoside (2-F-Glc) or 2-deoxy-2-fluoro-galactoside (2-F-Gal) bearing an activated leaving group, which forms a covalent intermediate with the enzyme of interest. Once the enzyme is inactivated excess inactivator is removed and the inactivated enzyme is subsequently incubated in the presence of several different potential reactivators. After a set period of time, reactivation is assessed by assaying the enzyme with a chromogenic or fluorogenic substrate. The reactivated enzyme is then compared to both the un-inhibited enzyme and a control in which no reactivator was used to assess the ability of a molecule to reactivate the enzyme, see Figure 4.2.

Acceptor specificity assays were performed with 6 of the 9 purified enzymes, C24-GH3-2 and C24-GH3-4 were excluded as C24-GH3-1 had a wider range of activity on modified substrates, and I01-GH1 was excluded as this was not inactivated with 2,4-dinitrophenyl 2-deoxy-2-fluoro-glucoside. The six enzymes that were interrogated were assayed in a plate-based format and incubated with a set of 87 potential reactivators, which included thiols, alcohols, glycosides, free sugars, and amino

acids. As O03-GH42 displayed a preference for galactosides as opposed to glucosides, accordingly, this enzyme was inhibited with DNP 2-F-Gal, as opposed to DNP 2-F-Glc and reactivation was assessed with MU-3-O-Me-Gal as opposed to pNP-Glc, which was used for the 5 other enzymes.

Table 4.4: Acceptor Specificity of Selected Wild-Type Hydrolases.

Acceptor	C08-GH3	O22-GH3	C24-GH3-1	D08-GH3	C11-GH1	O03-GH42
No Inhibitor	100	100	100	100	100	100
No Acceptor	10.0	26.5	17.3	8.7	0.4	52.7
Phenyl β-D-galactoside	-	**54.0**	22.0	-	-	-
Phenyl β-D-glucoside	-	-	-	-	0.8	-
pNP α-D-galactoside	69.5	-	-	14.7	1.0	-
pNP α-D-xyloside	30.5	-	24.9	12.0	2.4	-
pNP β-D-fucoside	30.3	-	-	-	1.1	-
pNP β-D-galactoside	49.2	36.9	22.2	13.0	1.9	-
pNP β-D-glucoside	**86.6**	30.2	42.2	**18.6**	**17.2**	-
pNP β-D-glucuronide	75.6	-	-	11.7	-	-
pNP β-D-mannoside	-	-	21.5	-	-	-
pNP β-D-xyloside	28.6	-	-	-	3.7	-
Cellobiose	-	-	-	-	0.7	-
Gentiobiose	-	-	-	-	0.8	-
Xylose	-	-	-	-	0.7	-
2-Mercaptoethanol	-	-	23.0	10.4	0.7	62.9
1-Hexanol	-	-	-	-	-	62.7
1-Pentanol	-	-	-	-	-	62.9
2-methoxyethanol	-	-	21.3	10.1	-	71.0
2-Phenylphenol	-	-	-	-	-	70.4
3-Mercapto-1-propanol	-	-	35.1	13.6	2.1	**77.4**
Ethanediol	-	28.2	21.7	-	-	60.6
Galactal	-	-	**80.5**	-	-	-
Methanol	-	-	24.5	-	-	-
Phenethyl alcohol	-	37.5	32.5	12.8	-	61.1

Rates are as a % of un-inhibited enzyme and only acceptors with a z-score > 3 in comparison to the no-acceptor control are shown. The top reactivator for each enzyme is bolded.
- : not a significant reactivator

Screening revealed distinct acceptor profiles for each of the enzymes assayed, see Table 4.4. In total twenty-four molecules were able to reactivate at least one inhibited enzyme faster than water alone. Many of the top reactivators were aryl-glycosides, with pNP-Glc being the top reactivator for C08-GH3, D08-GH3 and C11-GH1, while phenyl galactoside was the top reactivator for O22-GH3. Although C24-GH3-1 also was reactivated by aryl glycosides, including pNP-Glc and pNP-Gal, this enzyme was reactivated by a number of alcohols and its best reactivator was galactal.

O03-GH42 had a reactivator profile that was quite different from the other enzymes, as thiols and alcohols appeared to be preferred to aryl glycosides. Additionally, as this enzyme displayed rapid reactivation without any added acceptors, I also observed hydrolysis of *p*-nitrophenyl β-D-galactoside, *p*-nitrophenyl α-L-arabinoside and *p*-nitrophenyl β-D-fucoside.

4.3.4 Nucleophile Mutant Creation and Glycosynthase Tests

With both the donor and acceptor specificity information in hand, we next sought to create nucleophile variants of each enzyme and test for glycosynthase activity. The nucleophile residue of each of the wild-type enzymes was identified through multiple sequence alignment with well-characterized enzymes from the same family. The codon coding for the nucleophilic aspartate (in the case of C08-GH3, C24-GH3-1, D08-GH3 and O22-GH3) or glutamate (in the case of C11-GH1, I01-GH1 and O03-GH42) was mutated to a codon for either an alanine, serine or glycine. The 21 mutant genes were then transformed into an expression strain and expressed as for their cognate wild-type enzymes.

Initial glycosynthase tests were performed with the top three hits from the acceptor specificity test. The donors used were α-F-Glc (C08-GH3, D08-GH3, O22-GH3 and C24-GH3-1), α-F-Gal (O03-GH42 and C11-GH1) or α-F-6-PO_4Glc for I01_GH1, which was generated *in situ*. Of all nucleophile variants tested, only the C11-GH1 enzymes had any observable glycosynthase activity. The C11-GH1_E354S variant in particular was capable of transferring α-F-Gal onto pNP-Glc, pNP-β-Xyl and pNP-α-Xyl as determined by thin layer chromatography and mass spectrometry. C11-GH1_E354S was also capable of using both 6-N_3-α-F-Glc and 6-N_3-α-F-Gal as glycosynthase acceptors.

The lack of glycosynthase activity for a majority of the nucleophile variants led us to question how to improve the catalysts to obtain active glycosynthases. To date two GH3 enzymes have been successfully transformed into glycosynthases. The first of these EryBI D257G was able to catalyse the glucosylation of erythromycin, however the yields obtained were quite low, 14% [138]. The second active GH3 glycosynthase was derived from a thermostable β-glucosidase from *Thermotoga neapolitana* [244]. Nucleophile variants of this enzyme, TnBgl3B, were unable to catalyse glycosynthase reactions. However, when an additional mutation, W243F which had previously been seen

to result in increased transglycosylation of another GH3 β-glucosidase, [269] was introduced into the original nucleophile variants competent glycosynthases were created.

Inspired by this double mutation strategy we hoped that the introduction of a phenylanlanine in the place of the analogous tryptophan in the C08-GH3, C24-GH3-1, D08-GH3 and O22-GH3 enzymes could result in active glycosynthases. This was also a possibility as all four of these enzymes had a conserved tryptophan residue directly C-terminal to the active site nucleophile aspartate, in the same position as TnBgl3B. Mutagenesis was performed in the nucleophile serine variant for all four genes, resulting in double active site variant genes (O22_GH3_D231S_W232F, D08_GH3_D229S_W230F, C24_GH3_1_D271S_W272F and C08_GH3_D235S_W236F). The corresponding proteins were purified as for the wild-type enzymes. The four double variant proteins were then assayed as previously using 10 mM donor and acceptors. However, unlike the results obtained by Pozzo et. al. [244] no glycosynthase activity was observed, suggesting that mutation of the active site tryptophan isn't a general strategy for the creation of GH3 glycosynthases.

4.3.5 Product Characterization

To further characterize the glycosynthase products generated by C11_E354S I performed the reactions on a multi-milligram scale. These reactions were performed with 50 μmol of donor sugar and 250 μmol of acceptor. In total four different multi-milligram scale reactions were carried out, see Table 4.5. C11_E354S was able to catalyse the galactosylation of both pNP α- and β-D-xylopyranoside, and pNP β-D-glucopyranoside.

The Galactosylation of β-D-glucopyranoside resulted in both the 1,3- and 1,2-linked products. GH1 glycosynthase production of 1,3-linked pNP galactosylglucoside has been previously observed for both Abg_E358A [188] and Bgl3_E383A [87], however the 1,2-linked product has not been observed previously. Galactosylation of the xylosides resulted in 1,2-linkages when either pNP-α-Xyl or pNP-β-Xyl were used. The product containing the α-xyloside is particularly interesting as this could be used as a model substrate for xyloglucan decorations. The galactosyl xylosides produced here have different linkages than those produced by either Abg_E358A [188] or Bgl3_E383A [87]. Both these enzymes are able to catalyse the galactosylation of pNP-β-xyl, however in both cases the major regiochemical outcome was the 1,3-linked product.

We were also able to use C11_E354S to attach 6-azido-modified α-galactosyl fluoride. The glycosynthase reaction between 6-N_3-α-F-Gal and pNP-Glc resulted in three separate products, with similar yields, see Table 4.5. The 1,2-, 1,3- and 1,4-linked products were all observed with the 1,3-glycoside being the major product. This is somewhat surprising as the 1,4-linked product was not observed when the unmodified galactoside donor was used in a similar reaction. The regiochemical outcome is thus influenced by the presence of the 6-azido functional group. Taken together these results demonstrate the ability of C11_E354S to glycosylate with azido-modified donors and future research should focus on the scope of molecules that can act as competent acceptors.

Table 4.5: Stereochemical Outcome and Yield of C11_E354S Glycosynthase Reactions.

Enzyme	Donor	Acceptor	Product	Yield (%)
C11_E354S	α-F-Gal	pNP-Glc	Gal-(β-1,2)-Glc-β-pNP	15
			Gal-(β-1,3)-Glc-β-pNP	20
C11_E354S	α-F-Gal	pNP-α-Xyl	Gal-(β-1,2)-Xyl-α-pNP	50
C11_E354S	α-F-Gal	pNP-β-Xyl	Gal-(β-1,2)-Xyl-β-pNP	60
C11_E354S	6-N_3-α-F-Gal	pNP-Glc	6-N_3-Gal-(β-1,4)-Glc-β-pNP	23
			6-N_3-Gal-(β-1,2)-Glc-β-pNP	21
			6-N_3-Gal-(β-1,3)-Glc-β-pNP	37

4.4 Glycoside Hydrolase Family 1 Library

To further explore the prevalence of promiscuous hydrolase activities we exploited a library of 175 GH1 enzymes synthesized and characterized by Heins et. al. [117]. The genes within this library were chosen to maximize sequence diversity and are from eukaryal, archaeal, bacterial and metagenomic sources. The GH1 family contains members with activity on many different glycosides, yet the most abundant activity observed by Heins et. al. [117] was the hydrolysis of β-glucosides (59 of 105 expressed and purified enzymes). This abundance of β-glucosidases was the reason we selected this library for interrogation with modified β-glucosides. Additionally, the GH1 family contains many examples of successful glycosynthases [80, 87, 188, 236, 237, 243, 299], including C11_E354S detailed previously, implying that the hydrolases within this library can be converted to glycosynthases with good success rates.

4.4.1 Screening with Modified Glycosides

The GH1 library was screened, as for the fosmid library, with six substrates bearing a fluorogenic 4-methylumbelliferyl leaving group and either an azido or amino group at the 3, 4, or 6 position. Clones were also screened with MU-Glc to determine the number of β-glucosidases that could be detected from crude lysate. Screening revealed that 115 of the 175 GH1 enzymes cleaved the unmodified substrate MU-Glc (z-score > 9), a much higher number of active enzymes than observed by Heins et. al. [117], Figure 4.6. This increased number of active enzymes likely reflects the increased reaction time (18 hours in this study, 10 mins for Heins et al.[117]) and an increased enzyme concentration for those that had not been purified in significant concentrations.

Nearly two thirds of the clones tested on modified glucosides cleaved at least one substrate (106 of 175 with a z-score > 9), and 91 cleaved more than one substrate (Figure 4.7). Four of these clones (Genbank ID: BAJ01494.1, ABS04001.1, BAA74959.1, CAL97639.1) cleaved MU-6-N_3-Glc, yet had no activity on the parent MU-Glc. All other clones with activity on modified glucosides also cleaved the parent MU-Glc. In general a greater number of active clones were seen for the amino substituted glucosides (MU-4-NH_2-Glc: 99/175, MU-6-NH_2-Glc: 83/175, MU-3-NH_2-Glc : 64/175) than the azido substituted glucosides (MU-6-N_3-Glc: 91/175, MU-3-N_3-Glc: 2/175, MU-4-N_3-Glc:

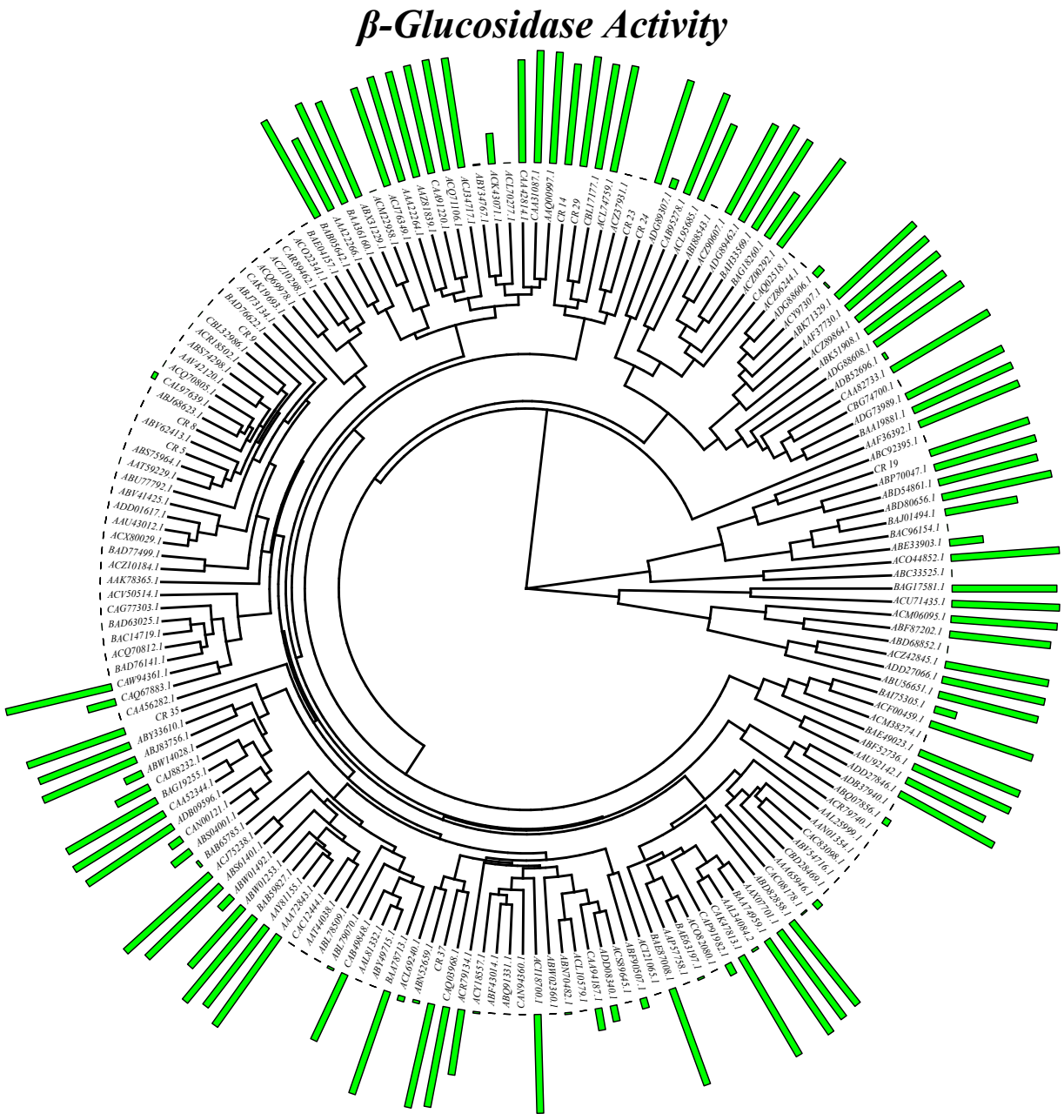

Figure 4.6: *GH1 Enzyme Library β-Glucosidase Activity.* Relative fluorescent intensity is represented by bars, with each leaf corresponding to a protein. Fluorescent values are given as the fraction of the maximum expected fluorescence. Genbank IDs are given at the tip of each leaf.

1/175). It is also worth noting that the absolute fluorescence for hits identified with the 3- and 4-substituted azido sugars was extremely low when compared to other hits and had fluorescence values less than 1 % of the anticipated maximum.

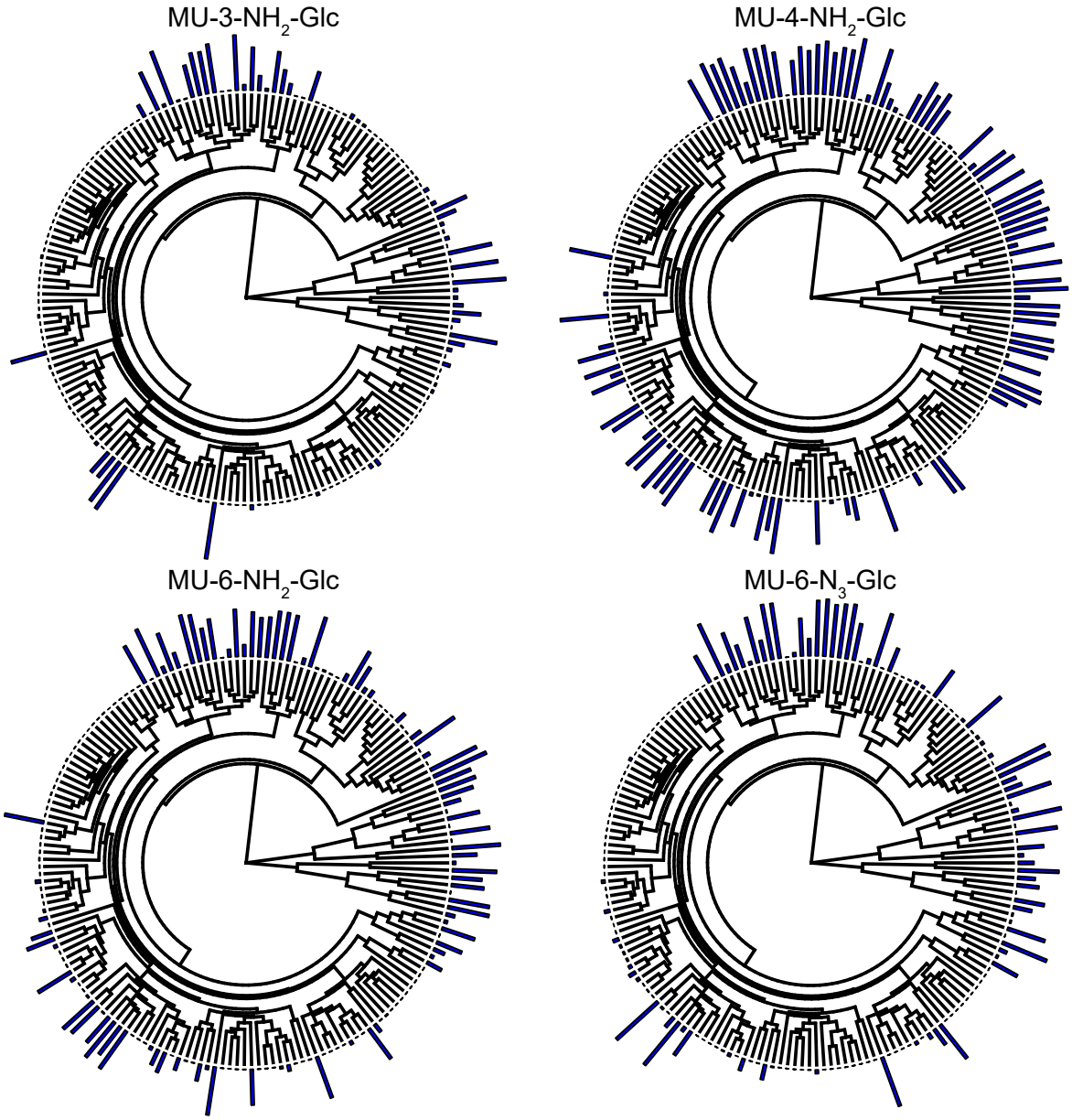

Figure 4.7: *Screening Results.* Relative fluorescence intensity is represented by bars, with each leaf corresponding to a protein. Fluorescent values are given as the fraction of the maximum expected fluorescence. Results for MU-3-N_3-Glc and MU-4-N_3-Glc are not shown.

The finding of relatively few enzymes that are capable of cleaving the MU-3-N_3-Glc and MU-4-N_3-Glc substrates is a reflection both of the considerable steric demand of an azide substituent compared to an amine or the parent hydroxyl, as well as of the importance of the hydrogen bonds normally formed between the enzyme and the substrate at those positions. The key residues involved in interactions with the 3- and 4-hydroxyls, His123, Gln20, Trp423 and Glu422 (numbering based on *Phanerochaete chrysosporium* GH1 [219]) are highly conserved, see Figure 4.8. The 6-azido substituent, however, is reasonably well tolerated, most likely because of the greater conformational flexibility possible at the 6-position, allowing the substituent to adopt an orientation that minimises steric repulsion. The readier acceptance of an amine substituent than an azide at C3 or C4 likely stems from its small size, as well as its hydrogen bonding potential. Likewise, amine substitution at the 6 position also seems to be broadly tolerated. The hit rates seen for the amino glucosides, in fact, nicely mirror the specificities determined for the GH1 β-glucosidase Abg by Namchuk and Withers [213] through measurement of kinetic parameters for hydrolysis of a set of mono-deoxyglycoside substrates in which each hydroxyl, individually, had been replaced by hydrogen. From these data, the contributions of interactions with each hydroxyl to transition state stabilization were determined, yielding $\Delta\Delta G^{o\ddagger}$ values of 7.4, 2.5 and 2.9 kJ/mol for the 3-, 4- and 6-hydroxyls, respectively, very much in line with the lower tolerance seen here for substitution at C3.

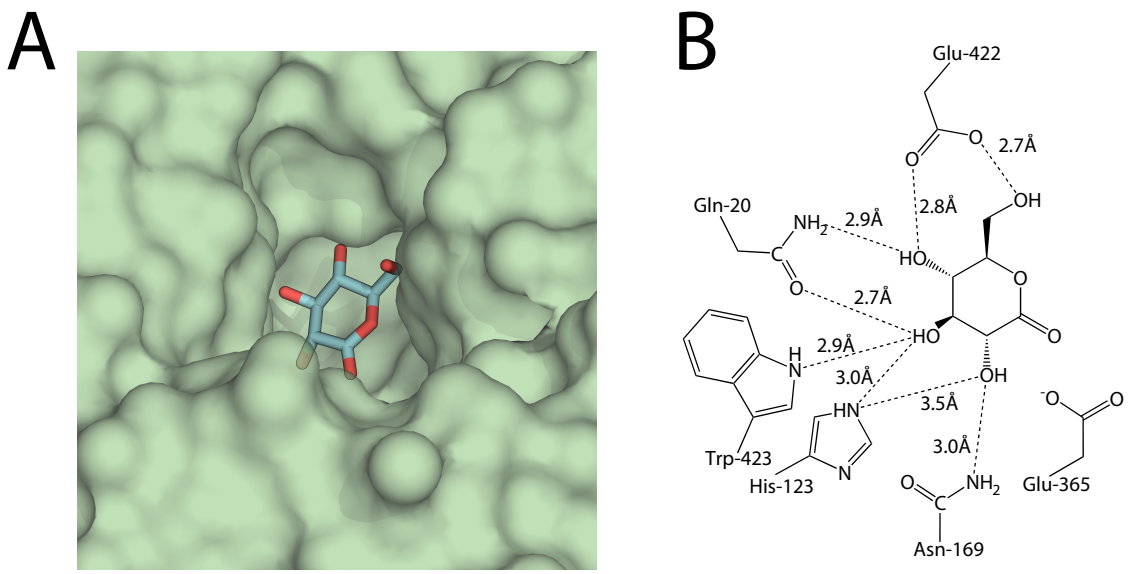

Figure 4.8: *Pha_GH1 in Complex With Gluconolactone and Substrate-Protein Bond Distances.* **A** The crystal structure of Pha_GH1 (PDB:2E40) in complex with gluconolactone clearly shows the 6-hydroxyl of the gluconolactone pointed toward the exterior of the tunnel like active site. The view shown is directed into the active site. Carbon atoms of gluconolactone are shown in green and oxygen atoms are shown in red. **B** Bond distances between the hydroxyls of gluconolactone and the conserved active site residues are shown. Crystal structure was determined by Nijikken and coworkers [219].

4.4.2 Kinetic Characterization of Hydrolases

From the 106 hits so identified we chose 8 enzymes, which between them encompassed all the activities sought, as candidates for transformation into glycosynthases (Table 4.6). As a first step, kinetic parameters were measured for each modified substrate with each of these purified wild-type enzymes (Table 4.7) with the exception of the Lac enzyme, for which activity on MU-3-NH_2-Glc, MU-4-N_3-Glc and MU-6-N_3-Glc was below detectable levels. Since this enzyme is primarily a 6-phospho-beta-glucosidase, it is not surprising that it has such sparing activity on non-phosphorylated modified glucosides. In many cases K_M values were too high to be reliably measured, so values of k_{cat}/K_M, the specificity constant, were determined instead through measurements at low substrate concentrations. In general, for those enzyme/substrate pairs where cleavage

was detected, higher specific activities were seen with 6-modified substrates (Table 4.7). However, in two cases k_{cat}/K_M values measured with 6-modified substrates were higher than for the parent glucoside. Also, interestingly, in some cases the 6-azidoglucoside was cleaved more rapidly than the 6-amino, and in others the reverse was seen. Additionally, all enzymes capable of cleaving both 4-amino and 3-amino glucosides displayed higher specific activities towards the 4-amino substrates. This again is very much in agreement with the specificity studies on Abg noted previously.

Table 4.6: Selected GH1 Genes and Their Activities.

			Modified Glucans						
GenbankID	Enzyme	Reference	MU-Glc	3-NH_2	4-NH_2	6-NH_2	3-N_3	4-N_3	6-N_3
AAZ81839.1	Ali_GH1	[168]	✓	✓	✓	✓	-	-	✓
ACO44852.1	Dei_GH1	-	✓	✓	✓	✓	-	-	✓
ACQ71106.1	Exi_GH1	-	✓	✓	✓	✓	✓	-	✓
CAQ67883.1	Lac_GH1	-	✓	✓	✓	-	✓	✓	-
ABF87202.1	Myx_GH1	-	✓	✓	✓	✓	-	-	✓
BAE87008.1	Pha_GH1	[301]	✓	-	✓	✓	-	-	✓
ABD82858.1	Sac_GH1	-	✓	✓	✓	✓	-	-	✓
AAF37730.1	The_GH1	[282]	✓	-	-	-	-	-	✓

4.4.3 Acceptor Specificity

To gain insight into the specificity of the +1 subsite, each of the 8 wild-type enzymes was subjected to acceptor specificity screening as was done for the metagenomic hits. All 8 enzymes were screened against a panel of 83 potential acceptors, including a variety of glycosides, free sugars and alcohols. As can be seen in Figure 4.9, each enzyme displayed a different pattern of acceptor specificity. The extent of reactivation of each enzyme also differed, with the maximum rates of reactivation for Ali_GH1, Pha_Gh1 and The_GH1 being fairly modest (0.4%, 2.1 % and 3.5% of the uninhibited rate, respectively) when compared to those of Dei_GH1, Exi_GH1, Lac_GH1, Myx_GH1 and Sac_GH1 (19.7%, 83.8 %, 59.4%, 47.2 % and 100% of the uninhibited rate, respectively). The majority of the best reactivators were aryl glycosides, with six of the eight enzymes being reactivated fastest by an aryl glucoside (pNP-Glc, pNP-α-Glc or MU-α-Glc). The exceptions to this were Pha_GH1, for which cellobiosides were best – suggesting a strong preference for cello-oligosaccharide acceptors, and Sac_GH1, which reactivated fastest with pNP β-D-fucopyranoside.

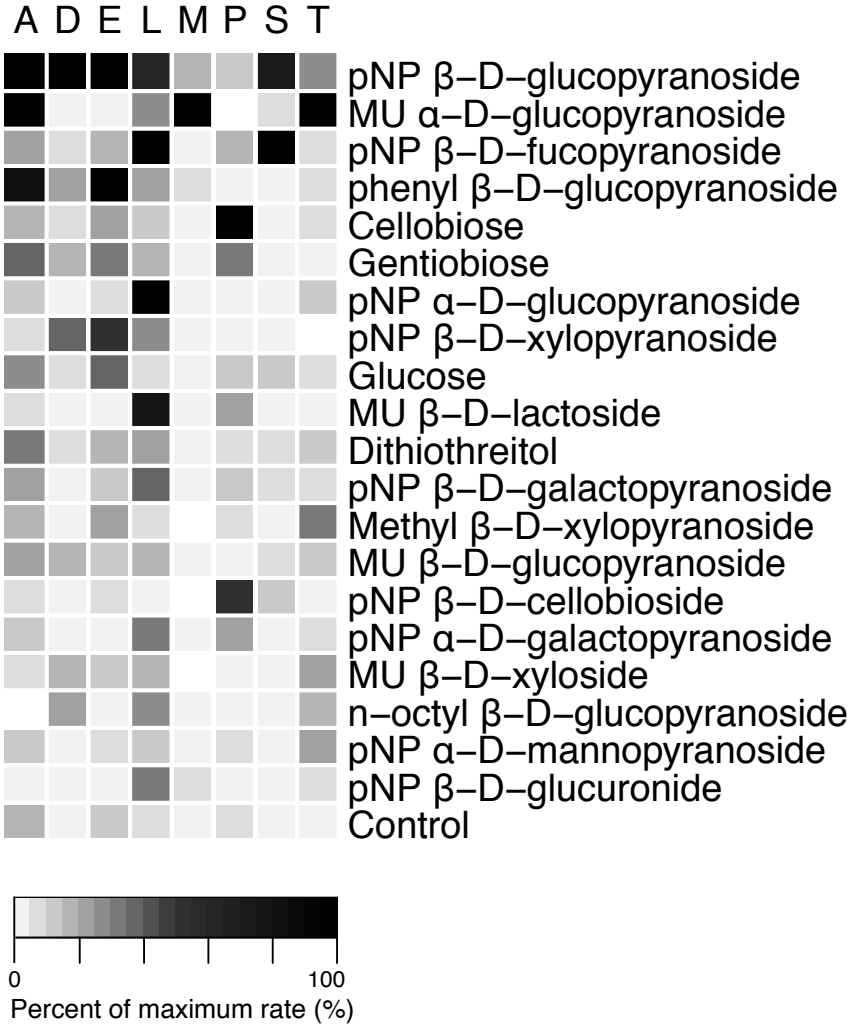

Figure 4.9: *GH1 Acceptor Specificity.* The top five reactivators and the relative initial rates observed are shown for each of the wild-type enzymes. Enzymes screened are abbreviated as follows: (A) Ali_GH1, (D) Dei_GH1, (E) Exi_GH1, (L) Lac_GH1, (M) Myx_GH1, (P) Pha_GH1, (S) Sac_GH1, and (T) The_GH1. Rates are scaled to the best reactivator for each given enzyme. The control is the activity seen when no reactivator was included.

Table 4.7: Kinetic Parameters for Selected GH1s

Enzyme	Substrate	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (s^{-1}, mM^{-1})
Ali_GH1	MU-Glc	28 ± 1	0.13 ± 0.01	220 ± 20
	MU-3-NH_2-Glc	-	-	$9 \times 10^{-4} \pm 2 \times 10^{-4}$
	MU-4-NH_2-Glc	-	-	2.8 ± 0.6
	MU-6-NH_2-Glc	-	-	$1.60 \times 10^{-1} \pm 6 \times 10^{-3}$
	MU-6-N_3-Glc	0.32 ± 0.04	0.04 ± 0.01	8 ± 3
Dei_GH1	MU-Glc	14 ± 1	0.13 ± 0.03	100 ± 30
	MU-3-NH_2-Glc	-	-	$7.5 \times 10^{-2} \pm 2 \times 10^{-3}$
	MU-4-NH_2-Glc	-	-	3.31 ± 0.05
	MU-6-NH_2-Glc	61 ± 2	0.20 ± 0.01	310 ± 20
	MU-6-N_3-Glc	-	-	$7.45 \times 10^{-1} \pm 1 \times 10^{-3}$
Exi_GH1	MU-Glc	46 ± 4	0.018 ± 0.009	3000 ± 1000
	MU-3-NH_2-Glc	0.0006 ± 0.0001	0.012 ± 0.001	0.48 ± 0.05
	MU-4-NH_2-Glc	-	-	8.4 ± 0.6
	MU-6-NH_2-Glc	117 ± 6	0.07 ± 0.01	1700 ± 300
	MU-3-N_3-Glc	-	-	$2.2 \times 10^{-5} \pm 3 \times 10^{-6}$
	MU-6-N_3-Glc	5.9 ± 0.7	0.24 ± 0.05	25 ± 6
Lac_GH1	pNP 6-PO_4-Glc	2.04 ± 0.04	1.34 ± 0.05	1.52 ± 0.06
	pNP Glc	-	-	$1.2 \times 10^{-4} \pm 2 \times 10^{-5}$
	MU-Glc	-	-	$8.70 \times 10^{-2} \pm 9 \times 10^{-4}$
	MU-3-NH_2-Glc	-	-	N.D.
	MU-4-NH_2-Glc	-	-	$1.05 \times 10^{-2} \pm 3 \times 10^{-4}$
	MU-4-N_3-Glc	-	-	N.D.
	MU-6-N_3-Glc	-	-	N.D.
Myx_GH1	MU-Glc	5.1 ± 0.2	$2.5 \times 10^{-3} \pm 7 \times 10^{-4}$	2000 ± 600
	MU-3-NH_2-Glc	-	-	$2 \times 10^{-3} \pm 1 \times 10^{-3}$
	MU-4-NH_2-Glc	-	-	1.05 ± 0.09
	MU-6-NH_2-Glc	4.5 ± 0.4	0.020 ± 0.007	220 ± 80
	MU-6-N_3-Glc	-	-	N.D.
Pha_GH1	MU-Glc	20.2 ± 0.8	0.073 ± 0.009	280 ± 40
	MU-4-NH_2-Glc	-	-	1.24 ± 0.04
	MU-6-NH_2-Glc	-	-	0.95 ± 0.04
	MU-6-N_3-Glc	1.40 ± 0.06	$2.0 \times 10^{-3} \pm 4 \times 10^{-4}$	700 ± 200
Sac_GH1	MU-Glc	9 ± 1	0.07 ± 0.02	160 ± 40
	MU-3-NH_2-Glc	-	-	$3.0 \times 10^{-3} \pm 9 \times 10^{-5}$
	MU-4-NH_2-Glc	-	-	$3.2 \times 10^{-2} \pm 2 \times 10^{-3}$
	MU-6-NH_2-Glc	-	-	$4.6 \times 10^{-3} \pm 2 \times 10^{-4}$
	MU-6-N_3-Glc	-	-	0.74 ± 0.03
The_GH1	MU-Glc	10 ± 1	0.07 ± 0.02	160 ± 50
	MU-3-NH_2-Glc	-	-	N.D.
	MU-4-NH_2-Glc	0.13 ± 0.01	0.29 ± 0.04	0.44 ± 0.07
	MU-3-N_3-Glc	-	-	N.D.
	MU-6-N_3-Glc	-	-	0.68 ± 0.03

4.4.4 Nucleophile Mutant Creation and Glycosynthase Tests

We sought to make glycosynthases from each of the eight hits, with the hopes that they would be competent at transferring modified glucosides onto a variety of molecules. The conserved nucleophilic glutamate residue, for each enzyme, was mutated to three different amino acids (Serine, Alanine and Glycine) in the hopes that one of these would be an active glycosynthase. All 24 variant enzymes were expressed and purified on a 50 mL scale. Initially enzymes were tested for glycosynthase activity using α-glucosyl fluoride (αF-Glc, 50 mM) as a donor and para-nitrophenyl glucoside as an acceptor (pNP-Glc, 10 mM). For six (Ali_GH1, Dei_GH1, Exi_GH1, Myx_GH1, Sac_GH1, The_GH1) of the eight enzymes, at least one variant acted as a glycosynthase with this donor/acceptor combination. Mutants of the other two (Lac_GH1 and Pha_GH1) did not yield products. However, when Pha_GH1 nucleophile variants were incubated with αF-Glc and pNP-cellobioside (pNP-C), one of the best reactivators for the wild-type enzyme, products were indeed seen. As Lac_GH1 has a much higher specificity constant for the hydrolysis of 6-phospho-glucosides when compared to that seen for glucosides (Table 4.7), we suspected that the nucleophile variants would be competent glycosynthases with 6-phospho-glucosyl donors. Unfortunately none of the Lac_GH1 variants had observable catalytic activity with 6-phospho-α-glucosyl fluoride in conjunction with any of the top five reactivators.

To choose the best glycosynthase variant from the three different variants we performed HPLC analysis of small scale reactions. Reaction mixtures contained an equal amount of donor and acceptor (αF-Glc, pNP-Glc or pNP-C at 5 mM). All glycine variants displayed hydrolytic activity, which we speculate is due to mis-incorporation of the wild-type glutamate, as has been reported previously for a GH1 glycosynthase [236]. In that study, mis-incorporation is seen for the gene containing the GGG codon for glycine, in our case we used GGA, however, in both of these cases the codons differ in only one base from that for glutamate (GAG, GAA). We suggest that future glycosynthase creation should utilise codons containing 2 substitutions (GGC, GGT) if the glycine variant is to be tested. Of the serine and alanine variants, the serine variant had the highest yield for the major product for all enzymes except for The_GH1 for which the alanine variant was selected. The differences between variants were, for most cases, within 5%, the only exception

being Sac_GH1, which had a 21 % higher yield for the serine variant than the alanine (Table 4.8).

Enzyme	Major Product Yield %
Ali_E354A	64 ± 0.2
Ali_E354S	65 ± 0.2
Dei_E346A	40 ± 0.1
Dei_E346S	45 ± 0.3
Exi_E350A	41 ± 0.2
Exi_E350S	54 ± 0.3
Myx_E357A	47 ± 0.4
Myx_E357S	52 ± 0.7
Sac_E368A	79 ± 0.9
Sac_E368S	100 ± 0.4
The_E388A	59 ± 0.7
The_E388S	55 ± 0.4
Pha_E365A	65 ± 1
Pha_E365S	70 ± 1

Table 4.8: Product Yields From Small Scale Glycosynthase Reactions.

4.4.5 Product Characterization

To identify the products of the most efficient glycosynthases, large-scale reactions were performed and products were purified by HPLC. The majority of the NMR experiments were performed by Dr. Feng Liu, and detailed chemical shift assignments are given in Appendix Section C.0.1. Initially large-scale glycosynthase reactions were carried out with α-F-Glc as acceptor and pNP-Glc or pNP-C as the donor. These products were then characterized by NMR and mass spectroscopy to reveal the glycosidic linkages (Table 4.9). Remarkably, a large set of different glycans was formed by the different glycosynthases despite both the donor and acceptor sugars being the same (except in the case of Pha_E365S). Of the seven competent glycosynthases, five selectively transferred a single sugar onto the acceptor, with Ali_E354S, Exi_E350S, and Myx_E357S preferentially forming β-1,3-linkages, while Pha_E365S and The_E388A preferentially formed β-1,4-linkages. The Sac_E368S glycosynthase also formed β-1,4-linkages, but this enzyme was also competent using the product as an acceptor to transfer an additional glucose, forming a trisaccharide. The final glycosynthase, Dei_E346S, also preferentially produced trisaccharides, but in this case the first transfer formed a β-1,3-linkage and the second a β-1,4 yielding the mixed trisaccharide (Glc-β-1,4-Glc-β-1,3-Glc-

β-1,4-pNP). This mixed-linkage product may be useful in dissecting the mechanism of hydrolases that function on mixed-linkage glucans.

Heins and coworkers performed a detailed characterization of the linkage specificity for a selected set of enzymes which included the Ali and The enzymes [117]. Both The_GH1 and Ali_GH1 had the fastest hydrolysis rates for laminaribiose (β-1,3-linked Glc-Glc), with sophorose hydrolysis second fastest for Ali_GH1 (β-1,2-linked Glc-Glc) and cellobiose (β-1,4-linked Glc-Glc) being second fastest for The_GH1. The glycosynthase product for Ali_GH1 was consistent with the hydrolysis rates, however The_E388A only synthesized β-1,4-linked products. Justification of this inconsistency may lie in the presence of a para-nitrophenyl-aglycone in the acceptor, which may interact with the +2 subsite, altering the acceptor orientation. Determining the regiochemical outcome of the reaction in which glucose as an acceptor would shed light on whether this is the case.

Enzyme	Donor	Acceptor	Product	Yield (%)
Ali_E354S	αF-Glc	pNP-Glc	Glc-(β-1,3)-Glc-β-pNP	65
			Glc-(β-1,4)-Glc-β-pNP	8
Dei_E346S	αF-Glc	pNP-Glc	Glc-(β-1,4)-Glc-(β-1,3)-Glc-β-pNP	45
			Glc-(β-1,4)-Glc-β-pNP	12
Exi_E350S	αF-Glc	pNP-Glc	Glc-(β-1,3)-Glc-β-pNP	54
			Glc-(β-1,4)-Glc-β-pNP	3
Myx_E357S	αF-Glc	pNP-Glc	Glc-(β-1,3)-Glc-β-pNP	52
			Glc-(β-1,4)-Glc-β-pNP	14
Sac_E368S	αF-Glc	pNP-Glc	Glc-(β-1,4)-Glc-(β-1,4)-Glc-β-pNP	100
Pha_E365S	αF-Glc	pNP-C	Glc-(β-1,4)-Glc-(β-1,4)-Glc-β-pNP	70
The_E388A	αF-Glc	pNP-Glc	Glc-(β-1,4)-Glc-β-pNP	59
			Glc-(β-1,4)-Glc-(β-1,4)-Glc-β-pNP	28
Sac_E368S	αF-3-NH_2-Glc	pNP-Glc	3-NH_2-Glc-(β-1,4)-Glc-β-pNP	74
Exi_E350S	αF-3-NH_2-Glc	pNP-Glc	3-NH_2-Glc-(β-1,3)-Glc-β-pNP	16
Sac_E368S	αF-4-NH_2-Glc	pNP-Glc	4-NH_2-Glc-(β-1,4)-Glc-β-pNP	63
The_E388A	αF-4-NH_2-Glc	pNP-Glc	4-NH_2-Glc-(β-1,4)-Glc-β-pNP	64
Sac_E368S	αF-6-NH_2-Glc	pNP-Glc	6-NH_2-Glc-(β-1,4)-Glc-β-pNP	37
Sac_E368S	αF-6-N_3-Glc	pNP-Glc	6-N_3-Glc-(β-1,4)-Glc-β-pNP	84
Exi_E350S	αF-6-N_3-Glc	pNP-Glc	6-N_3-Glc-(β-1,3)-Glc-β-pNP	42
			6-N_3-Glc-(β-1,4)-Glc-β-pNP	8
Pha_E365S	αF-6-N_3-Glc	pNP-C	6-N_3-Glc-(β-1,4)-Glc-(β-1,4)-Glc-β-pNP	27
Dei_E346S	αF-Glc	pNP-Xyl	Glc-(β-1,4)-Glc-(β-1,3)-Xyl-β-pNP	65
			Glc-(β-1,3)-Xyl-β-pNP	24
Exi_E350S	αF-Glc	pNP-Xyl	Glc-(β-1,3)-Glc-(β-1,3)-Xyl-β-pNP	50
			Glc-(β-1,3)-Xyl-β-pNP	49
Dei_E346S	αF-Glc	n-Octyl-Glc	Glc-(β-1,4)-Glc-(β-1,3)-Glc-β-Octyl	11
Sac_E368S	αF-Glc	DNP 2F-Glc	Glc-(β-1,4)-Glc-(β-1,4)-2F-Glc-β-DNP	28
			Glc-(β-1,4)-Glc-(β-1,4)-Glc-(β-1,4)-2F-Glc-β-DNP	21
Sac_E368S	αF-4-NH_2-Glc	DNP 2F-Glc	4-NH_2-Glc-(β-1,4)-2F-Glc-β-DNP	74

Table 4.9: Characterized GH1 Glycosynthase Products.

Enzyme	αF-3-NH_2-Glc	αF-4-NH_2-Glc	αF-6-NH_2-Glc	αF-6-N_3-Glc
Ali_E354S		✓	✓	
Dei_E346S	✓	✓	✓	
Exi_E350S	✓	✓	✓	✓
Myx_E357S	✓	✓	✓	
Pha_E365S	✓	✓	✓	✓
Sac_E368S	✓	✓	✓	✓
The_E388A		✓		

Table 4.10: Glycosynthase Activity with Azido and Amino Donor Sugars.

Having probed the general utility of the set of glycosynthases we turned our attention towards the modified glycosynthase donors. A range of α-glycosyl fluorides containing the same 3-, 4-, and 6-azido and amino modifications as those used to screen for hydrolase activity were synthesized by Dr. Hongming Chen to test as glycosynthase donors. Each of the seven competent glycosynthases was tested with each of the modified donor substrates, using either pNP-Glc or pNP-C as acceptor: all seven functioned with at least one modified donor (Table 4.10). Consistent with what had been seen in screening modified substrate activity, the 4-aminoglucosyl fluoride was accepted as a donor by all seven of these glycosynthases. The 3- and 6-aminoglucosyl fluorides were also accepted by many of the glycosynthases (5 of 7 and 6 of 7, respectively) corresponding fairly well with WT enzyme results on modified substrate. Three of the seven glycosynthases (Pha_E365S, Sac_E368S and Exi_E350S) were also able to transfer the 6-azidoglucosyl fluoride donor onto pNP-Glc or pNP-C, each producing a different azido-modified glucan (Table 4.9). Finally, none of the variants were able to carry out glycosyl transfers using the 3- or 4-azidoglucosyl fluorides as donors. This is not so surprising given the relatively low activities of these wild-type enzymes, along with the considerably lower activities of non-evolved glycosynthases relative to the wild-type parents carrying out the normal reaction.

Having established the donor and acceptor specificities of these glycosynthases we tested the ability of selected glycosynthases to generate useful conjugates of other glycans and non-sugar acceptors. Oligosaccharides of mixed sugar composition could be assembled, as demonstrated in the ability of Dei_E346S and Exi_E350S to transfer to pNP-xyloside, generating glucosyl-β-xylosides, with linkages that mirrored those seen when using pNP-G as acceptor (Table 4.9). Likewise, n-octyl β-glucoside served as an acceptor for Dei_E346S generating octyl oligosaccharides with

potential as detergents; addition of terminal aminosugars would allow simple assembly of cationic versions of these detergents. Within the sugar series, a particularly useful set of products are the mechanism-based inactivators generated by glycosynthase-catalyzed glycosylation of simple glucoside-based reactive entities such as 2,4-dinitrophenyl 2-deoxy-2-fluoro-β-glucoside (DNP 2-F-Glc). The Sac_E368S enzyme was selected for transfer onto this inhibitor – this enzyme had the highest transfer yields – resulting in the disaccharide version. More importantly, Sac_E368S was able to transfer a 4-aminoglucosyl moiety to create a disaccharide inhibitor bearing a functionalisable amine on the non-reducing end sugar (Table 4.9). This now allows facile, and mild derivatization of mechanism-based inhibitors and affinity labels via amide formation, allowing the attachment of fluorophores for detection, or of biotin for capture. Further, the amine could serve as the point of attachment of diverse substituents as a means of introducing novel specificity elements.

4.5 Discussion and Future Directions

The two clone libraries investigated in this study allowed for the interrogation of enzymatic promiscuity and creation of glycosynthases. The characteristics of the libraries allowed us to interrogate both the promiscuity across a range of different hydrolase-containing gene cassettes, and with a fine detail within one specific family. The identification of promiscuous genes in turn allowed for the creation of a panel of glycosynthases which can incorporate modified glucosides and galactosides.

The power of fosmid libraries lies in the ability to identify lengthy clusters of genes which may function synergistically. This however can also complicate the identification of the specific gene responsible for activity, especially when several potentially active genes are present. There are typically three ways forward to identify the genes responsible for activity: creation of a small insert libraries or knock-out libraries and sub-cloning the CAZymes. Small-insert libraries are created by first shearing the purified DNA, cloning these fragments into an expression vector then transforming this library into a suitable host. The re-screening of these small-insert libraries and subsequent sequencing should then reveal the gene(s) responsible for activity. Creating knock-out libraries involves integrating a selection marker randomly within a fosmid, then screening a library of clones with the integration for a loss of activity and finally sequencing the selected clones. These methods are feasible for small numbers of hits, but become intractable when tens or hundreds of fosmid clones are identified.

The ideal solution would be to create a library containing each of the potentially active genes from each fosmid in a high-expression vector. Creation of such a library would, without a doubt, be limited by the time consuming process of sub-cloning. However, technological advances may soon make this dream a reality. Development of laboratory automation equipment, such as the digital to biological converter [32], could allow for the rapid, automated sub-cloning of genes. Gene-synthesis automation coupled with decreasing costs, may soon allow for the realistic creation of libraries containing thousands of such clones at the click of a button. Technological advances should also enable the rapid creation of sets of phylogentically diverse enzymes from families other than GH1, allowing for a similar fine-detail characterization of promiscuous activity.

We were able to generate several new glycosynthases from promiscuous hydrolases, however,

several questions remain to be investigated. Are there additional mutations which will allow the creation of glycosythases from GH3 enzymes? How can we incorporate sterically bulky substituents at the 3- and 4-positions? How can we determine *a priori* whether an GH is a good candidate for transformation into a glycosynthase?

The first of these questions may be solved by directed evolution. This may be accomplished by using a screen similar to that performed by Kim et al [153] where detection of glycosynthase activity was coupled to the activity of an enzyme (Cel5A) that was able to cleave the glycosynthase product, but not the reactants. Subjecting the metagenomically identified promiscuous GH3s to this process should enable the evolution of glycosynthases and the identification of mutations which support this catalysis. Directed evolution of multiple GH3s in parallel may reveal mutations that universally support glycosynthase transformation for all GH3s.

We were able to use glycosynthases to incorporate amino modifications at the 3-, 4- and 6-positions of donor sugars and azido modifications at the 6-position. However, the creation of a glycosynthase capable of using either 3- or 4-azido or 3- or 4-methoxy sugars as donors remains elusive. Previous work by Shim et al. [272] produced a glycosynthase capable of transferring a 3-O-methyl-glucosyl moiety by means of directed evolution of an existing glycosynthase (Abg2F6) [153]. Their strategy involved the saturation mutagenesis at key primary protein interaction sites around the 3-hydroxyl group within the hydrolase and plate-based activity screens. This strategy may also be successful for the incorporation of azido-modified sugars. Another approach could be to target specific β-glucosidase families known to have relaxed interactions at either the 3- or 4-hydroxyl positions. The GH5 family, which typically has endo-activity, but contains members with β-glucosidase activity could be a useful starting point.

The question of what makes a good candidate for a glycosynthase has been addressed previously by Ducros et al [81]. Within this paper the authors suggested that measuring the reactivation rates of a 2-fluoro-glycosyl-enzyme intermediate (a proxy for the glycosynthase bound α-glycosyl fluoride) with acceptors and comparison of this rate to the reactivation rate with water could be useful metrics for determining whether a hydrolase will yield an efficient glycosynthase. They found that hydrolases with reactivation rates (k_{trans}) rates $> 10^{-2}$ min^{-1} and high selectivity for transfer to acceptor over water (k_{trans}/k_{H_2O}) > 20 acted as efficient glycosynthases. Although I did not

Table 4.11: Comparison of Enzyme Reactivation.

Enzyme	Family	Acceptor	Enzyme Reactivated (%) H_2O	Acceptor	Ratio Acceptor/H_2O	Active Glycosynthase?
Ali	GH1	pNP-Glc	0.06	0.31	5.6	Y
Dei	GH1	pNP-Glc	0.69	19.00	27.4	Y
Exi	GH1	pNP-Glc	9.15	74.68	8.2	Y
Lac	GH1	pNP-α-Glc	4.30	55.12	12.8	N
Myx	GH1	pNP-Glc	0.56	7.85	14.1	Y
Pha	GH1	pNP-C	0.13	1.04	8.2	Y
Sac	GH1	pNP-Glc	4.47	75.72	16.9	Y
The	GH1	pNP-Glc	0.17	0.71	4.1	Y
C08-GH3	GH3	pNP-Glc	1.65	12.69	7.7	N
O22-GH3	GH3	phenyl-Glc	34.44	35.87	1.0	N
C24-GH3-1	GH3	Galactal	11.89	43.34	3.6	N
D08-GH3	GH3	pNP-Glc	11.25	12.76	1.1	N
C11-GH1	GH1	pNP-Glc	0.45	16.75	37.3	Y
O03-GH42	GH42	3-Mercapto-1-propanol	52.66	24.70	0.5	N

directly measure such rates, the results of the acceptor specificity tests may give some insight into the selectivity for transfer (see Table 4.11). The enzymes that were successfully transformed into glycosynthases all had higher percentages of acceptor reactivation (total reactivation with acceptor - reactivation due to water) than water reactivation. Most had ratios of reactivation (Acceptor reactivation/ water reactivation) greater than 5. The hydrolases which could not be transformed into successful glycosynthases had fairly low ratios of the % enzyme reactivated (Table 4.11), with O03-GH42 even having a higher rate of hydrolysis than transfer to any acceptor. C08-GH3 was exceptional in that it had a ratio (7.7) comparable to enzymes that could be transformed into active glycosynthases, however, it may be that other factors such as a low transfer rate are limiting this enzyme from becoming an effective glycosynthase.

4.6 Conclusions

The variety of different bonds formed by this panel of glycosynthases truly speaks to the power of harnessing the diversity of enzymes present in nature. By exploring a wide variety of enzymes through hydrolase screening, we were able to rapidly identify enzymes with promiscuous hydrolase activity. Coupling this process to acceptor specificity screening enabled the identification of ideal substrates to use with each glycosynthase. Eight wild-type enzymes were transformed into competent glycosynthases, which were able to catalyse a variety of glycosylations. This set of

glycosynthases was able to generate disaccharides, trisaccharides, glycolipids and inhibitors containing azido or amino functional handles. The ability to synthesize glycans containing modified glucans will, going forward, enable the rapid diversification of molecules, to include a variety of functionalities such as fluorophores or specificity elements.

Chapter 5

Conclusions

In this thesis I harnessed high-throughput functional metagenomic screening to identify novel genes involved in carbohydrate degradation throughout oceanic, soil, coal-bed and man-made bioreactor environments. In addition I harnessed this technology to detail microbial mechanisms of carbohydrate degradation within the beaver digestive tract, and reveal new synergistic modes of degradation. Libraries of identified metagenomic clones and synthesized genes were then profiled to reveal promiscuous enzymes which, in turn, were developed into new synthetic tools.

The aim of this chapter is to provide an analysis and to integrate of the research within this thesis in light of current research in the field. In addition, the limitations and strengths of my approaches are analysed, as are possible future directions for investigation.

5.1 Relevant Research

Discovery of New Glycoside Hydrolases

Functional metagenomic screening offers a valuable method to discover biomass-degrading enzymes, to complement more traditional methods for enzyme discovery, including activity-based screening of isolates or isolate libraries, and genetic analysis of known carbohydrate-degrading organisms. Of the ten most recently discovered GH families (see Table 5.1), seven have been identified through genetic analysis of PULs, reflecting both the current interest in these gene clusters and the catalytic power of the human symbiont *B. thetaiotaomicron*. Two of the other three most recently identified families (GH144 & GH149) were identified through isolate activity screening, followed by protein purification [2, 158]. The last GH family (GH148) was identified through functional metagenomic screening of a fosmid-harbouring *E. coli* library. The DNA sample used to construct this library originated from a volcanic crater which had both high temperature (67 °C) and pH (9.3). This

library was screened with both MU-C and Carboxy-methyl Cellulose (CMC) and the clone of interest had low activity against CMC, but higher activity on β-glucans [10].

Table 5.1: The Ten Most Recently Defined Glycoside Hydrolase Families

Family	Activity	Discovery Method	Substrate	Reference
GH139	α-2-O-methyl-L-fucosidase	PUL genetic analysis	RG-II	Ndeh et al.[215]
GH140	endo-apiosidase	PUL genetic analysis	RG-II	Ndeh et al.[215]
GH141	β-L-arabinofuranosidase	PUL genetic analysis	RG-II	Ndeh et al.[215]
GH142	α-L-fucosidase; xylanase	PUL genetic analysis	RG-II	Ndeh et al.[215]
GH143	DHAase	PUL genetic analysis	RG-II	Ndeh et al.[215]
GH144	endo-β-1,2-glucanase	Isolate activity screening	β-1,2-glucan	Abe et al. [2]
GH145	α-L-rhamnosidase	PUL genetic analysis	AGP	Munoz-Munoz et al. [209]
GH146	β-L-arabinofuranosidase	PUL genetic analysis	RG-I	Luis et al.[183]
GH147	β-galactosidase	PUL genetic analysis	RG-I	Luis et al.[183]
GH148	β-1,3/β-1,4-glucanase	Metagenomic screening	CMC, β-glucan	Angelov et al. [10]
GH149	β-1,3-glucan phosphorylase	Isolate activity screening	laminaribiose	Kuhaudomlarp et al.[158]

CMC: Carboxymethyl-Cellulose, AGP: Arabinogalactan Protein, DHAase: 2-keto-3-deoxy-D-lyxo-heptulosaric acid hydrolase

Further investigation of the most recently described GH families can give us insight into successful strategies for discovery. Most of the new familes were identified with either difficult to purify and complex substrates (GH139-143 and GH135-147 which are active on RG-I, RG-II or AGP) or in the case of GH144 which is active on the uncommon β-1,2-glucan, substrates which have recently been made accessible through new synthetic schemes. The discovery of GH149 hinged on the mechanistic details of this family as activity assays were based on looking for reverse phosphorolysis products rather than degradation products. The discovery of GH148 is an outlier from this set, as it relied neither on new substrates – CMC was first employed in 1986 [262]– nor on mechanistic details, but rather was made possible by screening an extreme environment and investigating hits with low activity. Incorporating these successful strategies (complex natural substrates, probes based on mechanism, extreme environments and investigation of low activities) into functional metagenomic screens should allow for the continued discovery of new hydrolase families.

Glycoside hydrolase enzymes with activities previously unobserved within specific families have also been recently identified. The methods used to identify these enzymes has mirrored those used to identify new families. PUL genetic analysis [215], isolate screening [266, 307] transcriptomics [143] have all been used to identify new activities within known families. Additionally, phylogenetic analysis has been used to identify enzymes or enzyme subfamilies which have low sequence similarity, or are deeply branching. One such example employs a novel bioinformatic pipeline (SACCHARIS)

to identify uncharacterized subfamilies [144]. Application of this pipeline to the GH43 subfamily enabled the identification of a GH43 (*Bacteroides dorei* DSM 17855 [BdGH43b]) which is able to degrade α-D-glucans, a surprising activity as this family has hitherto only been known to degrade either β-D or α-L substrates [144].

Modified Glycan Synthesis

The glycosynthases generated in Chapter 4 enable facile incorporation of modified sugars bearing both azido and amino functional handles. Thus far only one other glycosynthase has been developed to incorporate azido functional handles. The *Hi*Cel7B E197A glycosynthase was used to synthesize a modified cellulose with 6-azido groups present at every second position [62]. This was accomplished by using a donor cellobioside possessing a single 6-azido modification on the reducing end glucose. This resulted in polymers with a degree of polymerization up to 34, which could be subsequently modified with click chemistry. Additionally, transglycosylation has been used to incorporate modified N-glycans bearing 6-azido functionalities. Ochiai and coworkers were able to use an oxazoline pentasaccharide donor sugar containing 6-azido mannose to remodel the N-glycan of a small natural glycoprotein [225].

Another successful avenue for the generation of glycosynthases that act on modified sugars is the use of directed evolution. As mentioned within Chapter 4, other members from the Withers group have had success subjecting Abg glycosynthase to directed evolution, and specifically screening for the incorporation of non-natural substrates with modified substituents at the C3- position [272]. To achieve this enhanced activity, a variant library of wild-type enzymes was first screened for hydrolytic activity with a 3-*O*-methyl-β-D-galactopyranoside. The mutations identified from this directed evolution were then introduced into the Abg 2F6 glycosynthase scaffold. This resulted in a 39-fold increase in glycosynthase activity when 3-*O*-methyl glucopyranosyl fluoride was used as the donor sugar. One could envisage a hybrid strategy employing both metagenomic dicovery, to first identify candidate GHs, and directed evolution to improve or modify their activities before conversion to a glycosynthase.

Glycosyl transferases have also been used to incorporate modified glycosides. Using a method they have termed glycorandomization, Jon Thorson and collegues have been able to repurpose GTs

to transfer amino and azido glycosides [182]. This technique employs a promiscuous nucleotidyl transferase to first synthesize modified nucleotide diphospho-sugars, which are then used as substrates for GTs. A number of different glucose analogues have been glycosylated using this method, including 3-, 4- and 6- amino glucose [97, 182]. Additionally 3-,4- and 6-azidoglucosyl moieties could be attached to erythromycin analogues [337] or vancomycin analogues [97].

5.2 Limitations and Future Directions

5.2.1 Diverse Searching

The functional metagenomic screening methods used in this thesis have enabled the identification of hundreds of new GH genes from diverse environments. This process is, however, susceptible to false negatives. Not every gene can be expressed in *E. coli* and not every glycoside hydrolase will be detected with our substrates. This invites the obvious question – how can we improve our functional screening to give ourselves a better chance of finding diverse plant biomass-degrading genes from an environment?

One problem affecting our ability to uncover biomass-degrading genes, discussed in Chapter 3, is under-sampling. This leads to poor representation of the rare taxa within the resulting library, decreasing the diversity of recovered hits. One solution to this problem is to simply create larger and larger libraries. This however leads to wasted resources as the most abundant hits are found over and over again. Also, there are technical limitations on the size of library that can be screened via plate-based methods within reasonable time frames and costs. Microfluidic technologies offer an alternative to plate-based screens and are able to more rapidly screen clone libraries [211]. Another potential solution to this problem could be to employ fluorescence activated cell sorting (FACS), which can be used to rapidly isolate sub-populations based on size, morphology or binding to specific probes. Alternatively, stable isotope probing, could be used to isolate the DNA from sub-populations. This technique relies on isotopic labelling of substrates, that when metabolized by bacteria are incorporated into their genomic DNA. This isotopically labelled DNA can then be separated via ultracentrifugation, and used to create metagenomic libraries.

The improvement of heterologous expression is another avenue which promises to improve func-

tional metagenomic screening. The screens performed in this thesis were conducted entirely in the *E. coli* EPI300 strain. This strain has several benefits: it grows rapidly, is genetically tractable and allows for copy number induction of fosmids. However, the sequence space that can be explored by this host is certainly limited. *E. coli* are limited to mesophilic growth, hindering access to thermophilic or psychotrophic enzymes, have comparatively few σ-factors and have biased codon usage [179, 302]. This has led to an exploration of alternative expression hosts and has spurred research into the creation of multiple host vectors. The Proteobacteria have been a particular area of interest with *E. coli*, *Agrobacterium tumefaciens*,[69] *Caulobacter vibrioides*,[69] *Rhizobium leguminosarum*, [178] *Ralstonia metallidurans*,[69] *Pseudomonas fluorescens*, [1] *Pseudomonas putida*,[53, 69] *Xanthomonas campestris*,[1] *Burkholderia graminis*, and [69] *Sinorhizobium meliloti*[263] all being used as screening hosts. Other bacterial hosts include *Thermus thermophilus* [170], belonging to the Deinococcus-Thermus phylum, the Firmicute *Bacillus subtilis* [26] and the Actinobacterium *Streptomyces albus*[136]. Although the number of hosts appears extensive, they lack phylogenetic diversity. Only 4 of the 30 accepted bacterial phyla have had a representative used in a functional metagenomic screen. Future advancement of functional metagenomic screening should lie in the development of new expression systems in hitherto under-utilized phyla.

Conspicuous by their absence from the list of functional metagenomic hosts are the Bacteroidetes. This phyla, as noted previously, have colonized virtually all types of environments and are well known for their ability to degrade complex carbohydrates [204]. Although I have been able to identify many fosmids with Bacteroidetes origins, expression of metagenomic DNA within a member of this phylum should increase our ability to detect carbohydrate degrading genes and PULs. One potential complication with using a Bacteroidetes as a host may be the presence of endogenous genes. Careful selection of a host with low background hydrolysis rates or creation of knock-out strains tuned to the specific screen should enable the use of a strain from this phylum.

In addition to developing hosts throughout the tree of life, it will be important to develop hosts that allow access to specific chemistries. Two recent publications have identified the role of hydrogen peroxide (H_2O_2) in the oxidative cleavage of carbohydrates [25, 161]. Detection of this activity in culture would likely require the presence of H_2O_2, however many bacteria (including *E. coli*) possess the enzyme catalase which functions to degrade hydrogen peroxide. The implementation of

functional screening in hosts known to produce hydrogen peroxide, such as *Lactobacillus acidophilus* [119], could enable the detection of such enzymes. Furthermore these strains may also be useful for the discovery of enzymes, such as lignin-peroxidases [322], which require reactive oxygen species.

It may also be beneficial to screen in hosts that are able to use uncommon amino-acids. Both selenocysteine and pyrolysine, thought of as the 21st and 22nd proteinogenic amino-acids, are essential to catalysis in certain enzymes [30, 258]. However, the synthesis and use of these amino acids does not occur in every branch of the tree of life. *E. coli* possess the machinery to incorporate selenocysteine, and they express several selenoproteins, yet they are unable to incorporate pyrolysine. *Desulfitobacterium hafniense*, an anaerobic Firmicute, is perhaps the only known bacterium that has both been isolated and is known to use both selenocysteine and pyrolysine [152, 283]. Use of *D. hafniense* as a metagenomic host would enable the detection of proteins incorporating these amino acids, which would otherwise go unseen. Although the development of metagenomic systems in new hosts may be technically difficult, it offers the potential to provide access to unexplored sequence space and new biocatalysts.

Another route forward, which circumvents the need for heterologous expression, is direct functional screening of environmental cells. This tactic also frees the researcher from the need to first purify and then insert metagenomic DNA into a vector and host strain. Two studies employing rapid screening technologies have made progress towards such direct functional screening of environmental cells [211, 250]. The first of these studies screened cells from a wheat stubble field in the North of France [211]. They used a microfluidic system to encapsulate cells in 20 pL droplets with a fluorogenic reporter (6,8-difluoro-7-hydroxycoumarin-4-methanesulfonate cellobioside) for cellobiosidase activity. Fluorescent droplets were then sorted on chip at a rate of over 100,000 bacteria in less than 20 min. DNA from the resulting cells was then used for 16S sequencing to reveal phylogeny of the hits and the sorted population was grown on agar. The second study screened surface water from Damariscotta Lake in the North-Eastern United States using a FACS-based method [250]. In this study fluoresceinamine-labeled laminarin was incubated with environmental cells, then FACS was used to sort those cells which bound to this substrate. The resulting hits were then subjected to single-cell whole-genome amplification, a powerful technique for revealing the genetic potential of environmental cells without prior culturing. This revealed 121 laminarin-binding

single amplified genomes (SAGs), five of which were sequenced. The SAG with the highest coverage (SAG AAA168-F10) contained 58 putative glycoside hydrolases and a host of other carbohydrate modifying enzymes.

The future of direct functional metagenomic screening should incorporate methodologies from both these studies. The use of droplet-based microfluidic screening with a reporter substrate offers the ability to directly detect a functional activity, which is a superior method to the FACS-based screen. This is because it offers a direct connection between a cell and its activty, unlike the FACS based screen which relied on the assumption that cells bound to substrates could also also contain the enzymes to hydrolyse them. Although the authors of the microfluidics-based study were able to sequence the 16S of the resulting hits, they failed to identify which genes were responsible for activity. Coupling microfluidics based screening and sorting to single-cell whole-genome amplification and sequencing will allow both rapid screening for activities and the identification of responsible genes.

5.2.2 Enzyme Profiling

Many of the hits identified throughout this thesis were subjected to plate-based assays to reveal the pH-dependence, thermal stability and substrate range of activity. This information is limited to relative values of initial rates as it is difficult to determine the exact concentration of an enzyme in crude lysate. To determine kinetic constants (k_{cat} & K_M) the concentration of the active enzyme must be known. Active site titrating reagents, such as chromogenic or fluorogenic 2-fluoro sugars [82, 101], offer one possible avenue towards determining enzyme concentration in crude lysate, however these substrates can only be effectively employed with retaining enzymes. The future of high-throughput enzyme characterization may therefore hinge on the rapid sub-cloning, expression and purification of proteins. As discussed in Chapter 4, robotic automation gene synthesis, protein expression and purification has been accomplished with a standalone system [32]. However for this technology to be routinely employed, throughput must be expanded and costs decreased.

The use of activated fluorogenic probes, such as the the chlorocoumarin glycosides employed in this thesis, enable rapid and sensitive screening. These substrates contain a fluorogenic molecule which is thought to occupy the +1 subsite of the active enzyme. However, positive subsite in-

teractions are undoubtedly important for catalysis. Furthermore, the identified enzymes likely have specificity towards different linkages, (β-1,2, β-1,3, β-1,4 etc.) and this information is not conveyed when the substrate contains a reporter molecule. As natural carbohydrates do not contain leaving groups that can be detected with the sensitivity of fluorophores, less sensitive and more time-consuming methods (Such as TLC, reducing sugar assays and HPAEC) must be used to characterize their activity. The use of such technologies becomes unreasonable as the number of enzymes being assayed approaches the hundreds or thousands.

The development of rapid, sensitive assays that utilize more natural substrates, that maintain +1 subsite interactions, will be an area of future research. Technologies based on Förster resonance energy transfer (FRET), biosensors, mass spectrometry and capillary electrophoresis show potential for rapid high-throughput characterization of activities. For example, a study by Yang and coworkers used FRET probes, containing two fluorophores installed on either side of a ganglioside, to interrogate endo-active hydrolases [329]. Generation of a suite of FRET-based probes which incorporate plant-biopolymer oligomers would allow for the rapid profiling of endo-acting enzymes. The development of biosensors has also been a topic of recent research, [123, 339] particularly in the context of metabolic engineering [338]. Recently, a biosensor has been developed for the detection of cellulases [162]. This work uses a genetic circuit that responds to the presence of cellobiose. When present, cellobiose derepresses the transcription of a fluorescent protein, resulting in a detectable readout. One could imagine biosensors being able to detect any molecule including the metabolites generated from plant biomass degradation. Future biosensor development will involve expanding the range of substrates that can be detected and improving the dynamic range of biosensors.

Another high-throughput characterization method is the use of Nanostructure initiator mass spectrometry (NIMS) to rapidly profile enzyme activity [108]. This method employs glycans containing fluorous tags with varying mass attached to the reducing end that are used to assay enzyme activity [222]. Accoustic deposition is then used to transfer small volumes (1 nL) of the reaction mixtures onto a chip containing a fluorous initiator. Matrix-assisted laser desorption/ionization (MALDI) is then used to detect the fluorous glycans depositied into this chip. This method has been used by Heins and coworkers [117] to detail the activity of the same panel of GH1 glycoside hydrolases used in Chapter 4 of this thesis. This work used a NIMS chip and acoustic deposition to

examine 10,080 experimental conditions with 4 different substrates. This revealed substrate preferences and temperature dependences for each of the 105 active enzymes that they assayed. Capillary electrophoresis has also been recently used to characterize enzymatically-released oligosaccharides [177]. This method re-purposed a DNA sequencer which can analyse 96 samples simultaneously to rapidly quantify sugars. This allowed the detection of sugars released from the action of xylanases on wheat flour arabinoxylan down to femtomolar ranges while differentiating between the activities of GH10 and GH11 xylanases. Future development of these methods promises to allow the rapid characterization of thousands of enzyme hits derived from metagenomic screening.

5.3 Closing

Functional metagenomic screening has the power to reveal those active genes within a microbial community that are used to shape their chemical landscape. In this thesis functional metagenomic screening has enabled the cataloging of new genes identified from diverse environments that degrade plant matter. It has also given us insight into complex carbohydrate metabolism within the beaver gut and feces. The diversity of genes discovered can serve as a starting point for both the profiling of enzyme promiscuity and the development of new catalysts. In this respect I have created 8 new competent glycosynthases from both metagenomic and synthetic gene libraries. Further refinement of these discovered and engineered catalysts will expand our carbohydrate synthesis and degradation toolkit. This, in turn, promises to open doors to more efficient degradation of plant biomass and the creation of complex molecular probes and inhibitors for carbohydrate-active enzymes.

Chapter 6

Methods

6.1 General Methods

All buffers and reagents were from Sigma-Aldrich Chemical Company unless otherwise stated. Custom DNA oligos used for sub-cloning and mutagenesis were synthesized by Integrated DNA Technologies. Sequence verification by means of targeted Sanger sequencing of mutants and sub-cloned genes was performed by Genewiz.

6.2 Data Accessioning

Nucleotide sequences for fosmids described in Chapter 2 have been deposited in Genbank (Accession ID: MH105917 - MH106139). Beaver feces data has been deposited in the NCBI BioProject portal (Bioproject ID: PRJNA261082), for assembled metagenomic reads (BioSample ID: SAMN04122864), unassembled metagenomic reads (BioSample ID: SAMN03389401), functionally identified fosmids (Biosample ID: SAMN03389402) and pyrotags (Biosample ID: SAMN03389403). Functionally identified beaver gut fosmids have been deposited in Genbank (Accession ID: MH106140 - MH106387).

6.3 Chapter 2 Experimental

6.3.1 Sampling

Soil

Soil samples were collected by Dr. Marcus Taupp from the long-term soil productivity site at Skulow Lake, British Columbia. Soil from the organic layer, mineral layer of eluviation, mineral

transition layer and mineral layer of accumulation at both undisturbed (Libraries: NO, NA, NB and NR) and harvested sites (Libraries: CO, CA, CB and SCR) were used to create fosmid libraries [114].

Ocean

Ocean water from the North-Eastern sub-Arctic Pacific Ocean was collected by Dr. Jody Wright. Water from Line P stations 4 and 12 was collected in February 2010 at depths between 10 and 2000 meters, and used to create fosmid libraries [326].

Coal Bed

Sampling of Hydrocarbon resource environments was a result of the Hydrocarbon Metagenomics Project (http://hydrocarbonmetagenomics.com/). Four separate samples were collected and used to create fosmid libraries. Two samples were derived from cuttings of coal bed cores sourced from Rockyford Standard (CO182) and Basal (CO183) coal zones in Alberta. Another two samples (CG23A and PWCG7) were collected from co-produced water from coal bed methane well heads located in the San Juan Basin, New Mexico [8].

Bioreactors

Three additional samples were derived from bioreactors. The first, a methanogenic naptha-degrading community (NapDC) was initially inoculated with mature fine tailings from the Syncrude Mildred Lake Settling Basin (Alberta, Canada). This culture was enriched for naphtha-degrading consortia by growing the culture with 0.2 % (v/v) hydrocarbon mixture naphtha as a sole carbon source [288]. The second enrichment culture was a methanogenic toluene-degrading culture (TolDC) derived from a shallow gas condensate-contaminated aquifer located beneath a natural gas production site in Weld County (Colorado, USA). This culture was enriched for toluene-degrading consortia by propagating the culture with 0.01 % toluene (v/v) as a sole carbon source, prior to DNA isolation [93, 104, 288]. The final sampled bioreactor (FOS62) was an designed for remediation of metal contaminated effluent from smelting operations. The bioreactor is located in Trail, British Columbia and contains a mixture of limestone, quartz sand and Celgar biosolids, a by-product of the pulp

and paper industry. The biosolids were used in an anaerobic digester by the Zellstoff Celgar mill company and therefore include bacterial biomass and partially degraded and composted cellulose and hemicellulose [4]. A homogenized core sample was collected, as described by Mewis et al [201].

6.3.2 Library Creation

Fosmid Libraries were created by Dr. Sangwon Lee. Once environmental DNA had been isolated and purified, fosmid libraries were generated. Fosmid library creation was performed as previously described using the CopyControl Fosmid Library Production Kit with pCC1FOS Vector Kit (Epi-Centre) [292]. Briefly, the DNA was end repaired to create 5 -phosphorylated blunt ends and then subjected to pulsed-field gel electrophoresis (PFGE) to size-select 35-60 kb DNA fragments. The DNA was recovered by gel extraction and ligated into the pCC1 vector. Linear concatemers of pCC1 and insert DNA were packaged into a phage and transduced into phage-resistant *E. coli* EPI300 cells. The successfully transduced clones were recovered on LB agar plates containing chloramphenicol (12.5 μg/mL) and picked into 384-well plates, containing 100 μL of LB chloramphenicol (12.5 μg/mL) and 10 % glycerol, with an automated colony-picking robot (Qpix2, GENETIX). Clones were grown overnight at 37 °C then stored at -80 °C. In total, fosmid library construction produced 309,504 individual clones from a diverse set of environments (Table 2.1).

6.3.3 Fosmid End-Sequencing

Bi-directional Sanger end-sequencing was performed on a subset of the libraries using the ABI BigDye kit (Applied Biosystems, Carlsbad, Ca) on all clones at Canada's Michael Smith Genome Sciences Centre, Vancouver, B. C. Canada. The primers used were pCC1 sequencing primers (forward: GGATGTGCTGCAAGGCGATTAAGTTGG, reverse: CTCGTATGTTGTGTGGAATTGTGAGC).

6.3.4 Annotation of End-Sequences

Open reading frames (ORFs) from fosmid end-sequences were predicted using Prodigal [134] implemented in the MetaPathways pipeline [155]. The 352,994 end-sequences yielded 400,561 ORFs >180

nucleotides in length which were annotated using LAST [150] implemented in the MetaPathways pipeline based on queries of the CAZy database [181] (retrieved 2014,09,04).

6.3.5 Functional Screening

Screening was performed generally according to procedures by Mewis et al. [201] with modifications. 384-well master plates were thawed at 37 °C for 20 minutes, after which they were replicated into 384-well plates (Corning 3680) containing 40 μL per well of LB chloramphenicol (12.5 μg/mL) with arabinose (100 μg/mL). Replicated plates were then incubated in a humid chamber at 37 °C for 16-18 hours. Plates were removed from the incubation chamber and 40 μL of Assay mix (1 % Triton X-100, 1 mM 4-methylumbelliferyl cellobioside, 100 mM potassium acetate, pH 5.5) was then added to each well. These plates were incubated at 37 °C for a further 16-18 hours in a humid chamber. Fluorescence was subsequently measured using a Varioskan (ThermoFisher) plate reader with the excitation wavelength = 365 nm and the emission wavelength = 450 nm. Fosmids chosen for sequencing (>3 standard deviations above the mean for each substrate) were validated by rescreening each clone in triplicate. These clones were rearrayed using an automated colony-picking robot (Qpix2, Molecular Devices), into a 384 well plate (Corning 3680) containing 80 μL of LB chloramphenicol (12.5 μg/mL) and 10% glycerol. This master plate was incubated overnight at 37 °C and then stored at -80 °C.

6.3.6 Fosmid DNA Isolation and Sequencing

The 384 well master plate was used to streak cultures onto LB chloramphenical (12.5 μg/mL) agar plates. Individual colonies were inoculated into 5 mL of terrific broth (TB) media and incubated with shaking for 18 hours at 37 °C. Fosmids were purified with a QIAprep Spin Miniprep Kit (QIAGEN), treated with PlasmidSafe ATP-dependent DNAse (Epicentre) and quantified using a Qbit fluorimeter (ThermoFisher). Purified DNA was prepared for sequencing on the Illumina MiSeq platform using Nextera XT library preparation kit and 96 sample Nextera V1 index kit. Bead-based normalization was used before pooling samples, and samples were sequenced using paired end 150 bp reads (2 x 150 bp mode). Fastq sequences were obtained from the sequencer and quality was assessed using FastQC. Raw sequences were trimmed to Q30 quality, and residual contaminating

E. coli genomic DNA was removed by alignment to the *E. coli* K12 reference genome using the bwa aligner [174]. Trimmed reads were assembled at a range of kmer values (64 to 160) using ABySS [275] and the kmer value that produced the fewest contigs of appropriate size (25 - 40 kb) was selected. The presence of pCC1 vector sequence at ends of fosmids signalled the proper contig to select. Wells that did not produce contigs with pCC1 vector present were end-sequenced and compared to all contigs produced from that well to identify the correct sequence. Assembly and quality control were done in part by Dr. Keith Mewis and Connor Morgan-Lang.

6.3.7 Fosmid Annotation

Open reading frames (ORFs) were predicted using Prodigal [134] implemented in the MetaPathways pipeline [155]. The 188 assembled fosmids yielded 4,969 ORFs >180 nucleotides in length which were annotated using LAST [150] implemented in the MetaPathways pipeline based on queries of the CAZy [181] (retrieved 2014-09-04), COG [291] (retrieved 2016-10-20), KEGG [146] (retrieved 2011-06-18) and refseq-nr [246] (retrieved 2014-01-18) databases.

6.3.8 GH Family Trees

All characterized protein sequences from GH1, GH3, GH5, GH8 and GH9 were downloaded from the CAZy database [295] in May of 2017 (Table 6.1). Sequences were clustered at 95% similarity with UCLUST [84]. Fosmid encoded proteins from the same families as above were compiled and clustered as for the characterized CAZy proteins. Representative sequences from both sets for proteins were aligned with COBALT [230] and poorly aligned regions were removed with trimAL [45] using a gap threshold of 0.95 and a conservation threshold of 0.8. The trimmed multiple sequence alignments were then used to generate phylogenetic trees based on maximum likelihood analysis using RAxML [284]. One hundred bootstrap cycles were performed using the Whelan and Goldman substitution model [316] and Γ distribution of heterogeneity as parameters. Trees were rerooted at the tree midpoint and visualised with the phytools [251] package in R.

Table 6.1: Sequences Used To Generate Trees

	CAZy Characterized		Fosmid Encoded	
CAZy Family	Sequences	Clusters (95%)	Sequences	Clusters (95%)
GH1	277	247	28	18
GH3	286	256	129	74
GH5	536	451	50	29
GH8	73	50	7	2
GH9	167	143	6	4

6.3.9 Fosmid-Encoded Activity Characterization

Colonies that were selected for characterization were inoculated into a 96 deep-well plate (Costar 3960) containing 800 μL of LB with chloramphenicol (12.5 μg/mL) and arabinose (100 μg/mL). After 18 hours of growth at 37° C with shaking, cells were harvested by centrifugation at 3,200 x *g* for 20 min. The supernatant was decanted, cell pellets were re-suspended in 100 μL of buffer (20 mM NaOAc, 10 mM NaCl, pH 6.0) and OD600 was recorded. The cell suspension was added to 100 μL of 2x lysis buffer (20 mM NaOAc, 10 mM NaCl, 2 % Triton X-100, 0.5 mg/mL lysozyme, cOmplete Protease Inhibitor-EDTA free(Sigma-Aldrich), pH 6.0) and incubated for 2 hours at 20 °C.

General Assay Procedure

To initiate reactions 20 μL of lysate was added into a 96-well plate containing 100 μL of buffer with substrate (20 mM sodium acetate, 10 mM NaCl, 240 μM substrate, pH 6.0). Reactions were performed using a Beckman Coulter Biomek FX workstation and run in triplicate at 20 °C. Samples (10 μL) were taken after set intervals and quenched by addition to 100 μL of stop buffer (1 M glycine, pH=10.4). The fluorescence of quenched reactions was determined with a Beckman Coulter DTX-880 Multimode Detector (λ_{ex}ex = 365 nm, λ_{em} = 465 nm). Initial rates, below 10 % of substrate consumption, were used to quantify enzyme activity.

Substrate Preference

To asses substrate preference assays were setup according to the general protocol, with eight 4-methylumbelliferyl (MU) glycoside substrates (MU cellobioside, MU lactoside, MU β-D-glucopyranoside,

MU β-D-galactopyranoside, MU β-D-xyloside, MU α-L-arabinoside, MU β-D-mannopyranoside and MU β-D-*N*-acetylglucosaminide).

pH Dependence

pH dependence assays with the optimal substrate, as determined by the substrate specificity assay, were conducted using the general protocol with buffer replaced by a one of a set of eight citrate-phosphate buffers (50 mM sodium phosphate, 25 mM sodium citrate, 10 mM NaCl) at a pH between 4-7.7, and repeated in pH 7-9.8 Glycyl-glycine buffers (20 mM) when necessary. Assays were conducted at 20 °C with sampling after 1, 4 and 16 hours or when appropriate they were incubated at at 37 °C for 2.5 hours. The optimum pH was recorded as the pH where the maximum velocity was observed.

Thermal Stability

The lysates were aliquoted and pre-incubated for 10 minutes at different temperatures using a gradient incubation protocol on a 96-well MyCycler thermal cycler (Biorad), at 37, 39, 42, 46, 52, 57, 60, 62 °C, and repeated at 37, 42, 52, 60, 65, 70, 80.4, 90 °C if found necessary. Assays with the best substrate, as determined by the substrate specificity assay, were then setup according to the general protocol, and conducted at 20 °C with sampling after 1, 4 and 16 hours or when appropriate they were incubated at at 37 °C for 2.5 hours. Data were fit to the van't Hoff equation (6.1) to deduce the denaturation midpoint temperature (T_m).

$$\ln K_D = \frac{\Delta S_D^\circ}{R} - \frac{\Delta H_D^\circ}{RT} \tag{6.1}$$

Where K_D is the is the equilibrium constant of denaturation, ΔS_D° is the standard-state entropy of denaturation, $\Delta H_D^\circ$is the standard-state enthalpy of denaturation, R is the ideal gas constant and T is absolute temperature.

6.4 Chapter 3 Experimental

6.4.1 Sample Collection

Fecal samples were collected by Dr. Kevin Mehr and myself on April 24^{th}, 2012 from two beaver that were being cared for at the Critter Care Wildlife Society located in Langley, British Columbia, Canada. Animals were fed branches from a variety of woody plant species, native to the Pacific Northwest. Due to the difficulty of obtaining fresh fecal matter, as beavers defecate underwater, samples were collected from material that had accumulated at the enclosure water outflow grating, within 12 hours of cleaning. The enclosure was open to the environment and not heated. The temperature fluctuated between 7 and 15 °C in the time before sample collection. As both beavers shared the same enclosure, it was not possible to identify which animal the samples came from. Samples were frozen in a slurry of dry-ice and ethanol and transported to the laboratory on dry-ice and were stored at -80 °C.

Intestinal samples were collected by Dr. Keith Mewis and myself from beavers freshly trapped (< 24 hours) by Allan Starkey in Maple Ridge, British Columbia, Canada between January 18th, 2014 and April 14th, 2014. Six beavers were dissected in Maple Ridge to remove the entire digestive tract (esophagus to rectum) and transported on ice to UBC. Beaver 1 had suffered trauma that resulted in rupture of the stomach and possible contamination from other gut sites. Digestive tracts were dissected and chyme or feces was collected from five locations (stomach, small intestine, cecum, proximal colon, and rectum). Collected samples were frozen in a slurry of dry-ice and ethanol and stored at -80 °C.

6.4.2 DNA Extraction

High molecular weight DNA was extracted as described previously [169]. Four grams of beaver feces or chyme was thawed and extracted in two gram duplicates. The samples were ground by mortar and pestle under liquid nitrogen, and extracted three times with extraction buffer (100 mM sodium phosphate pH 7.0, 100 mM Tris-HCl, 100 mM EDTA, 0.5 M NaCl, 1 % hexadecyltrimethylammonium bromide, 2 % sodium dodecyl sulfate) at 65 °C with rotation. The resulting supernatant was washed with chloroform-isoamyl alcohol. Finally, DNA was purified by isopropanol precipitation

and quantified using the PicoGreen assay (Invitrogen).

6.4.3 PCR Amplification of Ribosomal SSU Gene Sequences

Following DNA isolation, the V6-V8 region of the small subunit ribosomal RNA (rRNA) gene was PCR amplified with the universal three-domain primers 926F (5-AAA CTY AAA KGA ATT GAC GG-3) and 1392R (5-ACG GGC GGT GTG TRC-3). Reverse primer sequences were modified to include the 454 adaptor sequence and a 5 base-pair (bp) barcode for multiplexing during sequencing. Reactions were run in duplicate under the following PCR conditions: initial denaturation cycle at 95°C for 3 minutes; 25 cycles of 95°C for 30 seconds, primer annealing at 55°C for 45 seconds, and extension at 72 °C for 90 seconds; and a final extension cycle at 72°C for 10 minutes. Each 50 μL reaction contained: 1-10 ng template DNA, 0.6 μL Taq polymerase (Bioshop Canada Inc., 5 U/μL), 5 μL 10X reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (Invitrogen), 200 nM each of forward and bar-coded reverse primers, and 33.4 μL nuclease free water (Fisher). Duplicate reactions were pooled and purified using a QIAquick PCR Purification Kit (Qiagen) and quantified using the PicoGreen assay (Invitrogen). Samples were diluted to 10 ng/μL and pooled in equal concentrations.

6.4.4 Sequencing and Assembly

Amplicon pools from both the fecal and intestinal samples were sent to The McGill University and Génome Québec Innovation Centre for 454 pyrosequencing using the Roche 454 GS FLX Titanium (454 Life Sciences, Branford, CT, USA) technology according to the manufacturer's instructions.

Metagenomic DNA from the beaver feces was sent to the same facility and was sequenced with the same platform. This resulted in 616,811 reads of average length 761 bp, and total length of 469.2 Mbp. Sequences were trimmed by Dr. Keith Mewis to Q30 quality score using prinseq lite+ [265] and assembled using MIRA [54], resulting in 75,523 contigs with an N50 of 1787 bp and 130.5 Mbp of consensus sequence.

Metagenomic DNA from intestinal samples was sequenced at the UBC Pharmaceutical Sciences Sequencing Center on an Illumina MiSeq using paired-end, 300 bp Nextera XT chemistry (Illumina, San Diego, CA). Fifteen samples (five sites from three different beavers) were indexed and pooled

for sequencing. DNA sequences from intestinal samples was trimmed by Dr. Keith Mewis to Q30 quality score prinseq lite+ [265]. Dr. Keith Mewis and Connor Morgan-Lang attempted to assemble this data with ABySS [275], IDBA-UD [235] and SPAdes [18], however all assemblies generated contained less than 5 % of the unassembled data. Paired-end reads were merged by Dr. Keith Mewis using FLASH [189], with parameters specifying a minimum 20 base pair overlap with 95 % similarity, to generate reads up to 580 bp of high quality. Assembly attempts with these combined reads remained poor, with no samples showing N50 values (a weighted median statistic such that 50 % of the entire assembly is contained in contigs equal to or larger than this value) above 1000 bp.

6.4.5 Analysis of Pyrotag Data

The software package Quantitative Insights Into Microbial Ecology (QIIME) [46] was used to analyze both the fecal and gut pyrotag sequences. As a quality control step, sequences with quality scores less than Q25, those containing ambiguous bases, or identified homopolymer runs, or chimeric sequences, or with length less than 200 bp were removed. The remaining high quality sequences (see Table 6.2 for breakdown by sample) were clustered at the 97% identity threshold with a maximum e-value cut-off of 1 x 10^{-10} using UCLUST, implemented in QIIME software [46]. Singletons were omitted from downstream analyses, leaving a total of 1,044 operational taxonomic units (OTUs) made up of 154,028 pyrotag sequences from both fecal and gut samples. Taxonomic assignment for each OTU cluster was performed using the Basic Local Alignment Tool (BLAST) [7] and the SILVA database version 111 (www.arb-silva.de) [248] with a confidence level of 0.8 and a maximum e-value cut-off of 1 x 10^{-3}. OTU abundance was normalized to the total number of reads recovered, and expressed as a normalized percentage for analysis.

6.4.6 Analysis of Metagenomic Sequences

Genes from the beaver fecal metagenome were predicted from both assembled and unassembled data using Prodigal [134] within the MetaPathways software package. The assembled metagenome yielded 151,180 open reading frames (ORFs) larger than 60 amino acids. Using the LAST algorithm [150] within MetaPathways [155], these ORFs were compared to the KEGG [146], COG [291],

RefSeq [246], and MetaCyc [49] databases. The unassembled intestinal metagenomes were queried in an identical fashion, revealing a total of 4,910,871 predicted ORFs larger than 60 amino acids, Table 6.3

6.4.7 Fosmid Library Creation

For large insert library construction, DNA was further purified by cesium chloride gradient ultracentrifugation [325]. The large insert libraries were constructed as described previously [292] using the CopyControl Fosmid Library Production Kit with pCC1FOS Vector Kit (EpiCentre). A library of 12 x 384-well plates of clones (4,608 individual clones) was generated for the Fecal library. Additionally, DNA from the intestinal tract of beaver 2 was used to make libraries derived from the cecum (17 x 384-well plates, 6,528 clones), the proximal colon (39 x 384-well plates, 14,976 clones) and the rectum (58 x 384 well plates, 22,272 clones). We also attempted to make libraries from the stomach and small intestine DNA from beaver 2, however this DNA was highly fragmented and I was unsuccessful.

Clones were picked with an automated colony picking robot (Qpix2, Molecular Devices) and inoculated into plates containing 100 μL of LB chloramphenicol (12.5 μg/mL) and 10% glycerol. These plates were incubated overnight at 37 °C then stored at -80 °C.

6.4.8 Functional Screening

Screening was performed generally according to procedures by Mewis et al. [201] with modifications. Screening was carried out in phosphate buffer (final concentration 25 mM sodium phosphate pH 6.0), with 100 μM each of three fluorogenic substrates (6-chloro-4-methylumbelliferyl cellobioside, 6-chloro-4-methylumbelliferyl xylobioside, and 6-chloro-4-methylumbelliferyl β-D-xylopyranoside) were pooled to screen for multiple activities simultaneously. Screening was performed at a temperature of 37 °C, which is the body temperature of *Castor canadensis* [83]. Wells with fluorescence above a specific threshold (z-score $>$ 3 for the fecal library, and robust z-score $>$ 40 for the gut libraries) were selected for validation and re-screening of these clones was performed in triplicate. Fosmid containing clones chosen for sequencing (validated with a z-score $>$3 for each substrate) were rearrayed using an automated colony-picking robot (Qpix2, Molecular Devices), into a 96 well

plate (Costar 3370) containing 200 μL of LB chloramphenicol (12.5 μg mL^{-1}) and 10 % glycerol. This master plate was incubated overnight at 37 °C and then stored at -80 °C.

Screening was performed similarly for the beaver intestinal fosmids, except for the use 6-chloro-4-methylumbelliferyl β-D-mannoside in the place of 6-chloro-4-methylumbelliferyl β-D-xylopyranoside.

6.4.9 Fosmid Preparation and Sequencing

The 96 well master plate was used to inoculate a 96 deep-well plate (Costar) containing 1.65 mL LB with chloramphenicol (12.5 μg/mL) and arabinose (100 μg/mL). This deep-well plate was incubated with shaking (37 °C, 320 rpm) for 20 hours, after which the plate was centrifuged at 1500 × g for 10 minutes and the supernatant was decanted. Fosmids were purified from the pelleted cells using a Montage Plasmid MiniprepHTS 96 Kit (Millipore), treated with PlasmidSafe ATP-dependent DNAse (Epicentre) and quantified using the PicoGreen assay (Invitrogen).

Purified DNA was prepared for sequencing on the Illumina MiSeq platform using Nextera XT library preparation kit and 96 sample Nextera V1 index kit. Bead-based normalization was used before pooling samples, and samples were sequenced using paired end 150 bp reads (2 x 150 bp mode). FastQ sequences were obtained from the sequencer and quality was assessed using FastQC. Raw sequences were trimmed to Q30 quality, and residual contaminating *E. coli* genomic DNA was removed by alignment to the *E. coli* K12 reference genome using the bwa aligner [174]. Trimmed reads were assembled at a range of kmer values (64 to 160) using ABySS [275] and the kmer value that produced the fewest contigs of appropriate size (25 - 40 kb) was selected. The presence of pCC1 vector sequence at ends of fosmids signalled the proper contig to select. Wells that did not produce contigs with pCC1 vector present were end-sequenced and compared to all contigs produced from that well to identify the correct sequence.

6.4.10 Fosmid Annotation

Open reading frames (ORFs) were predicted using Prodigal [134] implemented in the MetaPathways pipeline [155]. The assembled metagenome yielded 151,180 ORFs >180 nucleotides in length which were annotated using LAST [150] implemented in the MetaPathways pipeline based on queries of the CAZy [181] (retrieved 2014,0904), COG [291] (retrieved 2016-10-20), KEGG [146] (retrieved

2011-06-18) and refseq-nr [246] (retrieved 2014-01-18) databases.

6.4.11 Fosmid Encoded Enzyme Specificities

Fosmid hits from the fecal library were further characterized once chosen for sequencing. The frozen master plate was used to inoculate a deep well plate (Costar) containing 0.8 mL of LB containing chloramphenicol (12.5 μg/mL) and arabinose (100 μg/mL). This expression plate was incubated at 37 °C for 18 hours with shaking at 225 rpm. Cells were harvested by centrifugation at 3220 x g for 20 min. After supernatant was decanted, cell pellets were re-suspended in 200 μL of buffer (50 mM sodium phosphate, 10 mM NaCl, pH 6.0) and OD600 was recorded. This cell suspension was added the same volume of lysis buffer (50 mM sodium phosphate, 10 mM NaCl, 2 % triton, 0.5 mg/mL lysozyme, cOmplete Protease Inhibitor- EDTA free (Roche), pH 6.0) and incubated for 1h at 20 °C.

Activity assays were performed in 96-well plates (Costar) which contained 40 mM sodium phosphate, 200 μM substrate and 20 μL of cell lysis. Substrates assayed for activity were: 6-chloro-4-methylumbelliferyl cellobioside, 6-chloro-4-methylumbelliferyl xylobioside, and 6-chloro-4-methylumbelliferyl β-D-xylopyranoside, methylumbelliferyl cellobioside, methylumbelliferyl β-D-glucopyranoside, methylumbelliferyl xylobioside, methylumbelliferyl β-D-xylopyranoside, methylumbelliferyl lactopyranoside, methylumbelliferyl β-D-galactopyranoside, methylumbelliferyl β-D-mannopyranoside, methylumbelliferyl α-L-arabinofuranoside and methylumbelliferyl *N*-acetyl-β-D-glucosaminide. Reactions were setup on a Beckman Coulter Biomek FX workstation and run in triplicate at 20 °C. Samples (10 μL) were taken after 1, 2, 4, 6 hours and quenched with stop buffer (1 M glycine, pH=10.4) and analyzed by fluorescence spectroscopy on a Beckman Coulter DTX-880 Multimode Detector (λ_{ex} = 365, bandwidth 25 nm, λ_{em} = 465, bandwidth 35 nm).

6.4.12 Sub-Cloning of Genes

To further investigate the GH43 genes belonging to uncharacterized subfamilies present on the beaver fecal fosmids I sub-cloned and expressed the protein products. One GH43 gene from each of the uncharacterized subfamilies 2, 7 and 28 was chosen for cloning, expression and characterization. The three genes (12_H03-12, 12_H03-13, 12_J03-18), were inserted into a pET28 vector by use of the

Polymerase Incomplete Primer Extension method [154]. Purified fosmids were used as a template for insert amplifications, while purified pET28 was used as the vector template. Each PCR reaction contained 10 μL of Phusion reaction buffer, 1.5 μL of dNTPs (10 mM), 1 μL forward primer (10 μM), 1 μL reverse primer (10 μM) 2 μL of template DNA (5 ng/μL) 0.5 μL Phusion polymerase and 34 μL of water. The insert PCR was performed with the following parameters: Initial denaturation at 95 °C for 2 minutes followed by 25 cycles of denaturation at 95 °C (30 s), annealing between 57 °C and 70 °C (30 s) and extension at 72 °C (1 min). Vector PCR was performed as above, except the annealing temperature was 55 °C and the extension time was 3.5 minutes. The primers used are detailed in Table 6.4. PCR products were mixed and transformed into DH5α cells, plasmids were sequence verified, then transformed into BL21(DE3) cells for expression.

6.4.13 Mutagenesis

The variant enzymes H03-13_E507A and H03-13_E209A were produced by means of modified QuikChange mutagenesis [180]. PCR was first performed for 12 cycles with one of the sense or anti-sense primers these two reactions were subsequently pooled and an additional 16 cycles of PCR were performed. Each PCR reaction contained 10 μL of Phusion GC reaction buffer, 2.5 μL of dNTPs (10 mM), 2.5 μL sense or anti-sense primer (10 μM), 10 μL of template DNA (5 ng/μL) 1 μL Phusion polymerase and 24 μL of water. The PCR cycling parameters were: Initial denaturation at 98 °C for 30 s followed cycles of denaturation at 98 °C (10 s), annealing between 60 °C and 70 °C (30 s) and extension at 72 °C (3 min and 15 seconds). The primers used are detailed in Table 6.5. PCR reactions were digested with the endonuclease DpnI (ThermoFisher) for 1 hour at 37 °C. This digestion reaction was subsequently cleaned up with a GeneJet PCR purification kit (ThermoFisher) and DNA was eluted into water. The cleaned up DNA (10 μL) was then used to transform DH5α cells, plasmids were sequence verified, then transformed into BL21(DE3) cells for expression.

6.4.14 Protein Expression and Purification

Proteins were purified with use of polyhistidine tags and Ni-NTA resin columns. Cultures of 50 mL LBE-5052 [285], containing 50 μg/L of kanamycin were inoculated with the expression host

and cells were grown for 18 hours at 37 °C (12_H03-12 and 12_J03-18) or 30 °C (12_H03-13, H03-13_E507A and H03-13_E209A) with shaking. Cultures were centrifuged (3,200 x g, 4 °C, 20 min), the supernatant was removed and cell pellets were stored at -80 °C until purification. To purify proteins 2.5 mL of lysis mix (1 x BugBuster [Novagen], 20 mM HEPES, 300 mM NaCl, 20 mM Imidazole, pH 7.0) was used to resuspend thawed cell pellets. This suspension was incubated at 20 °C for 20 minutes, after which the lysate was clarified by centrifugation (3220 x g, 4 °C, 20 min) and loaded onto columns containing 1 mL of HisPur resin (ThermoScientific). Columns were washed with 20 mL of Buffer A (20 mM HEPES, 300 mM NaCl, 20 mM Imidazole, pH 7.0) and protein was eluted with 4 mL of Buffer B (20 mM HEPES, 300 mM NaCl, 500 mM Imidazole, pH 7.0). Proteins were buffer-exchanged into storage buffer (20 mM HEPES, 300 mM NaCl, pH 7.0) with Amicon 30 kDa filter columns and stored at 4 °C. Protein concentrations were determined based on absorbance at 280 nm.

6.4.15 Protein Characterization

Each purified enzyme was tested for activity on the following model substrates: p-nitrophenyl β-D-xylopyranoside,, p-nitrophenyl β-D-xylopyranoside, p-nitrophenyl α-L-arabinofuranoside, 4-methylumbelliferyl β-D-xylopyranoside, 6-chloro-4-methylumbelliferyl β-D-xylopyranoside and 4-methylumbelliferyl α-L-arabinofuranoside. Purified enzyme was added (final concentrations of 200 nM) to a solution of 100 μM substrate, 50 mM HEPES, 50 mM NaCl, pH 7.0. These assays were incubated at 37 °C for 18 hours after which absorbance (λ = 400 nm) and fluorescence (λ_{ex} = 365 nm, λ_{em} = 450 nm) were detected with a BioTek synergy H1 plate reader.

Kinetic parameters were determined using 6-chloro-4-methylumbelliferyl β-D-xylopyranoside. Assays were performed in 96 well plates (Corning 3370) containing the substrate (2.5 μM - 100 μM), buffer (50 mM HEPES, 50 mM NaCl, pH 7.0) and purified enzyme. Reactions were performed at 30 °C and fluorescence (λ_{ex}= 365 nm and λ_{em}= 450 nm, gain = 65) was monitored using a Synergy H1 plate reader (BioTek). The quantity of fluorophore generated was determined by means of a calibration curve of 6-chlorocoumarin within an identical buffer system. All reactions were performed in triplicate. Rate measurements were used to calculate kinetic parameters with the software program GraFit 7.0 software.

Enzyme activity was also determined on the oligosaccharides, 3^2-α-L-arabinofuranosyl-xylobiose (A3X), 2^3-α-L-arabinofuranosyl-xylotriose (A2XX) and a mixture of both 2^3-α-L-arabinofuranosyl-xylotetraose and 3^3-α-L-arabinofuranosyl-xylotetraose (XA3XX/XA2XX). Purified enzymes (final concentration of 0.5 μM per enzyme) were added to a solution of 4 mM substrate in HEPES buffer (50 mM HEPES, 50 mM NaCl, pH 7.0). Assays were incubated at 25 °C for 18 hours, then subsequently boiled for 10 min to inactivate the enzymes. Products were analyzed with the use of a high performance anion-exchange chromatography equipped with a pulsed amperometric detector (HPAEC-PAD). This system was quipped with a CARBOPAC^{TM} PA-200 analytical anion exchange column (Dionex). The elution conditions were: 0-4 min 20 mM NaOH; 4-13 min, 20 mM NaOH with a 0 - 84 mM sodium acetate gradient; 13-14 min, 20 mM NaOH with a 84-120 mM sodium acetate gradient; 14-16 min with 20 mM NaOH and 120 mM sodium acetate. The standards used to identify the chromatographic peaks were arabinose, xylose, xylobiose, and xylotetraose.

Table 6.2: Beaver Pyrotag Counts.

		Number of Sequences		
Beaver	Site	High-quality	Singletons Removed	OTUs*
0	Feces	12,250	11,575	355
1	Stomach	3,874	3,659	431
	Small Intestine	2,167	2,115	299
	Cecum	2,784	2,593	370
	Proximal Colon	5,584	5,222	478
	Rectum	15,708	15,015	588
2	Stomach	4,904	3,594	288
	Small Intestine	5,332	5,178	357
	Cecum	251	241	105
	Proximal Colon	17,078	16,367	512
	Rectum	3,889	3,690	344
3	Stomach	3,595	3,580	53
	Small Intestine	7,899	7,718	272
	Cecum	1,352	1,264	178
	Proximal Colon	3,180	3,163	165
	Rectum	4,081	4,033	271
4	Stomach	3,574	3,564	36
	Small Intestine	4,351	4,339	27
	Cecum	4,879	4,818	281
	Proximal Colon	4,408	4,300	280
	Rectum	4,215	4,130	294
5	Stomach	5,425	4,482	120
	Small Intestine	2,626	2,620	21
	Cecum	5,272	5,050	392
	Proximal Colon	4,558	4,290	372
	Rectum	5,532	5,412	373
6	Stomach	5,055	3,927	104
	Small Intestine	8,087	8,066	31
	Cecum	2,077	2,028	276
	Proximal Colon	4,341	4,156	370
	Rectum	3,966	3,839	319
Total		**162,294**	**154,028**	**1,044**

*singletons and mitochondria/chloroplasts removed from OTUs

Table 6.3: Beaver Intestinal Metagenomes

Beaver	Site	File Size (Mbp)	Predicted ORFs
1	Stomach	101.8	1,203,77
	Small Intestine	69.5	47,018
	Cecum	385.4	689,640
	Proximal Colon	295.0	545,684
	Rectum	360.4	597,864
2	Stomach	47.1	23,566
	Small Intestine	282.2	178,664
	Cecum	390.4	592,572
	Proximal Colon	240.8	377,942
	Rectum	334.2	338,381
3	Stomach	393.1	179,079
	Small Intestine	333.0	170,004
	Cecum	243.8	325,306
	Proximal Colon	278.4	335,192
	Rectum	374.8	509,959
	Total	**4,129.9**	**4,910,871**

Table 6.4: Sub-Cloning Primers

Primer	Sequence
H03_12_Δ1-23_IPF	CTTTAAGAAGGAGATATACCATGCAGGTGGGGCAACCCTGGAT
H03_12_Δ1-23_IPR	GATCTCAATGGTGATGGTGATGGTGAGGTTCCCTCCTCATCCTCC
H03_12_Δ1-23_VPF	AGGATGAGGAGGGAACCTCACCATCACCATCACCAT
H03_12_Δ1-23_VPR	AATCCAGGGTTGCCCCACCTGCATGGTATATCTCCTTCTTAAAG
H03_13_Δ1-21_IPF	CTTTAAGAAGGAGATATACCATGCAAAACCCGCTCATCCACTC
H03_13_Δ1-21_IPR	GATCTCAATGGTGATGGTGATGGTGTTTTACATCCACAGTGATATTCC
H03_13_Δ1-21_VPF	ATCACTGTGGATGTAAAACACCATCACCATCACCAT
H03_13_Δ1-21_VPR	CGAGTGGATGAGCGGGTTTTGCATGGTATATCTCCTTCTTAAAG
12_J03_IPF	CTTTAAGAAGGAGATATACCATGAAAACCTACTGCAACCCG
12_J03_IPR	GATCTCAATGGTGATGGTGATGGTGGCCCTCCATCTTTACAATTTC
12_J03_VPF	ATTGTAAAGATGGAGGGCCACCATCACCATCACCAT
12_J03_VPR	GCGGGTTGCAGTAGGTTTTCATGGTATATCTCCTTCTTAAAG

Table 6.5: Mutagenesis Primers

Primer	Sequence
H03_13_E507A_F	GATGTGCGCACCGCCGGAATGTCATAC
H03_13_E507A_R	GTATGACATTCCGGCGGTGCGCACATC
H03_13_E209A_F	CGAAGGCTTCAAGGCAGGGCCCTTCGCCTTC
H03_13_E209A_R	GAAGGCGAAGGGCCCTGCCTTGAAGCCTTCG

6.5 Chapter 4 Experimental

6.5.1 Screening: Metagenomic Hit Library

Screening was performed using master-plates generated from the screening of numerous libraries (including all of those detailed in Chapter 2 and the fecal library generated in Chapter 3). Screening methods followed the procedures detailed in section 6.3.5, with no modifications except for the substrate used. Instead of the substrates used to originally identify the clones a panel of azido-, amino- and methoxy glycosides were used. The fluorogenic substrates used were the amino-glycosides: 4-methylumbelliferyl 3-amino-3-deoxy-β-D-glucopyranoside (MU-3-NH_2-Glc), 4-methylumbelliferyl 4-amino-4-deoxy-β-D-glucopyranoside (MU-4-NH_2-Glc), and 4-methylumbelliferyl 6-amino-6-deoxy-β-D-glucopyranoside (MU-6-NH_2-Glc); the azido-glycosides: 4-methylumbelliferyl 3-azido-3-deoxy-β-D-glucopyranoside (MU-3-N_3-Glc), 4-methylumbelliferyl 4-azido-4-deoxy-β-D-glucopyranoside (MU-4-N_3-Glc), 4-methylumbelliferyl 6-azido-6-deoxy-β-D-glucopyranoside (MU-6-N_3-Glc), and 4-methylumbelliferyl 6-azido-6-deoxy-β-D-galactopyranoside (MU-6-N_3-Gal); and the methoxy-glycosides: 3-methoxy-β-D-galactopyranoside (MU-3-O-Me-Gal) and 3-methoxy-β-D-glucopyranoside (MU-3-O-Me-Glc).

6.5.2 Sub-Cloning of Genes

The genes selected for further investigation were inserted into a pET28 vector with a C-terminal His-tag by use of the Polymerase Incomplete Primer Extension method [154]. Signal sequences were predicted using SignalP [218] and primers were designed to exclude these amino acids. Purified fosmids were used as a template for insert amplifications, while purified pET28 was used as the vector template. Each PCR reaction contained 10 μL of Phusion reaction buffer, 1.5 μL of dNTPs (10 mM), 1 μL forward primer (10 μM), 1 μL reverse primer (10 μM) 2 μL of template DNA (5 ng/μL) 0.5 μL Phusion polymerase and 34 μL of water. The insert PCR was performed with the following parameters: Initial denaturation at 95 °C for 2 minutes followed by 25 cycles of denaturation at 95 °C (30 s), annealing between 57 °C and 70 °C (30 s) and extension at 72 °C (1 min). Vector PCR was performed as above, except the annealing temperature was 55 °C and the extension time was 3.5 minutes. The primers used are detailed in Table 6.6. PCR products

were mixed and transformed into DH5α cells, plasmids were sequence verified, then transformed into BL21(DE3) cells for expression.

Table 6.6: Primers Used for Sub-Cloning Fosmid Derived Genes

Primer	Sequence
C11_GH1_pET28_fwd	GTACCATATGGTGGCTTTTTCGGATAAATTTTTGTG
C11_GH1_pET28_rev	GTACCTCGAGTTACAGATTTTTTCCGTTCCTGCTG
Beaver_09_O03_GH42_IPF	CTTTAAGAAGGAGATATACCATGTACGAAAAAGTATGGAAACAGG
Beaver_09_O03_GH42_IPR	GATCTCAATGGTGATGGTGATGGTGTATCGCCGTCTTCACGATCG
Beaver_09_O03_GH42_VPF	ATCGTGAAGACGGCGATACACCATCACCATCACCAT
Beaver_09_O03_GH42_VPR	GTTTCCATACTTTTTCGTACATGGTATATCTCCTTCTTAAAG
FOS62_40_O22-25_IPF	CTTTAAGAAGGAGATATACCATGAAACACAACATTGAAGAAATC
FOS62_40_O22-25_IPR	GATCTCAATGGTGATGGTGATGGTGATTACTGAGTCCCAAAGA
FOS62_40_O22-25_VPF	TCTTTGGGACTCAGTAATCACCATCACCATCACCAT
FOS62_40_O22-25_VPR	TTCTTCAATGTTGTGTTTCATGGTATATCTCCTTCTTAAAG
NapDC_14_D08-33_IPF	CTTTAAGAAGGAGATATACCATGTCCGATTCTGTGCTATCCA
NapDC_14_D08-33_IPR	GATCTCAATGGTGATGGTGATGGTGAGCCTGGCTGTGCACCTG
NapDC_14_D08-33_VPF	CAGGTGCACAGCCAGGCTCACCATCACCATCACCAT
NapDC_14_D08-33_VPR	GGATAGCACAGAATCGGACATGGTATATCTCCTTCTTAAAG
TolDC_15_C08-23_IPF	CTTTAAGAAGGAGATATACCATGCTTCATTACCTTTCCCGC
TolDC_15_C08-23_IPR	GATCTCAATGGTGATGGTGATGGTGCATCTCCAAGCGCAGGCT
TolDC_15_C08-23_VPF	AGCCTGCGCTTGGAGATGCACCATCACCATCACCAT
TolDC_15_C08-23_VPR	GCGGGAAAGGTAATGAAGCATGGTATATCTCCTTCTTAAAG
C24_GH3-1_Δ40_IPF	CTTTAAGAAGGAGATATACCATGAAGTTTGCACATGATTTTC
C24_GH3-1_Δ40_IPR	GATCTCAATGGTGATGGTGATGGTGGAGATCTTCCCCACGATT
C24_GH3-1_Δ40_VPF	AATCGTGGGGAAGATCTCCACCATCACCATCACCAT
C24_GH3-1_Δ40_VPR	AAAATCATGTGCAAACTTCATGGTATATCTCCTTCTTAAAG
C24_GH3-2_Δ34_IPF	CTTTAAGAAGGAGATATACCATGGAACACGATGAAAAGC
C24_GH3-2_Δ34_IPR	GATCTCAATGGTGATGGTGATGGTGTTTCCCGTTGATTAGAAT
C24_GH3-2_Δ34_VPF	ATTCTAATCAACGGGAAACACCATCACCATCACCAT
C24_GH3-2_Δ34_VPR	CTGCTTTTCATCGTGTTCCATGGTATATCTCCTTCTTAAAG
C24_GH3-3_Δ31_IPF	CTTTAAGAAGGAGATATACCATGAGCGCGGCTTCTTTTG
C24_GH3-3_Δ31_IPR	GATCTCAATGGTGATGGTGATGGTGGGTGAATTCCAGGTAATCGAG
C24_GH3-3_Δ31_VPF	GATTACCTGGAATTCACCCACCATCACCATCACCAT
C24_GH3-3_Δ31_VPR	TGCAAAAGAAGCCGCGCTCATGGTATATCTCCTTCTTAAAG
C24_GH3-4_Δ29_IPF	CTTTAAGAAGGAGATATACCATGAAACACAACATTGAAGAAATC
C24_GH3-4_Δ29_IPR	GATCTCAATGGTGATGGTGATGGTGATTACTGAGTCCCAAAGA
C24_GH3-4_Δ29_VPF	TCTTTGGGACTCAGTAATCACCATCACCATCACCAT
C24_GH3-4_Δ29_VPR	TTCTTCAATGTTGTGTTTCATGGTATATCTCCTTCTTAAAG

6.5.3 Protein Expression and Purification: Metagenome Hit Library

Proteins were purified with use of polyhistidine tags and Ni-NTA resin columns. Cultures of 50 mL LBE-5052 [285], containing 50 μg/L of kanamycin were inoculated with the expression host and cells were grown for either 18 hours at 37 °C (O03_GH42_His6, C11_GH1_His6, D08_GH3_His6) or at 30 °C for 6 hours followed by 48 hours at 18 °C (O22_GH3_His6, C08_GH3_His6, C24_GH3-1_Δ1-40_His6, C24_GH3-3_Δ1-31_His6, C24_GH3-4_Δ1-29_His6) with shaking. Cultures were centrifuged (3,200 x g, 4 °C, 20 min), the supernatant was removed and cell pellets were stored at -80 °C until

purification. To purify proteins 2.5 mL of lysis mix (1 x BugBuster [Novagen], 20 mM HEPES, 300 mM NaCl, 50 mM Imidazole, pH 7.0) was used to resuspend thawed cell pellets. This suspension was incubated at 20 °C for 20 minutes, after which the lysate was clarified by centrifugation (3220 x g, 4 °C, 20 min) and loaded onto columns containing 1 mL of HisPur resin (ThermoScientific). Columns were washed with 20 mL of Buffer A (20 mM HEPES, 300 mM NaCl, 50 mM Imidazole, pH 7.0) and protein was eluted with 4 mL of Buffer B (20 mM HEPES, 300 mM NaCl, 500 mM Imidazole, pH 7.0). Proteins were buffer-exchanged into storage buffer (20 mM HEPES, 300 mM NaCl, pH 7.0) with Amicon 30 kDa filter columns and stored at 4 °C. Protein concentrations were determined based on absorbance at 280 nm. The extinction coefficients used were: C11_GH1 ε = 128,480 $M^{-1}cm^{-1}$, C08_GH3 ε = 134,105 $M^{-1}cm^{-1}$, C24_GH3-1_Δ1-40 ε = 113,680 $M^{-1}cm^{-1}$, C24_GH3-2_Δ1-34 ε = 84,925 $M^{-1}cm^{-1}$, C24_GH3-4_Δ1-29 ε = 89,395 $M^{-1}cm^{-1}$, D08_GH3 ε = 82,655 $M^{-1}cm^{-1}$, O03_GH42 ε = 170,225 $M^{-1}cm^{-1}$, O22_GH3 ε = 79,355 $M^{-1}cm^{-1}$. All variant enzymes were expressed and purified as for the wild-type enzymes.

Both *E. coli* ATP-Dependent Glucokinase (EcGlk)[203] and *Klebsiella pneumoniae* β-glucoside kinase (BglK) [296] were expressed on a 50 mL scale, as above, and purified using the same his-tag/Ni-NTA procedure as above.

6.5.4 Wild-Type Enzyme Kinetics: Metagenomic Hit Library

Kinetic parameters for wild-type enzymes were determined using the fluorogenic screening substrates. Assays were performed in 96-well plates (Corning 3370) containing the fluorogenic glycoside (0.5 μM - 1 mM), buffer (20 mM HEPES, 300 mM NaCl, pH 7.0) and purified enzyme. Assays for I01-GH1 were performed both with and without the presence of EcGlk [203], ATP (10 mM) and $MgSO_4$ (10 mM). Reactions were performed at 37 °C and fluorescence (λ_{ex}=365 nm and λ_{em}= 450 nm) was monitored using a Synergy H1 plate reader (BioTek). The quantity of fluorophore generated was determined by means of a calibration curve of 4-methylumbelliferyl within an identical buffer system. All reactions were performed in triplicate. Rate measurements were used to calculate kinetic parameters with the software program GraFit 7.0 software.

6.5.5 Production of Mutants: Metagenomic Hits

All mutants were generated using a modified QuikChange mutagenesis protocol [180]. For each variant generated, PCR was first performed for 12 cycles with one of the sense or anti-sense primers. These two reactions were subsequently pooled and an additional 16 cycles of PCR were performed. Each PCR reaction contained 10 μL of Phusion GC reaction buffer, 2.5 μL of dNTPs (10 mM), 2.5 μL sense or anti-sense primer (10 μM), 10 μL of template DNA (5 ng/μL) 1 μL Phusion polymerase and 24 μL of water. The PCR cycling parameters were: Initial denaturation at 98 °C for 30 s followed by cycles of denaturation at 98 °C (10 s), annealing between 60 °C and 70 °C (30 s) and extension at 72 °C (3 min and 15 seconds). Wild-type plasmids were used as the template to generate all nucleophile variants. The double mutants (O22_GH3_D321S_W232F, D08_GH3_D229S_W230F, C24_GH3-1_D271S_W272F and C08_GH3_D235S_W236F) were generated using the serine nucleophile mutant as PCR template. The primers used for PCR are detailed in Table 6.7. After PCR, reactions were digested with the endonuclease DpnI (ThermoFisher) for 1 hour at 37 °C. This digestion reaction was subsequently purified with a GeneJet PCR purification kit (ThermoFisher) and DNA was eluted into water. The purified DNA (10 μL) was then used to transform DH5α cells, plasmids were sequence verified, then transformed into BL21(DE3) cells for expression.

6.5.6 Acceptor Specificity: Metagenomic Hits

Acceptor specificity screening generally followed the procedures detailed by Blanchard et. al. [29] First, between 0.5 and 3 nanomoles of purified wild-type enzyme was incubated with 1 mM of 2,4-dinitrophenyl 2-deoxy-2-fluoro-β-D-glucopyranoside (DNP 2F-Glc). In the case of O03-GH42 the inactivator 2,4-dinitrophenyl 2-deoxy-2-fluoro-β-D-galactopyranoside was used as this enzyme is not inactivated with the glucoside. These reactions were incubated at 20 °C until the wild-type enzyme displayed less than 95 % activity (typically 30 min). The inactivated enzyme was then washed three times with storage buffer to remove excess inactivator using a Viva spin 500 (10k) centrifugal filter unit (Vivaproducts). An aliquot of the inactivated enzyme was then transferred to a 96 well plate containing an array of potential reactivators at concentrations of 20 mM or 40 % of a saturated solution. This reactivation plate was then incubated for 1 hour at 25 °C. Reactivation

rates were then assessed by adding para-nitrophenyl glucoside (pNP-G) to a final concentration of 1 mM to each reaction and monitoring absorbance at 400 nm using a Synergy H1 plate reader (BioTek). For the O03-GH42 enzyme MU-3-O-Me-Gal was used instead of pNP-Glc to determine reactivation rates and the resulting fluorescence (λ_{ex} = 365, λ_{em} = 450 nm) was detected using a Synergy H1 plate reader (BioTek). Acceptor specificity assay was not performed for I01-GH1 as this enzyme wasn't inhibited by DNP 2F-Glc.

The acceptors used were: 1-Adamantanemethanol, α,α-D-trehalose, α-lactose, α-L-rhamnose, β-gentiobiose, 1,3-propanediol, 1,5-anhydro-D-glucitol, 1-butanol, 1-ethynylcyclohexanol, 1-hexanol, 1-naphthol, 1-octanol, 1-pentanol, 1-propanol, 1-pyrenemethanol, 2-mercaptoethanol, 2-methoxyethanol 2-naphthol, 2-propanol, 3-mercapto-1-propanol 4-(hexyloxy)phenol, 4-methylumbelliferyl cellobioside, 4-methylumbelliferyl β-D-galactopyranoside, 4-methylumbelliferyl β-D-glucopyranoside, 4-methylumbilliferyl β-D-xylopyranoside, 4-vinylphenol, 5-hexyne-1-ol, asparagine, caffeic acid, cyclohexanol, D/L-threitol, D-allose, D-arabitol, D-cellobiose, D-fructose, D-galactose, D-lyxose, D-mannitol, D-mannose, D-ribose, D-tagatose, D-xylose, ethanediol, ethanol, galactal, galactitol, gallic acid, glucal, glucose, inositol, L-arabinose, L-arabitol, L-arginine, L-cysteine, L-erythritol, L-fucose, L-serine, L-sorbose, L-threonine, L-tyrosine, maltose, maltotriose, methanol, o-Phenylphenol, p-nitrophenyl α-D-galactopyranoside p-nitrophenyl α-D-mannopyranoside, p-nitrophenyl α-D-xylopyranoside, p-nitrophenyl α-L-arabinopyranoside, p-nitrophenyl β-D-cellobioside, p-nitrophenyl β-D-fucopyranoside, p-nitrophenyl β-D-galactopyranoside, p-nitrophenyl β-D-glucopyranoside, p-nitrophenyl β-D-glucuronide, p-nitrophenyl β-D-lactopyranoside, p-nitrophenyl β-D-mannopyranoside, p-nitrophenyl β-D-xylopyranoside, phenethyl alcohol, phenol, phenyl β-D-galactopyranoside, phenyl β-D-glucopyranoside, phloroglucinol, p-methoxyphenol, p-phenylphenol, raffinose, resorcinol, sorbitol, and sucrose.

6.5.7 Glycosynthase Reactions: Metagenomic Hits

Glycosynthase activity for each of the generated nucleophile variants was first assessed with the unmodified α-glycosyl fluoride as a donor. In the case of the nucleophile variants derived from C08-GH3, C24-GH3-1, D08-GH3 and O22-GH3 the donor used was α-D-glucopyranosyl fluoride

(αF-Glc), while the donor used for C11-GH1 and O03-GH42 was α-D-galactopyranosyl fluoride (αF-gal). The donor sugar used for I0S-GH1, 6-phospho-α-D-glucopyranosyl fluoride (6-PO_4-αF-Glc) was generated *in situ* by *Escherichia coli* ATP-Dependent Glucokinase (EcGlk) [203] in the presence of ATP (10 mM) and $MgSO_4$ (10 mM). The acceptors used in the assay were chosen as the top three hits from the acceptor specificity assay. Reactions were performed on a 100 μL scale with 10 mM donor sugar and 10 mM acceptor in reaction buffer (100 mM HEPES, 100 mM NaCl, pH 7.0). Reactions were incubated at 37 °C for 18 hours, after which point they were monitored by thin-layer chromatography (TLC). TLC was performed on aluminum-backed sheets of Silica Gel $60F_{254}$ (E. Merck) of thickness 0.2 mm. The plates were visualised using UV light (254nm) and/or by exposure to 10% ammonium molybdate (2M in H_2SO_4) followed by charring. Reactions displaying product spots were sent for mass spectrometry analysis and selected for multi-milligram scale reactions. Enzymes which displayed activity were then assayed with the appropriate amino-, azido- or methoxy-α-fluorosugar as a donor.

6.5.8 Multi-milligram Scale Reactions: Metagenomic Hits

Large-scale reactions contained 1.5 μM of the applicable glycosynthase, 2 mM donor sugar (α-D-glucopyranosyl fluoride (αF-Glc), 6-azido-6-deoxy-α-D-galactoyranosyl fluoride (αF-6-N_3-Gal) or 6-azido-6-deoxy-α-D-glucopyranosyl fluoride (αF-6-N_3-Glc), 10 mM acceptor molecule (pNP β-D-glucopyranoside (pNP-Glc), or pNP α-D-xylopyranoside (pNP-α-xyl) and buffer (100 mM Hepes, 100 mM NaCl, pH 7.0). These reactions were set up on a 25 mL scale and incubated at 25 °C for 18 hours with gentle agitation. Reactions were terminated by boiling for 10 minutes. This was followed by centrifugation to remove precipitated protein and storage at -80 °C. Reactions were then lyophilized. Solid products were suspended in 500 μL of 5 % acetonitrile in water, passed through a Millipore Ultrafree MC centrifugal column (PVDF, 0.22 μM), then loaded on a C-18 column (ZORBAX Eclipse XDB-C18, 9.4 mm x 250 mm, Agilent). A gradient of 5-10% acetonitrile in water was used to elute the product. Absorbance at 300 nm was monitored and fractions corresponding to major products were pooled. Products were lyophilized then prepared appropriately for mass spectrometry and NMR

6.5.9 Screening: GH1 library

Two 96 deep well plates containing 800 μL of LBE-5052 [285] (50 μg/L of carbenicilin) were inoculated with (5 μL per well) the library of GH1 enzymes described in Heins et al [117]. Plates were incubated at 37 °C for 18 hours with shaking at 225 rpm. Cells were harvested by centrifugation at 3220 x *g* for 30 minutes. The supernatant was then removed and 300 μL of lysis buffer (0.3 mg/mL lysozyme, 1 % triton X-100, cOmplete protease inhibitor [1 tablet/50 mL] and benzonase), was added to each well. The plates were incubated, with shaking for 2 hours at 25 °C. Lysate from each well (20 μL) was added to 96 well plates containing 280 μL of reaction buffer (20 mM sodium acetate, pH 6.0 and 107 μL of a fluorogenic substrate). The fluorogenic substrates used were: MU-3-NH_2-Glc, MU-4-NH_2-Glc, MU-6-NH_2-Glc, 3-N_3-MU-Glc, MU-4-N_3-Glc, MU-6-N_3-Glc, and MU-Glc. After 18 hours reactions were terminated by diluting 20 μL with 100 μL of stop buffer (1 M Glycine, pH 10.0). Fluorescence (λ_{ex}=360 nm and λ_{em}= 465 nm) was then measured with a Synergy H1 plate reader (BioTek).

6.5.10 Protein Purification: GH1 Library

Both variant and wild-type enzymes were purified with polyhistidine tags and Ni-NTA resin columns. Cultures of 50 mL LBE-5052[285] containing 50 μg/L of carbenicilin were inoculated with 5 μL of the expression host. Reactions were incubated with shaking (250 rpm) at 37 °C for 18 hours. Cultures were centrifuged (3220 x g, 4 °C, 20 min), the supernatant was removed and cell pellets were stored at -80 °C until purification. To purify proteins 2.5 mL of lysis mix (1 x BugBuster [Novagen], 20 mM HEPES, 300 mM NaCl, 20 mM Imidazole, pH 7.0) was used to resuspend thawed cell pellets. This suspension was incubated at 20 °C for 20 minutes, after which the lysate was clarified by centrifugation (3220 x g, 4 °C, 20 min) and loaded onto columns containing 1 mL of HisPur resin (ThermoScientific). Columns were washed with 20 mL of Buffer A (20 mM HEPES, 300 mM NaCl, 20 mM Imidazole, pH 7.0) and protein was eluted with 4 mL of Buffer B (20 mM HEPES, 300 mM NaCl, 500 mM Imidazole, pH 7.0). Proteins were buffer-exchanged into storage buffer (20 mM HEPES, 300 mM NaCl, pH 7.0) with Amicon 30 kDa filter columns and stored at 4 °C. Protein concentrations were determined based on absorbance at 280 nm. Extinction coefficients

are as follows : Ali_GH1: ε = 121,365 $M^{-1}cm^{-1}$, Dei_GH1: ε = 110,365 $M^{-1}cm^{-1}$, Exi_GH1: ε = 113,135 $M^{-1}cm^{-1}$, Lac_GH1: ε = 113,470 $M^{-1}cm^{-1}$, Myx_GH1: ε = 108,540 $M^{-1}cm^{-1}$, Pha_GH1: ε = 125,375 $M^{-1}cm^{-1}$, Sac_GH1: ε = 123,675 $M^{-1}cm^{-1}$, The_GH1: ε = 108,540 $M^{-1}cm^{-1}$.

6.5.11 Acceptor Specificity Screening: GH1 Library

Acceptor specificity screening generally followed the procedures detailed by Blanchard et. al. [29] First, between 0.5 and 3 nanomoles of purified wild-type enzyme was incubated with 1 mM of dintrophenyl 2-deoxy-2-fluoro-β-D-glucopyranoside (DNP 2F-Glc). This reaction was incubated at 20 °C until the wildtype enzyme displayed greater than 95 % inhibition. The inactivated enzyme was then washed with storage buffer to remove excess inactivator using a Viva spin 500 (10 k) centrifugal filter unit (Vivaproducts). An aliquot of the inactivated enzyme was then transferred to a 96 well plate containing an array of potential reactivators at concentrations of 20 mM or 40 % of a saturated solution. This reactivation plate was then incubated for 1 hour at 25 °C. Reactivation rates were then assessed by adding para-nitrophenyl glucoside (pNP-G) to a final concentration of 1 mM to each reaction and monitoring absorbance at 400 nm using a Synergy H1 plate reader (BioTek)

The acceptors used were: 1-propanol, 2-Naphthol, 2-propanol, 4-hydroxycoumarin, 4-methylumbelliferyl β-D-xyloside, 4-methylumbelliferyl a-D-glucopyranoside, 4-methylumbelliferyl α-L-arabinopyranoside, 4-methylumbelliferyl β-D-cellobiopyranoside, 4-methylumbelliferyl β-D-galactopyranoside, 4-methylumbelliferyl β-D-glucopyranoside, 4-methylumbelliferyl β-D-glucuronide dihydrate, 4-methylumbelliferyl lactoside, 4-methylumbelliferyl *N*-acetyl-glucosaminide, 8-hydroxy-quinoline, α-L-fucose, α-L-rhamnose, β-gentiobiose, caffeic acid, citric acid, cyclohexanol, D-araboascorbic acid, D-galactose, D-glucosamine, D-maltose, D-trehalose, D-xylose, D-arabitol, cellobiose, D-galactosamine, D-fructose, D-fructose 1,6-diphosphate, D-galactal, D-galacturonic acid, D-glucoheptose, D-gluconic acid, D-gluconic acid lactone, D-glucose, D-glucose 6-phosphate, D-glucuronic acid, D-gulonic-γ-lactone, D-lyxose, D-mannitol, D-mannose, D-mannose-6 phosphate, D-tagatose, dithiothreitol, gallic acid, geraniol, glycine, inositol, L-fucose, L-arabinose, L-ascorbic acid, lactose, L-cysteine, levulinic acid, L-xylose, maltotriose, methyl α-D-mannopyranoside, methyl α-L-rhamnoside, methyl β-D-

galactopyranoside, methyl β-D-xylopyranoside, *N*-acetyl-D-glucosamine, *N*-acetyl-mannosamine, *N*-acetylneuraminic acid, octyl-β-D-glucopyranoside, palatinose, phenethyl alcohol, phenyl β-D-galactoside, phenyl β-D-glucopyranoside, p-nitrophenyl β-D-glucuronide, p-nitrophenyl α-D-mannopyranoside, p-nitrophenyl β-D-fucopyranoside, p-nitrophenyl *N*-acetyl-β-D-glucosaminide, p-nitrophenyl α-L-arabinopyranoside, p-nitrophenyl β-D-glucopyranoside, p-nitrophenyl β-D-xylopyranoside, p-nitrophenyl β-cellobioside, p-nitrophenyl β-D-galactopyranoside, p-nitrophenyl α-D-galactopyranoside, p-nitrophenyl α-D-glucopyranoside, quercetin, raffinose hydrate, sialic acid, and sodium azide.

6.5.12 Wild-Type Enzyme Kinetics: GH1 Library

Kinetic parameters for wild-type enzymes were determined using fluorogenic or chromogenic substrates. Assays with fluorogenic substrates were performed in 96 well plates (Corning 3370) containing the 4-methylumbelliferyl glycoside (0.5 μM - 1 mM), buffer (20 mM HEPES, 300 mM NaCl, pH 7.0) and purified enzyme. Reactions were performed at 30 °C and fluorescence (λ_{ex}=360 nm and λ_{em}= 465 nm, gain = 75) was monitored using a Synergy H1 plate reader (BioTek). The quantity of fluorophore generated was determined by means of a calibration curve of 4-methylumbelliferyl alcohol within an identical buffer system. Rate measurements for chromogenic reagents (para-nitrophenyl 6-deoxy-6-phospho-β-D-glucopyranoside, para-nitrophenyl β-D-glucopyranoside) were performed using a Cary3000 spectrophotometer (Agilent). Reaction buffer was the same as for the fluorogenic substrates and temperature was also maintained at 30 °C. Reactions were monitored at 400 nm (ε = 9.42 mM^{-1} cm^{-1}). All reactions were performed in triplicate. Rate measurements for both chromogenic and fluorogenic substrates were used to calculate kinetic parameters with the software program GraFit 7.0 software.

6.5.13 Production of Mutants: GH1 Library

Nucleophile mutants were generated using the same protocol used for the generation of mutants from metagenome-sourced hydrolases. Table 6.8 details the primers used for QuikChange mutagenesis.

6.5.14 Glycosynthase Reactions: GH1 Library

The inhibitor 2,4-dinitrophenyl 2-deoxy-2-fluoro-glucopyranoside were synthesized as previously described [175, 320]. All modified and unmodified α-glycosyl fluorides were synthesized by Dr. Hong-Ming Chen.

To determine the best nucleophile variant for each enzyme, glycosynthase reactions were performed in triplicate, and analysed by HPLC. Glycosynthase reactions (50 μL scale) contained 20 μM enzyme, 5 mM α-glucosyl fluoride, 5 mM pNP-glucopyranoside or 5 mM pNP cellobioside in reaction buffer (100 mM HEPES, 100 mM NaCl, pH 7.0). Reactions were incubated at 25 °C for 18 hours, after which point they were diluted with 500 μL of ethanol and centrifuged to remove precipitated protein. An aliquot (1 μL) of the reaction was loaded on a C-18 column (Poroshell 120 EC-C18, 4.6 mm x 50 mm, Agilent). A gradient of 0-10% acetonitrile in water was used to elute the product and the absorbance at 300 nm was monitored and peak area was quantified. Activity with amino and azido donor sugars was performed on a 25 μL scale with 20 μM enzyme, 50 mM donor sugar, 10 mM pNP-glucopyranoside or 10 mM pNP cellobioside in reaction buffer (100 mM HEPES, 100 mM NaCl, pH 7.0). Reactions were incubated at 25 °C for 18 hours, after which point they were monitored by thin-layer chromatography (TLC). TLC was performed on aluminium-backed sheets of Silica Gel $60F^{254}$ (E. Merck) of thickness 0.2mm. The plates were visualised using UV light (254nm) and/or by exposure to 10% ammonium molybdate (2M in H_2SO_4) followed by charring. Reactions displaying product spots were sent for mass spectrometry analysis and selected for multi-milligram scale reactions. Reactions with alternate acceptors were performed as above.

6.5.15 Multi-milligram Scale Reactions: GH1 Library

Large-scale reactions contained 20 μM of the applicable glycosynthase, 10 mM donor sugar (α-D-glucopyranosyl fluoride (αF-Glc), 3-amino-3-deoxy-α-D-glucopyranosyl fluoride (αF-3-NH_2-Glc), 4-amino-4-deoxy-α-D-glucopyranosyl fluoride (αF-4-NH_2-Glc), 6-amino-6-deoxy-α-D-glucopyranosyl fluoride (αF-6-NH_2-Glc), 3-azido-3-deoxy-α-D-glucopyranosyl fluoride (αF-3-N_3-Glc), 4-azido-4-deoxy-α-D-glucopyranosyl fluoride (αF-4-N_3-Glc) or 6-azido-6-deoxy-α-D-glucopyranosyl fluoride (αF-6-N_3-Glc), 10 mM acceptor molecule (pNP glucopyranoside (pNP-Glc), pNP cellobioside (pNP-

C), pNP xylopyranoside (pNP-xyl), n-octyl glucoside or DNP 2-deoxy-2-fluoro-glucopyranoside (DNP 2F-Glc) and buffer (100 mM Hepes, 100 mM NaCl, pH 7.0). These reactions were set up on a 3.2 mL scale and incubated at 25 °C for 18 hours with gentle agitation. Reactions were terminated by boiling for 5 minutes. This was followed by centrifugation to remove precipitated protein and storage at -80 °C. Reactions were then lyophilized. Solid products were suspended in 500 μL of 5 % acetonitrile in water, passed through a Millipore Ultrafree MC centrifugal column (PVDF -.22 μM), then loaded on a C-18 column (ZORBAX Eclipse XDB-C18, 9.4 mm x 250 mm, Agilent). A gradient of 5-10% acetonitrile in water was used to elute the product. Absorbance at 300 nm was monitored and fractions corresponding to major products were pooled. Products were lyophilized then prepared appropriately for mass spectrometry and NMR

6.5.16 Mass Spectrometry and NMR Spectroscopy of Products

Proton and carbon NMR spectra were recorded on Bruker Advance 400inv, 400dir and a 300 Fourier Transform spectrometer fitted with a 5mm BBI-Z probe. All spectra were recorded using an internal deuterium lock and are referenced internally using the residual solvent peak. Carbon and proton chemical shifts are quoted in parts per million (ppm) downfield of tetramethylsilane. Coupling constants (J) are given in Hertz (Hz). Carbon NMR spectra were acquired with broadband proton decoupling and were recorded with DEPT. Mass spectra were measured on a Waters/Micromass LCT using electrospray ionisation (ESI) using methanol as solvent.

Table 6.7: Primers Used for Mutagenesis of Metagenome Sourced Hydrolases

Primer	Sequence
B09O03_GH42_E309A_F	GTTTCTGCTCATGGCGCAGACGCCGAGC
B09O03_GH42_E309A_R	GCTCGGCGTCTGCGCCATGAGCAGAAAC
B09O03_GH42_E309S_F	CGTTTCTGCTCATGAGCCAGACGCCGAGCGTG
B09O03_GH42_E309S_R	CACGCTCGGCGTCTGGCTCATGAGCAGAAACG
B09O03_GH42_E309G_F	GTTTCTGCTCATGGGCCAGACGCCGAGCG
B09O03_GH42_E309G_R	CGCTCGGCGTCTGGCCCATGAGCAGAAAC
FOS6240O22_GH3_D231A_F	GTTACGTGATGACGGCGTGGGGCGCAATGAAC
FOS6240O22_GH3_D231A_R	GTTCATTGCGCCCCACGCCGTCATCACGTAAC
FOS6240O22_GH3_D231S_F	GTTACGTGATGACGAGCTGGGGCGCAATG
FOS6240O22_GH3_D231S_R	CATTGCGCCCCAGCTCGTCATCACGTAAC
FOS6240O22_GH3_D231G_F	GTTACGTGATGACGGGCTGGGGCGCAATGAAC
FOS6240O22_GH3_D231G_R	GTTCATTGCGCCCCAGCCCGTCATCACGTAAC
FOS6241C11_GH1_E354A_F	CCTGCCGCTTATTATTACCGCAAACGGGATGGCGGACAACGAC
FOS6241C11_GH1_E354A_R	GTCGTTGTCCGCCATCCCGTTTGCGGTAATAATAAGCGGCAGG
FOS6241C11_GH1_E354S_F	CCTGCCGCTTATTATTACCTCAAACGGGATGGCGGACAACGAC
FOS6241C11_GH1_E354S_R	GTCGTTGTCCGCCATCCCGTTTGAGGTAATAATAAGCGGCAGG
FOS6241C11_GH1_E354G_F	CCTGCCGCTTATTATTACCGGAAACGGGATGGCGGACAACGAC
FOS6241C11_GH1_E354G_R	GTCGTTGTCCGCCATCCCGTTTCCGGTAATAATAAGCGGCAGG
NapDC14D08_GH3_D229A_F	GTTTTGTGGTTTCTGCGTGGGGAGCTGTGCATG
NapDC14D08_GH3_D229A_R	CATGCACAGCTCCCCACGCAGAAACCACACAAAAC
NapDC14D08_GH3_D229S_F	GAAGGTTTTGTGGTTTCTAGCTGGGGAGCTGTGCATGAC
NapDC14D08_GH3_D229S_R	GTCATGCACAGCTCCCCAGCTAGAAACCACACAAAACCTTC
NapDC14D08_GH3_D229G_F	GTTTTGTGGTTTCTGGCTGGGGAGCTGTGCATG
NapDC14D08_GH3_D229G_R	CATGCACAGCTCCCCAGCCAGAAACCACACAAAAC
TolDC15C08_GH3_D235A_F	CGGTCGTCTCCGCGTGGTTTGCGAC
TolDC15C08_GH3_D235A_R	GTCGCAAACCACGCGGAGACGACCG
TolDC15C08_GH3_D235S_F	CGCGGTCGTCTCCAGCTGGTTTGCGAC
TolDC15C08_GH3_D235S_R	GTCGCAAACCAGCTGGAGACGACCGCG
TolDC15C08_GH3_D235G_F	CGGTCGTCTCCGGCTGGTTTGCGAC
TolDC15C08_GH3_D235G_R	GTCGCAAACCAGCCGGAGACGACCG
CA23302C24_GH3_1_D271A_F	CGTCATGATGTCCGCGTGGTTTGCGACTTAC
CA23302C24_GH3_1_D271S_F	GGCGTCATGATGTCCAGCTGGTTTGCGACTTAC
CA23302C24_GH3_1_D271G_F	GTCATGATGTCCGGCTGGTTTGCGAC
CA23302C24_GH3_1_D271A_R	GTAAGTCGCAAACCACGCGGACATCATGACG
CA23302C24_GH3_1_D271S_R	GTAAGTCGCAAACCAGCTGGACATCATGACGCC
CA23302C24_GH3_1_D271G_R	GTCGCAAACCAGCCGGACATCATGAC
CG23A23I01_GH1_E374A_F	CAAGATTTATATTACCGAGCGTGGTCTTGGTGATGAAGATC
CG23A23I01_GH1_E374A_R	GATCTTCATCACCAAGACCACGCTCGGTAATATAAATCTTG
CG23A23I01_GH1_E374S_F	GTCAAGATTTATATTACCGAAGCTGGTCTTGGTGATGAAGATCC
CG23A23I01_GH1_E374S_R	GGATCTTCATCACCAAGACCAGCTTCGGTAATATAAATCTTGAC
CG23A23I01_GH1_E374G_F	CAAGATTTATATTACCGAGGCTGGTCTTGGTGATGAAGATC
CG23A23I01_GH1_E374G_R	GATCTTCATCACCAAGACCAGCCTCGGTAATATAAATCTTG
O22_GH3_D231S_W232F_F	CGTGATGACGAGCTTTGGCGCAATGAACAAC
O22_GH3_D231S_W232F_R	GTTGTTCATTGCGCCAAAGCTCGTCATCACG
D08_GH3_D229S_W230F_F	GTTTTGTGGTTTCTAGCTTTGGAGCTGTGCATGACAG
D08_GH3_D229S_W230F_R	CTGTCATGCACAGCTCCAAAGCTAGAAACCACAAAAC
C24_GH3_1_D271S_W272F_F	GCGTCATGATGTCCAGCTTCTTTGCGACTTACGACGGTG
C24_GH3_1_D271S_W272F_R	CACCGTCGTAAGTCGCAAAGAAGCTGGACATCATGACGC
C08_GH3_D235S_W236F_F	GCGGTCGTCTCCAGCTTCTTTGCGACCCATTCCAC
C08_GH3_D235S_W236F_R	GTGGAATGGGTCGCAAAGAAGCTGGAGACGACCGC

Table 6.8: Primers Used in QuikChange Mutagenesis of Selected GH1 Enzymes

Primer Name	Sequence
AliGH1_E354A_F	CATTCCGATCTACATTACTGCAAATGGCGCAGCCTTTGATG
AliGH1_E354A_R	CATCAAAGGCTGCGCCATTTGCAGTAATGTAGATCGGAATG
AliGH1_E354G_F	CATTCCGATCTACATTACTGGAAATGGCGCAGCCTTTGATG
AliGH1_E354G_R	CATCAAAGGCTGCGCCATTTCCAGTAATGTAGATCGGAATG
AliGH1_E354S_F	CATTCCGATCTACATTACTTCAAATGGCGCAGCCTTTGATG
AliGH1_E354S_R	CATCAAAGGCTGCGCCATTTGAAGTAATGTAGATCGGAATG
DeiGH1_E346A_F	CACCGATGTACATTACCGCAAATGGTGCAGCCTATC
DeiGH1_E346A_R	GATAGGCTGCACCATTTGCGGTAATGTACATCGGTG
DeiGH1_E346G_F	CACCGATGTACATTACCGGAAATGGTGCAGCCTATC
DeiGH1_E346G_R	GATAGGCTGCACCATTTCCGGTAATGTACATCGGTG
DeiGH1_E346S_F	CACCGATGTACATTACCTCAAATGGTGCAGCCTATC
DeiGH1_E346S_R	GATAGGCTGCACCATTTGAGGTAATGTACATCGGTG
MyxGH1_E357A_F	CCCTTTGTACATTACAGCAAATGGTTGCGCCTATG
MyxGH1_E357A_R	CATAGGCGCAACCATTTGCTGTAATGTACAAAGGG
MyxGH1_E357G_F	CCCTTTGTACATTACAGGAAATGGTTGCGCCTATG
MyxGH1_E357G_R	CATAGGCGCAACCATTTCCTGTAATGTACAAAGGG
MyxGH1_E357S_F	GCCCTTTGTACATTACATCAAATGGTTGCGCCTATGC
MyxGH1_E357S_R	GCATAGGCGCAACCATTTGATGTAATGTACAAAGGGC
PhaGH1_E365A_F	CAGTTTATGTGACAGCAAATGGCTTCCCTG
PhaGH1_E365A_R	CAGGGAAGCCATTTGCTGTCACATAAACTG
PhaGH1_E365G_F	CAGTTTATGTGACAGGAAATGGCTTCCCTG
PhaGH1_E365G_R	CAGGGAAGCCATTTCCTGTCACATAAACTG
PhaGH1_E365S_F	GATAAGCCAGTTTATGTGACATCAAATGGCTTCCCTGTTAAAGG
PhaGH1_E365S_R	CCTTTAACAGGGAAGCCATTTGATGTCACATAAACTGGCTTATC
SacGH1_E368A_F	CTGATATCTATATCACTGCAAACGGTTGCGCCCTGC
SacGH1_E368A_R	GCAGGGCGCAACCGTTTGCAGTGATATAGATATCAG
SacGH1_E368G_F	CTGATATCTATATCACTGGAAACGGTTGCGCCCTGC
SacGH1_E368G_R	GCAGGGCGCAACCGTTTCCAGTGATATAGATATCAG
SacGH1_E368S_F	CTGATATCTATATCACTTCAAACGGTTGCGCCCTG
SacGH1_E368S_R	CAGGGCGCAACCGTTTGAAGTGATATAGATATCAG
TheGH1_E388A_F	GTTGTACATCACCGCAAACGGTGCAGCCTTCGAAG
TheGH1_E388A_R	CTTCGAAGGCTGCACCGTTTGCGGTGATGTACAAC
TheGH1_E388G_F	CCGTTGTACATCACCGGAAACGGTGCAGCCTTCGAAG
TheGH1_E388G_R	CTTCGAAGGCTGCACCGTTTCCGGTGATGTACAACGG
TheGH1_E388S_F	CTTACCGTTGTACATCACCTCAAACGGTGCAGCCTTCGAAG
TheGH1_E388S_R	CTTCGAAGGCTGCACCGTTTGAGGTGATGTACAACGGTAAG
ExiGH1_E350A_F	CTATCTATATCACTGCAAACGGTGCCGCGTTC
ExiGH1_E350A_R	GAACGCGGCACCGTTTGCAGTGATATAGATAG
ExiGH1_E350G_F	CTATCTATATCACTGGAAACGGTGCCGCGTTC
ExiGH1_E350G_R	GAACGCGGCACCGTTTCCAGTGATATAGATAG
ExiGH1_E350S_F	GCCTATCTATATCACTAGCAACGGTGCCGCGTTCG
ExiGH1_E350S_R	CGAACGCGGCACCGTTGCTAGTGATATAGATAGGC
LacGH1_E366A_F	CATGGTTTGTTGCCGCAAATGGTATTGGCG
LacGH1_E366A_R	CGCCAATACCATTTGCGGCAACAAACCATG
LacGH1_E366G_F	CATGGTTTGTTGCCGGAAATGGTATTGGCG
LacGH1_E366G_R	CGCCAATACCATTTCCGGCAACAAACCATG
LacGH1_E366S_F	GCCATGGTTTGTTGCCAGCAATGGTATTGGCGTGG
LacGH1_E366S_R	CCACGCCAATACCATTGCTGGCAACAAACCATGGC

Bibliography

[1] T. Aakvik, K. F. Degnes, R. Dahlsrud, F. Schmidt, R. Dam, L. Yu, U. Völker, T. E. Ellingsen, and S. Valla. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiology Letters*, 296(2): 149–158, 2009.

[2] K. Abe, M. Nakajima, T. Yamashita, H. Matsunaga, S. Kamisuki, T. Nihira, Y. Takahashi, N. Sugimoto, A. Miyanaga, H. Nakai, T. Arakawa, S. Fushinobu, and H. Taguchi. Biochemical and structural analyses of a bacterial endo-β-1,2-glucanase reveal a new glycoside hydrolase family. *Journal of Biological Chemistry*, 292(18):7487–7506, 2017.

[3] K. E. Achyuthan, A. M. Achyuthan, P. D. Adams, S. M. Dirk, J. C. Harper, B. A. Simmons, and A. K. Singh. Supramolecular self-assembled chaos: polyphenolic lignin's barrier to cost-effective lignocellulosic biofuels. *Molecules*, 15:8641–8688, 2010.

[4] M. Al, L. J. Evans, W. Douglas Gould, W. F. A. Duncan, and S. Glasauer. The long term operation of a biologically based treatment system that removes As, S and Zn from industrial (smelter operation) landfill seepage. *Applied Geochemistry*, 26(11):1886–1896, 2011.

[5] M. Aleksiuk. The Seasonal Food Regime of Arctic Beavers. *Ecology*, 51(2):264–270, 1970.

[6] E. Allerdings, J. Ralph, H. Steinhart, and M. Bunzel. Isolation and structural identification of complex feruloylated heteroxylan side-chains from maize bran. *Phytochemistry*, 67(12): 1276–1286, 2006.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[8] D. An, S. M. Caffrey, J. Soh, A. Agrawal, D. Brown, K. Budwill, X. Dong, P. F. Dunfield, J. Foght, L. M. Gieg, S. J. Hallam, N. W. Hanson, Z. He, T. R. Jack, J. Klassen, K. M. Konwar, E. Kuatsjah, C. Li, S. Larter, V. Leopatra, C. L. Nesbo, T. Oldenburg, A. P. Page, E. Ramos-Padron, F. F. Rochman, A. Saidi-Mehrabad, C. W. Sensen, P. Sipahimalani, Y. C. Song, S. Wilson, G. Wolbring, M. L. Wong, and G. Voordouw. Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common. *Environ Science & Technology*, 47(18):10708–17, 2013.

[9] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[10] A. Angelov, V. T. T. Pham, M. belacker, S. Brady, B. Leis, N. Pill, J. Brolle, M. Mechelke, M. Moerch, B. Henrissat, and W. Liebl. A metagenome-derived thermostable β-glucanase with an unusual module architecture which defines the new glycoside hydrolase family GH148. *Scientific Reports*, 7(1), 2017.

[11] M. M. Appeldoorn, M. A. Kabel, D. Van Eylen, H. Gruppen, and H. A. Schols. Characterization of Oligomeric Xylan Structures from Corn Fiber Resistant to Pretreatment and Simultaneous Saccharification and Fermentation. *Journal of Agricultural and Food Chemistry*, 58(21):11294–11301, 2010.

[12] Z. Armstrong and S. G. Withers. Synthesis of Glycans and Glycopolymers Through Engineered Enzymes. *Biopolymers*, 99(10):666–674, 2013.

[13] Z. Armstrong, K. Mewis, C. Strachan, and S. J. Hallam. Biocatalysts for biomass deconstruction from environmental genomics. *Current Opinion in Chemical Biology*, 29:18–25, 2015.

[14] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen,

W. M. de Vos, S. Brunak, J. Dor, M. Antoln, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mrieux, R. Melo Minardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the Human Gut Microbiome. *Nature*, 473(7346):174–180, 2011.

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

[16] V. Bågenholm, S. K. Reddy, H. Bouraoui, J. Morrill, E. Kulcinskaja, C. M. Bahr, O. Aurelius, T. Rogers, Y. Xiao, D. T. Logan, E. C. Martens, N. M. Koropatkin, and H. Stålbrand. Galactomannan Catabolism Conferred by a Polysaccharide Utilization Locus of *Bacteroides ovatus*. *Journal of Biological Chemistry*, 292(1):229–243, 2017.

[17] S. G. Ball and M. K. Morell. FROM BACTERIAL GLYCOGEN TO STARCH: Understanding the Biogenesis of the Plant Starch Granule. *Annual Review of Plant Biology*, 54(1): 207–233, 2003.

[18] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[19] G. Bastien, G. Arnal, S. Bozonnet, S. Laguerre, F. Ferreira, R. Faure, B. Henrissat, F. Lefevre, P. Robe, O. Bouchez, C. Noirot, C. Dumon, and M. O'Donohue. Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics. *Biotechnology for Biofuels*, 6(1):78, 2013.

[20] M. W. Bauer, L. E. Driskill, W. Callen, M. A. Snead, E. J. Mathur, and R. M. Kelly. An

endoglucanase, EglA, from the hyperthermophilic archaeon *Pyrococcus furiosus* hydrolyzes beta-1,4 bonds in mixed-linkage (1→3),(1→4)-beta-D-glucans and cellulose. *Journal of Bacteriology*, 181(1):284–90, 1999.

[21] N. T. Baxter, J. J. Wan, A. M. Schubert, M. L. Jenior, P. Myers, P. D. Schloss, and K. E. Wommack. Intra- and Interindividual Variations Mask Interspecies Variation in the Microbiota of Sympatric Peromyscus Populations. *Applied and Environmental Microbiology*, 81(1): 396–404, 2015.

[22] E. W. Beals. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data. *Advances in Ecological Research*, 14:1–55, 1984.

[23] C. Bera, V. Broussolle, E. Forano, and G. Gaudet. Gene sequence analysis and properties of EGC, a family E (9) endoglucanase from Fibrobacter succinogenes BL2. *FEMS Microbiology Letters*, 136(1):79–84, 1996.

[24] A. Bhat, S. Riyaz-Ul-Hassan, N. Ahmad, N. Srivastava, and S. Johri. Isolation of cold-active, acidic endocellulase from Ladakh soil by functional metagenomics. *Extremophiles*, 17(2): 229–239, 2013.

[25] B. Bissaro, s. K. Rhr, G. Mller, P. Chylenski, M. Skaugen, Z. Forsberg, S. J. Horn, G. Vaaje-Kolstad, and V. G. H. Eijsink. Oxidative cleavage of polysaccharides by monocopper enzymes depends on H2O2. *Nature Chemical Biology*, 13(10):1123–1128, 2017.

[26] S. Biver, S. Steels, D. Portetelle, and M. Vandenbol. Bacillus subtilis as a tool for screening soil metagenomic libraries for antimicrobial activities. *Journal of Microbiology and Biotechnology*, 23(6), 2013.

[27] S. Biver, A. Stroobants, D. Portetelle, and M. Vandenbol. Two promising alkaline beta-glucosidases isolated by functional metagenomics from agricultural soil, including one showing high tolerance towards harsh detergents, oxidants and glucose. *Journal of Industrial Microbiology & Biotechnology*, 41(3):479–488, 2014.

[28] M. K. Bjursell, E. C. Martens, and J. I. Gordon. Functional Genomic and Metabolic Studies of the Adaptations of a Prominent Adult Human Gut Symbiont, *Bacteroides thetaiotaomicron*, to the Suckling Period. *Journal of Biological Chemistry*, 281(47):36269–36279, 2006.

[29] J. E. Blanchard and S. G. Withers. Rapid screening of the aglycone specificity of glycosidases: Applications to enzymatic synthesis of oligosaccharides. *Chemistry & Biology*, 8(7):627–633, 2001.

[30] A. Bock, K. Forchhammer, J. Heider, W. Leinfelder, G. Sawers, B. Veprek, and F. Zinoni. Selenocysteine: the 21st amino acid. *Molecular Microbiology*, 5(3):515–20, 1991.

[31] W. Boerjan, J. Ralph, and M. Baucher. Lignin biosynthesis. *Annual Review of Plant Biology*, 54:519–46, 2003.

[32] K. S. Boles, K. Kannan, J. Gill, M. Felderman, H. Gouvis, B. Hubby, K. I. Kamrud, J. C. Venter, and D. G. Gibson. Digital-to-biological converter for on-demand production of biologics. *Nature Biotechnology*, 35(7):672–675, 2017.

[33] U. T. Bornscheuer and R. J. Kazlauskas. Catalytic Promiscuity in Biocatalysis: Using Old Enzymes to Form New Bonds and Follow New Pathways. *Angewandte Chemie International Edition*, 43(45):6032–6040, 2004.

[34] E. Bouhajja, M. McGuire, M. R. Liles, G. Bataille, S. N. Agathos, and I. F. George. Identification of novel toluene monooxygenase genes in a hydrocarbon-polluted sediment using sequence- and function-based screening of metagenomic libraries. *Applied Microbiology and Biotechnology*, 101(2):797–808, 2017.

[35] F. J. Brenner. Foods Consumed by Beavers in Crawford County, Pennsylvania. *The Journal of Wildlife Management*, 26(1):104, 1962.

[36] R. Brunecky, M. Alahuhta, Q. Xu, B. S. Donohoe, M. F. Crowley, I. A. Kataeva, S. J. Yang, M. G. Resch, M. W. Adams, V. V. Lunin, M. E. Himmel, and Y. J. Bomble. Revealing nature's cellulase diversity: the digestion mechanism of Caldicellulosiruptor bescii CelA. *Science*, 342 (6165):1513–6, 2013.

[37] M. S. Buckeridge, H. Pessoa dos Santos, and M. A. S. Tin. Mobilisation of storage cell wall polysaccharides in seeds. *Plant Physiology and Biochemistry*, 38(1-2):141–156, 2000.

[38] R. R. Buech. Ontogeny and diurnal cycle of fecal reingestion in the North American beaver (*Castor canadensis*). *Journal of Mammalogy*, 65(2):347–350, 1984.

[39] L. Burhenne, J. Messmer, T. Aicher, and M.-P. Laborie. The effect of the biomass components lignin, cellulose and hemicellulose on TGA and fixed bed pyrolysis. *Journal of Analytical and Applied Pyrolysis*, 101:177–184, 2013.

[40] W. F. Burkholder, R. M. Weiner, L. E. Taylor, B. Henrissat, L. Hauser, M. Land, P. M. Coutinho, C. Rancurel, E. H. Saunders, A. G. Longmire, H. Zhang, E. A. Bayer, H. J. Gilbert, F. Larimer, I. B. Zhulin, N. A. Ekborg, R. Lamed, P. M. Richardson, I. Borovok, and S. Hutcheson. Complete Genome Sequence of the Complex Carbohydrate-Degrading Marine Bacterium, *Saccharophagus degradans* Strain 2-40T. *PLOS Genetics*, 4(5):e1000087, 2008.

[41] P. E. Busher. Food Caching Behavior of Beavers (*Castor canadensis*): Selection and Use of Woody Species. *American Midland Naturalist*, 135(2):343–348, 1996.

[42] M. Busse-Wicher, A. Li, R. L. Silveira, C. S. Pereira, T. Tryfona, T. C. F. Gomes, M. S. Skaf, and P. Dupree. Evolution of xylan substitution patterns in gymnosperms and angiosperms: implications for xylan interaction with cellulose. *Plant Physiology*, page pp.00539.2016, 2016.

[43] K. H. Caffall and D. Mohnen. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research*, 344(14):1879–1900, 2009.

[44] P. Capek, J. Alföldi, and D. Likov. An acetylated galactoglucomannan from *Picea abies* L. Karst. *Carbohydrate Research*, 337(11):1033–1037, 2002.

[45] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–3, 2009.

[46] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello,

N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[47] A. Carroll and C. Somerville. Cellulosic Biofuels. *Annual Review of Plant Biology*, 60(1): 165–182, 2009.

[48] A. Cartmell, E. C. Lowe, A. Basl, S. J. Firbank, D. A. Ndeh, H. Murray, N. Terrapon, V. Lombard, B. Henrissat, J. E. Turnbull, M. Czjzek, H. J. Gilbert, and D. N. Bolam. How members of the human gut microbiota overcome the sulfation problem posed by glycosaminoglycans. *Proc Natl Acad Sci*, 114(27):7037–7042, 2017.

[49] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(Database issue):D742–53, 2012.

[50] P. Chang, X. Chen, C. Smyrniotis, A. Xenakis, T. Hu, C. Bertozzi, and P. Wu. Metabolic Labeling of Sialic Acids in Living Animals with Alkynyl Sugars. *Angewandte Chemie International Edition*, 48(22):4030–4033, 2009.

[51] P. V. Chang, D. H. Dube, E. M. Sletten, and C. R. Bertozzi. A Strategy for the Selective Imaging of Glycans Using Caged Metabolic Precursors. *Journal of the American Chemical Society*, 132(28):9516–9518, 2010.

[52] H. M. Chen, Z. Armstrong, S. J. Hallam, and S. G. Withers. Synthesis and evaluation of a series of 6-chloro-4-methylumbelliferyl glycosides as fluorogenic reagents for screening metagenomic libraries for glycosidase activity. *Carbohydrate Research*, 421:33–9, 2016.

[53] J. Cheng and T. C. Charles. Novel polyhydroxyalkanoate copolymers produced in Pseudomonas putida by metagenomic polyhydroxyalkanoate synthases. *Applied Microbiology and Biotechnology*, 100(17):7611–7627, 2016.

[54] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. Müller, T. Wetter, and S. Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, 14(6):1147–1159, 2004.

[55] S. R. Chhabra and R. M. Kelly. Biochemical characterization of *Thermotoga maritima* endoglucanase Cel74 with and without a carbohydrate binding module (CBM). *FEBS Letters*, 531(2):375–80, 2002.

[56] C. Choi. Tierra del Fuego: the beavers must die. *Nature*, 453(7198):968–968, 2008.

[57] S. P. S. Chundawat, G. T. Beckham, M. E. Himmel, and B. E. Dale. Deconstruction of Lignocellulosic Biomass to Fuels and Chemicals. *Annual Review of Chemical and Biomolecular Engineering*, 2(1):121–145, 2011.

[58] L. Clarke and J. Carbon. A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire *E. coli* genome. *Cell*, 9(1):91–99, 1976.

[59] J. B. Clayton, P. Vangay, H. Huang, T. Ward, B. M. Hillmann, G. A. Al-Ghalith, D. A. Travis, H. T. Long, B. V. Tuan, V. V. Minh, F. Cabana, T. Nadler, B. Toddes, T. Murphy, K. E. Glander, T. J. Johnson, and D. Knights. Captivity humanizes the primate microbiome. *Proc Natl Acad Sci*, 113(37):10376–10381, 2016.

[60] B. Cobucci-Ponzano and M. Moracci. Glycosynthases as tools for the production of glycan analogs of natural products. *Natural Product Reports*, 29(6):697–709, 2012.

[61] B. Cobucci-Ponzano, A. Strazzulli, M. Rossi, and M. Moracci. Glycosynthases in Biocatalysis. *Advanced Synthesis & Catalysis*, 353(13):2284–2300, 2011.

[62] V. Codera, K. J. Edgar, M. Faijes, and A. Planas. Functionalized Celluloses with Regular Substitution Pattern by Glycosynthase-Catalyzed Polymerization. *Biomacromolecules*, 17(4): 1272–1279, 2016.

[63] P.-Y. Colin, B. Kintses, F. Gielen, C. M. Miton, G. Fischer, M. F. Mohamed, M. Hyvönen, D. P. Morgavi, D. B. Janssen, and F. Hollfelder. Ultrahigh-throughput discovery of promis-

cuous enzymes by picodroplet functional metagenomics. *Nature Communications*, 6:10008, 2015.

[64] S. Comtet-Marre, N. Parisot, P. Lepercq, F. Chaucheyras-Durand, P. Mosoni, E. Peyretaillade, A. R. Bayat, K. J. Shingfield, P. Peyret, and E. Forano. Metatranscriptomics Reveals the Active Bacterial and Eukaryotic Fibrolytic Communities in the Rumen of Dairy Cow Fed a Mixed Diet. *Frontiers in Microbiology*, 8, 2017.

[65] D. J. Cosgrove. Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology*, 6(11): 850–61, 2005.

[66] D. J. Cosgrove and M. C. Jarvis. Comparative structure and biomechanics of plant primary and secondary cell walls. *Frontiers in Plant Science*, 3, 2012.

[67] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*, 326(5960): 1694–1697, 2009.

[68] M. Cotta and R. Forster. *The Family Lachnospiraceae, Including the Genera Butyrivibrio, Lachnospira and Roseburia*. Springer, 2006.

[69] J. W. Craig, F. Y. Chang, J. H. Kim, S. C. Obiajulu, and S. F. Brady. Expanding Small-Molecule Functional Metagenomics through Parallel Screening of Broad-Host-Range Cosmid Environmental DNA Libraries in Diverse Proteobacteria. *Applied and Environmental Microbiology*, 76(5):1633–1641, 2010.

[70] A. Currier, W. D. Kitts, and C. I. Cellulose Digestion in The Beaver (*Castor canadensis*). *Canadian Journal of Zoology*, 38:1109–1116, 1960.

[71] F. Cuskin, E. C. Lowe, M. J. Temple, Y. Zhu, E. A. Cameron, N. A. Pudlo, N. T. Porter, K. Urs, A. J. Thompson, A. Cartmell, A. Rogowski, B. S. Hamilton, R. Chen, T. J. Tolbert, K. Piens, D. Bracke, W. Vervecken, Z. Hakki, G. Speciale, J. L. Munz-Munz, A. Day, M. J. Pea, R. McLean, M. D. Suits, A. B. Boraston, T. Atherly, C. J. Ziemer, S. J. Williams, G. J.

Davies, D. W. Abbott, E. C. Martens, and H. J. Gilbert. Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature*, 517(7533):165–169, 2015.

[72] P. M. Danby and S. G. Withers. Advances in Enzymatic Glycoside Synthesis. *ACS Chemical Biology*, 11(7):1784–1794, 2016.

[73] R. Daniel. The metagenomics of soil. *Nature Reviews Microbiology*, 3(6):470–478, 2005.

[74] L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 2013.

[75] M. M. de O. Buanafina. Feruloylation in Grasses: Current and Future Perspectives. *Molecular Plant*, 2(5):861–872, 2009.

[76] P. M. de Souza and P. de Oliveira Magalhaes. Application of microbial alpha-amylase in industry - A review. *Brazilian Journal of Microbiology*, 41(4):850–61, 2010.

[77] T. C. Delport, M. L. Power, R. G. Harcourt, K. N. Webster, S. G. Tetu, and H. Goodrich-Blair. Colony Location and Captivity Influence the Gut Microbial Community Composition of the Australian Sea Lion (Neophoca cinerea). *Applied and Environmental Microbiology*, 82 (12):3440–3449, 2016.

[78] A. Demirba. Calculation of higher heating values of biomass fuels. *Fuel*, 76(5):431–434, 1997.

[79] D. S. Domozych, M. Ciancia, J. U. Fangel, M. D. Mikkelsen, P. Ulvskov, and W. G. T. Willats. The Cell Walls of Green Algae: A Journey through Evolution and Diversity. *Frontiers in Plant Science*, 3, 2012.

[80] J. Drone, H.-y. Feng, C. Tellier, L. Hoffmann, V. Tran, C. Rabiller, and M. Dion. *Thermus thermophilus* Glycosynthases for the Efficient Synthesis of Galactosyl and Glucosyl β-(1$\rightarrow$3)-Glycosides. *European Journal of Organic Chemistry*, 2005(10):1977–1983, 2005.

[81] V. Ducros, C. Tarling, D. Zechel, A. M. Brzozowski, T. P. Frandsen, I. von Ossowski, M. Schlein, S. G. Withers, and G. J. Davies. Anatomy of Glycosynthesis. *Chemistry & Biology*, 10(7):619–628, 2003.

[82] T. Duo, E. D. Goddard-Borger, and S. G. Withers. Fluoro-glycosyl acridinones are ultrasensitive active site titrating agents for retaining β-glycosidases. *Chemical Communications*, 50(66):9379–9382, 2014.

[83] A. P. Dyck and R. A. MacArthur. Seasonal patterns of body temperature and activity in free-ranging beaver (*Castor canadensis*). *Canadian Journal of Zoology*, 70(9):1668–1672, 1992.

[84] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–1, 2010.

[85] L. Eichinger, J. A. Pachebat, G. Glöckner, M. A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, B. Tunggal, S. Kummerfeld, M. Madera, B. A. Konfortov, F. Rivero, A. T. Bankier, R. Lehmann, N. Hamlin, R. Davies, P. Gaudet, P. Fey, K. Pilcher, G. Chen, D. Saunders, E. Sodergren, P. Davis, A. Kerhornou, X. Nie, N. Hall, C. Anjard, L. Hemphill, N. Bason, P. Farbrother, B. Desany, E. Just, T. Morio, R. Rost, C. Churcher, J. Cooper, S. Haydock, N. van Driessche, A. Cronin, I. Goodhead, D. Muzny, T. Mourier, A. Pain, M. Lu, D. Harper, R. Lindsay, H. Hauser, K. James, M. Quiles, M. Madan Babu, T. Saito, C. Buchrieser, A. Wardroper, M. Felder, M. Thangavelu, D. Johnson, A. Knights, H. Loulseged, K. Mungall, K. Oliver, C. Price, M. A. Quail, H. Urushihara, J. Hernandez, E. Rabbinowitsch, D. Steffen, M. Sanders, J. Ma, Y. Kohara, S. Sharp, M. Simmonds, S. Spiegler, A. Tivey, S. Sugano, B. White, D. Walker, J. Woodward, T. Winckler, Y. Tanaka, G. Shaulsky, M. Schleicher, G. Weinstock, A. Rosenthal, E. C. Cox, R. L. Chisholm, R. Gibbs, W. F. Loomis, M. Platzer, R. R. Kay, J. Williams, P. H. Dear, A. A. Noegel, B. Barrell, and A. Kuspa. The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, 435(7038):43–57, 2005.

[86] A. Escalante, A. Gonalves, A. Bodin, A. Stepan, C. Sandström, G. Toriz, and P. Gatenholm.

Flexible oxygen barrier films from spruce xylan. *Carbohydrate Polymers*, 87(4):2381–2387, 2012.

[87] M. Faijes, M. Saura-Valls, X. Prez, M. Conti, and A. Planas. Acceptor-dependent regioselectivity of glycosynthase reactions by Streptomyces E383A β-glucosidase. *Carbohydrate Research*, 341(12):2055–2065, 2006.

[88] Y. Feng, C.-J. Duan, H. Pang, X.-C. Mo, C.-F. Wu, Y. Yu, Y.-L. Hu, J. Wei, J.-L. Tang, and J.-X. Feng. Cloning and identification of novel cellulase genes from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases. *Applied Microbiology and Biotechnology*, 75(2):319–328, 2007.

[89] B. Fernndez-Gmez, M. Richter, M. Schler, J. Pinhassi, S. G. Acinas, J. M. Gonzlez, and C. Pedrs-Ali. Ecology of marine Bacteroidetes: a comparative genomics approach. *The ISME Journal*, 7(5):1026–1037, 2013.

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

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

[92] R. J. Forster, A. Salgado-Flores, L. H. Hagen, S. L. Ishaq, M. Zamanzadeh, A.-D. G. Wright, P. B. Pope, and M. A. Sundset. Rumen and Cecum Microbiomes in Reindeer (Rangifer tarandus tarandus) Are Changed in Response to a Lichen Diet and May Affect Enteric Methane Emissions. *PLOS ONE*, 11(5):e0155213, 2016.

[93] S. J. Fowler, X. Dong, C. W. Sensen, J. M. Suflita, and L. M. Gieg. Methanogenic toluene metabolism: community structure and intermediates. *Environmental Microbiology*, 14(3): 754–64, 2012.

[94] K. E. H. Frandsen, T. J. Simmons, P. Dupree, J.-C. N. Poulsen, G. R. Hemsworth, L. Ciano, E. M. Johnston, M. Tovborg, K. S. Johansen, P. von Freiesleben, L. Marmuse, S. Fort, S. Cottaz, H. Driguez, B. Henrissat, N. Lenfant, F. Tuna, A. Baldansuren, G. J. Davies, L. Lo Leggio, and P. H. Walton. The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases. *Nature Chemical Biology*, 12(4):298–303, 2016.

[95] E. A. Franzosa, X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, and C. Huttenhower. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci*, 111 (22):E2329–E2338, 2014.

[96] S. C. Fry, B. H. W. A. Nesselrode, J. G. Miller, and B. R. Mewburn. Mixed-linkage $(1\rightarrow 3, 1\rightarrow 4)-\beta$-D-glucan is a major hemicellulose of *Equisetum* (horsetail) cell walls. *New Phytologist*, 179(1):104–115, 2008.

[97] X. Fu, C. Albermann, J. Jiang, J. Liao, C. Zhang, and J. S. Thorson. Antibiotic optimization via in vitro glycorandomization. *Nature Biotechnology*, 21(12):1467–1469, 2003.

[98] M. Fujita, S. Shoda, K. Haneda, T. Inazu, K. Takegawa, and K. Yamamoto. A novel disaccharide substrate having 1,2-oxazoline moiety for detection of transglycosylating activity of endoglycosidases. *Biochimica et Biophysica Acta*, 1528(1):9–14, 2001.

[99] E. M. Gabor, W. B. L. Alkema, and D. B. Janssen. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology*, 6(9): 879–886, 2004.

[100] P. Gallezot. Conversion of biomass to selected chemical products. *Chemical Society Reviews*, 41(4):1538–1558, 2012.

[101] Z. Gao, M. Niikura, and S. G. Withers. Ultrasensitive Fluorogenic Reagents for Neuraminidase Titration. *Angewandte Chemie International Edition*, 56(22):6112–6116, 2017.

[102] V. Garcia-Campayo and P. Beguin. Synergism between the cellulosome-integrating protein

CipA and endoglucanase CelD of *Clostridium thermocellum*. *Journal of Biotechnology*, 57 (1-3):39–47, 1997.

[103] J. P. Giddens, J. V. Lomino, M. N. Amin, and L. X. Wang. Endo-F3 glycosynthase mutants enable chemoenzymatic synthesis of core fucosylated tri-antennary complex-type glycopeptides and glycoproteins. *The Journal of Biological Chemistry*, 2016.

[104] L. M. Gieg, R. V. Kolhatkar, M. J. McInerney, R. S. Tanner, S. H. Harris, K. L. Sublette, and J. M. Suflita. Intrinsic Bioremediation of Petroleum Hydrocarbons in a Gas Condensate-Contaminated Aquifer. *Environmental Science & Technology*, 33(15):2550–2560, 1999.

[105] H. J. Gilbert. The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiology*, 153(2):444–455, 2010.

[106] E. D. Goddard-Borger, B. Fiege, E. M. Kwan, and S. G. Withers. Glycosynthase-mediated assembly of xylanase substrates and inhibitors. *ChemBioChem*, 12(11):1703–11, 2011.

[107] J. Greenblatt and R. Schleif. Arabinose C protein: regulation of the arabinose operon in vitro. *Nature New Biology*, 233(40):166–70, 1971.

[108] M. Greving, X. Cheng, W. Reindl, B. Bowen, K. Deng, K. Louie, M. Nyman, J. Cohen, A. Singh, B. Simmons, P. Adams, G. Siuzdak, and T. Northen. Acoustic deposition with NIMS as a high-throughput enzyme activity assay. *Analytical and Bioanalytical Chemistry*, 403(3):707–711, 2012.

[109] R. J. Gruninger, T. A. McAllister, and R. J. Forster. Bacterial and Archaeal Diversity in the Gastrointestinal Tract of the North American Beaver (*Castor canadensis*). *PLOS ONE*, 11 (5):e0156457, 2016.

[110] F. Gullfot, F. M. Ibatullin, G. Sundqvist, G. J. Davies, and H. Brumer. Functional characterization of xyloglucan glycosynthases from GH7, GH12, and GH16 scaffolds. *Biomacromolecules*, 10(7):1782–8, 2009.

[111] H. S. Hahm, M. Hurevich, and P. H. Seeberger. Automated assembly of oligosaccharides containing multiple cis-glycosidic linkages. *Nature Communications*, 7:12482, 2016.

[112] J. Handelsman. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685, 2004.

[113] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, 1998.

[114] M. Hartmann, S. Lee, S. J. Hallam, and W. W. Mohn. Bacterial, archaeal and eukaryal community structures throughout soil horizons of harvested and naturally disturbed forest stands. *Environmental Microbiology*, 11(12):3045–62, 2009.

[115] J. J. Hatton, T. J. Stevenson, C. L. Buck, and K. N. Duddleston. Diet affects arctic ground squirrel gut microbial metatranscriptome independent of community structure. *Environmental Microbiology*, 19(4):1518–1535, 2017.

[116] T. Hattori, M. Ogata, Y. Kameshima, K. Totani, M. Nikaido, T. Nakamura, H. Koshino, and T. Usui. Enzymatic synthesis of cellulose II-like substance via cellulolytic enzyme-mediated transglycosylation in an aqueous medium. *Carbohydrate Research*, 353:22–6, 2012.

[117] R. A. Heins, X. Cheng, S. Nath, K. Deng, B. P. Bowen, D. C. Chivian, S. Datta, G. D. Friedland, P. D'Haeseleer, D. Wu, M. Tran-Gyamfi, C. S. Scullin, S. Singh, W. Shi, M. G. Hamilton, M. L. Bendall, A. Sczyrba, J. Thompson, T. Feldman, J. M. Guenther, J. M. Gladden, J.-F. Cheng, P. D. Adams, E. M. Rubin, B. A. Simmons, K. L. Sale, T. R. Northen, and S. Deutsch. Phylogenomically Guided Identification of Industrially Relevant GH1 β-Glucosidases through DNA Synthesis and Nanostructure-Initiator Mass Spectrometry. *ACS Chemical Biology*, 9(9):2082–2091, 2014.

[118] W. Helbert, J. Sugiyama, M. Ishihara, and S. Yamanaka. Characterization of native crystalline cellulose in the cell walls of Oomycota. *Journal of Biotechnology*, 57(1-3):29–37, 1997.

[119] R. Hertzberger, J. Arents, H. L. Dekker, R. D. Pridmore, C. Gysler, M. Kleerebezem, M. J. T. de Mattos, and G. T. Macfarlane. H_2O_2 Production in Species of the Lactobacillus acidophilus Group: a Central Role for a Novel NADH-Dependent Flavin Reductase. *Applied and Environmental Microbiology*, 80(7):2229–2239, 2014.

[120] M. Hess, A. Sczyrba, R. Egan, T. W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, and E. M. Rubin. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–7, 2011.

[121] M. E. Himmel. *Biomass Recalcitrance: deconstructing the plant cell wall for bioenergy*. Blackwell Publishing, Oxford, 2008. ISBN 1405163607.

[122] M. E. Himmel, S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady, and T. D. Foust. Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science*, 315(5813):804–807, 2007.

[123] J. C. H. Ho, S. V. Pawar, S. J. Hallam, and V. G. Yadav. An Improved Whole-Cell Biosensor for the Discovery of Lignin-Transforming Enzymes in Functional Metagenomic Screens. *ACS Synthetic Biology*, 2017.

[124] Y. Honda and M. Kitaoka. A Family 8 Glycoside Hydrolase fromBacillus haloduransC-125 (BH2105) Is a Reducing End Xylose-releasing Exo-oligoxylanase. *Journal of Biological Chemistry*, 279(53):55097–55103, 2004.

[125] S. Horn, W. Durka, R. Wolf, A. Ermala, A. Stubbe, M. Stubbe, and M. Hofreiter. Mitochondrial genomes reveal slow rates of molecular evolution and the timing of speciation in beavers (*Castor*), one of the largest rodent species. *PLOS ONE*, 6(1):e14622, 2011.

[126] S. J. Horn, G. Vaaje-Kolstad, B. Westereng, and V. G. Eijsink. Novel enzymes for the degradation of cellulose. *Biotechnology for Biofuels*, 5(1):45, 2012.

[127] M. Hosokawa, Y. Hoshino, Y. Nishikawa, T. Hirose, D. H. Yoon, T. Mori, T. Sekiguchi, S. Shoji, and H. Takeyama. Droplet-based microfluidics for high-throughput screening of a metagenomic library for isolation of microbial enzymes. *Biosensors and Bioelectronics*, 67: 379–85, 2015.

[128] M. Hrmova, J. N. Varghese, R. De Gori, B. J. Smith, H. Driguez, and G. B. Fincher. Catalytic

Mechanisms and Reaction Intermediates along the Hydrolytic Pathway of a Plant β-D-glucan Glucohydrolase. *Structure*, 9(11):1005–1016, 2001.

[129] Y. S. Y. Hsieh and P. J. Harris. Xyloglucans of Monocotyledons Have Diverse Structures. *Molecular Plant*, 2(5):943–965, 2009.

[130] Y. S. Y. Hsieh and P. J. Harris. Structures of xyloglucans in primary cell walls of gymnosperms, monilophytes (ferns sensu lato) and lycophytes. *Phytochemistry*, 79:87–101, 2012.

[131] T. L. Hsu, S. R. Hanson, K. Kishikawa, S. K. Wang, M. Sawa, and C. H. Wong. Alkynyl sugar analogs for the labeling and visualization of glycoconjugates in cells. *Proc Natl Acad Sci*, 104(8):2614–2619, 2007.

[132] Y. Huang, G. Krauss, S. Cottaz, H. Driguez, and G. Lipps. A highly acid-stable and thermostable endo-β-glucanase from the thermoacidophilic archaeon *Sulfolobus solfataricus*. *Biochemical Journal*, 385(Pt 2):581–8, 2005.

[133] L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A new view of the tree of life. *Nature Microbiology*, 1(5), 2016.

[134] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.

[135] K. Ininbergs, B. Bergman, J. Larsson, and M. Ekman. Microbial metagenomics in the Baltic Sea: Recent advancements and prospects for environmental monitoring. *Ambio*, 44(S3):439–450, 2015.

[136] H. Iqbal, L. Low-Beinart, J. Obiajulu, and S. Brady. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. *Journal of the American Chemical Society*, 138(30):9341–9344, 2016.

[137] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, 1961.

[138] D. L. Jakeman and A. Sadeghi-Khomami. A β-(1,2)-Glycosynthase and an Attempted Selection Method for the Directed Evolution of Glycosynthases. *Biochemistry*, 50(47):10359–10366, 2011.

[139] M. C. Jarvis and D. C. Apperley. Chain conformation in concentrated pectic gels: evidence from ^{13}C NMR. *Carbohydrate Research*, 275(1):131–145, 1995.

[140] H. Jiang, B. P. English, R. B. Hazan, P. Wu, and B. Ovryn. Tracking Surface Glycans on Live Cancer Cells with Single-Molecule Sensitivity. *Angewandte Chemie International Edition*, 54 (6):1765–1769, 2015.

[141] L. Johansson, L. Virkki, S. Maunu, M. Lehto, P. Ekholm, and P. Varo. Structural characterization of water soluble β-glucan of oat bran. *Carbohydrate Polymers*, 42(2):143–148, 2000.

[142] M. H. Johansson and O. Samuelson. Reducing end groups in brich xylan and their alkaline degradation. *Wood Science and Technology*, 11(4):251–263, 1977.

[143] D. R. Jones, M. S. Uddin, R. J. Gruninger, T. T. M. Pham, D. Thomas, A. B. Boraston, J. Briggs, B. Pluvinage, T. A. McAllister, R. J. Forster, A. Tsang, L. B. Selinger, and D. W. Abbott. Discovery and characterization of family 39 glycoside hydrolases from rumen anaerobic fungi with polyspecific activity on rare arabinosyl substrates. *Journal of Biological Chemistry*, 292(30):12606–12620, 2017.

[144] D. R. Jones, D. Thomas, N. Alger, A. Ghavidel, G. D. Inglis, and D. W. Abbott. SACCHARIS: an automated pipeline to streamline discovery of carbohydrate active enzyme activities within polyspecific families and de novo sequence datasets. *Biotechnology for Biofuels*, 11(1), 2018.

[145] A. Kabisch, A. Otto, S. König, D. Becher, D. Albrecht, M. Schler, H. Teeling, R. I. Amann,

and T. Schweder. Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes *Gramella forsetii* KT0803. *The ISME Journal*, 8(7):1492–1502, 2014.

[146] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[147] P. Kanokratana, L. Eurwilaichitr, K. Pootanakit, and V. Champreda. Identification of glycosyl hydrolases from a metagenomic library of microflora in sugarcane bagasse collection site and their cooperative action on cellulose degradation. *Journal of Bioscience and Bioengineering*, 2014.

[148] A. E. Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7):497–504, 2013.

[149] O. Khersonsky and D. S. Tawfik. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1):471–505, 2010.

[150] S. M. Kielbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–93, 2011.

[151] J. H. Kim, R. Resende, T. Wennekes, H. M. Chen, N. Bance, S. Buchini, A. G. Watts, P. Pilling, V. A. Streltsov, M. Petric, R. Liggins, S. Barrett, J. L. McKimm-Breschkin, M. Niikura, and S. G. Withers. Mechanism-based covalent neuraminidase inhibitors with broad-spectrum influenza antiviral activity. *Science*, 340(6128):71–5, 2013.

[152] S.-H. Kim, C. Harzman, J. K. Davis, R. Hutcheson, J. B. Broderick, T. L. Marsh, and J. M. Tiedje. Genome sequence of *Desulfitobacterium hafniense* DCB-2, a Gram-positive anaerobe capable of dehalogenation and metal reduction. *BMC Microbiology*, 12(1):21, 2012.

[153] Y.-W. Kim, S. S. Lee, R. A. J. Warren, and S. G. Withers. Directed Evolution of a Glycosynthase from *Agrobacterium sp.* Increases Its Catalytic Activity Dramatically and Expands Its Substrate Repertoire. *Journal of Biological Chemistry*, 279(41):42787–42793, 2004.

[154] H. E. Klock and S. A. Lesley. *The Polymerase Incomplete Primer Extension (PIPE) Method Applied to High-Throughput Cloning and Site-Directed Mutagenesis*, volume 498. Humana Press, 2009.

[155] K. M. Konwar, N. W. Hanson, M. P. Bhatia, D. Kim, S. J. Wu, A. S. Hahn, C. Morgan-Lang, H. K. Cheung, and S. J. Hallam. MetaPathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics*, 31(20):3345–7, 2015.

[156] N. M. Koropatkin, E. A. Cameron, and E. C. Martens. How glycan metabolism shapes the human gut microbiota. *Nature Reviews Microbiology*, 2012.

[157] S. K. Kraun, J. Schckel, B. Westereng, L. G. Thygesen, R. N. Monrad, V. G. H. Eijsink, and W. G. T. Willats. A new generation of versatile chromogenic substrates for high-throughput analysis of biomass-degrading enzymes. *Biotechnology for Biofuels*, 8(1), 2015.

[158] S. Kuhaudomlarp, N. J. Patron, B. Henrissat, M. Rejzek, G. Saalbach, and R. A. Field. Identification of *Euglena gracilis* β-1,3-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149. *Journal of Biological Chemistry*, page jbc.RA117.000936, 2018.

[159] P. S. Kumar, M. R. Brooker, S. E. Dowd, and T. Camerlengo. Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLOS ONE*, 6(6):e20956, 2011.

[160] M. Kurogochi, M. Mori, K. Osumi, M. Tojino, S. Sugawara, S. Takashima, Y. Hirose, W. Tsukimura, M. Mizuno, J. Amano, A. Matsuda, M. Tomita, A. Takayanagi, S. Shoda, and T. Shirai. Glycoengineered Monoclonal Antibodies with Homogeneous Glycan (M3, G0, G2, and A2) Using a Chemoenzymatic Approach Have Different Affinities for Fc gamma RIIIa and Variable Antibody-Dependent Cellular Cytotoxicity Activities. *PLOS ONE*, 10(7): e0132848, 2015.

[161] S. Kuusk, B. Bissaro, P. Kuusk, Z. Forsberg, V. G. H. Eijsink, M. Srlie, and P. Vljame. Kinetics of H2O2-driven degradation of chitin by a bacterial lytic polysaccharide monooxygenase. *Journal of Biological Chemistry*, 293(2):523–531, 2018.

[162] K. K. Kwon, S.-J. Yeom, D.-H. Lee, K. J. Jeong, and S.-G. Lee. Development of a novel cellulase biosensor that detects crystalline cellulose hydrolysis using a transcriptional regulator. *Biochemical and Biophysical Research Communications*, 495(1):1328–1334, 2018.

[163] S. Lagaert, S. Van Campenhout, A. Pollet, T. M. Bourgois, J. A. Delcour, C. M. Courtin, and G. Volckaert. Recombinant Expression and Characterization of a Reducing-End Xylose-Releasing Exo-Oligoxylanase from Bifidobacterium adolescentis. *Applied and Environmental Microbiology*, 73(16):5374–5377, 2007.

[164] M. Lang, T. Kamrat, and B. Nidetzky. Influence of ionic liquid cosolvent on transgalactosylation reactions catalyzed by thermostable beta-glycosyl hydrolase CelB from *Pyrococcus Furiosus*. *Biotechnology and Bioengineering*, 95(6):1093–100, 2006.

[165] J. Larsbrink, T. E. Rogers, G. R. Hemsworth, L. S. McKee, A. S. Tauzin, O. Spadiut, S. Klinter, N. A. Pudlo, K. Urs, N. M. Koropatkin, A. L. Creagh, C. A. Haynes, A. G. Kelly, S. N. Cederholm, G. J. Davies, E. C. Martens, and H. Brumer. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature*, 506(7489): 498–502, 2014.

[166] C. L. Lauber, M. Hamady, R. Knight, and N. Fierer. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Applied and Environmental Microbiology*, 75(15):5111–5120, 2009.

[167] S. T. Laughlin and C. R. Bertozzi. Metabolic labeling of glycans with azido sugars and subsequent glycan-profiling and visualization via Staudinger ligation. *Nature Protocols*, 2 (11):2930–2944, 2007.

[168] B. D. Lauro, M. Rossi, and M. Moracci. Characterization of a β-glycosidase from the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius*. *Extremophiles*, 10(4):301–310, 2006.

[169] S. Lee and S. J. Hallam. Extraction of high molecular weight genomic DNA from soils and sediments. *Journal of Visualized Experiments*, (33), 2009.

[170] B. Leis, A. Angelov, M. Mientus, H. Li, V. T. Pham, B. Lauinger, P. Bongen, J. Pietruszka, L. G. Goncalves, H. Santos, and W. Liebl. Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. *Frontiers in Microbiology*, 6:275, 2015.

[171] A. Levasseur, E. Drula, V. Lombard, P. M. Coutinho, and B. Henrissat. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels*, 6(1):41, 2013.

[172] A. Levasseur, E. Drula, V. Lombard, P. M. Coutinho, and B. Henrissat. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels*, 6(1):41, 2013.

[173] R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon. Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–51, 2008.

[174] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[175] K.-Y. Li, J. Jiang, M. D. Witte, W. W. Kallemeijn, H. van den Elst, C.-S. Wong, S. D. Chander, S. Hoogendoorn, T. J. M. Beenakker, J. D. C. Code, J. M. F. G. Aerts, G. A. van der Marel, and H. S. Overkleeft. Synthesis of Cyclophellitol, Cyclophellitol Aziridine, and Their Tagged Derivatives. *European Journal of Organic Chemistry*, 2014(27):6030–6043, 2014.

[176] L. L. Li, S. Taghavi, S. M. McCorkle, Y. B. Zhang, M. G. Blewitt, R. Brunecky, W. S. Adney, M. E. Himmel, P. Brumm, C. Drinkwater, D. A. Mead, S. G. Tringe, and D. Lelie. Bioprospecting metagenomics of decaying wood: mining for new glycoside hydrolases. *Biotechnology for Biofuels*, 4(1):23, 2011.

[177] X. Li, P. Jackson, D. V. Rubtsov, N. Faria-Blanc, J. C. Mortimer, S. R. Turner, K. B. Krogh, K. S. Johansen, and P. Dupree. Development and application of a high throughput carbo-

hydrate profiling technique for analyzing plant cell wall polysaccharides and carbohydrate active enzymes. *Biotechnology for Biofuels*, 6(1):94, 2013.

[178] Y. Li, M. Wexler, D. J. Richardson, P. L. Bond, and A. W. B. Johnston. Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of *Rhizobium leguminosarum* and of *Escherichia coli* reveals different classes of cloned trp genes. *Environmental Microbiology*, 7(12):1927–1936, 2005.

[179] W. Liebl, A. Angelov, J. Juergensen, J. Chow, A. Loeschcke, T. Drepper, T. Classen, J. Pietruzska, A. Ehrenreich, W. R. Streit, and K.-E. Jaeger. Alternative hosts for functional (meta)genome analysis. *Applied Microbiology and Biotechnology*, 98(19):8099–8109, 2014.

[180] H. Liu and J. H. Naismith. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnology*, 8(1):91, 2008.

[181] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, 42(Database issue):D490–5, 2014.

[182] H. C. Losey, J. Jiang, J. B. Biggins, M. Oberthr, X.-Y. Ye, S. D. Dong, D. Kahne, J. S. Thorson, and C. T. Walsh. Incorporation of Glucose Analogs by GtfE and GtfD from the Vancomycin Biosynthetic Pathway to Generate Variant Glycopeptides. *Chemistry & Biology*, 9(12):1305–1314, 2002.

[183] A. S. Luis, J. Briggs, X. Zhang, B. Farnell, D. Ndeh, A. Labourel, A. Basl, A. Cartmell, N. Terrapon, K. Stott, E. C. Lowe, R. McLean, K. Shearer, J. Schckel, I. Venditto, M.-C. Ralet, B. Henrissat, E. C. Martens, S. C. Mosimann, D. W. Abbott, and H. J. Gilbert. Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nature Microbiology*, 2017.

[184] L. R. Lynd, P. J. Weimer, W. H. van Zyl, and I. S. Pretorius. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiology and Molecular Biology Reviews*, 66(3):506–77, table of contents, 2002.

[185] S. Mabeau and B. Kloareg. Isolation and Analysis of the Cell Walls of Brown Algae: *Fucus spiralis, F. ceranoides, F. vesiculosus, F. serratus, Bifurcaria bifurcata* and *Laminaria digitata*.

[186] M. S. Macauley, G. E. Whitworth, A. W. Debowski, D. Chin, and D. J. Vocadlo. O-GlcNAcase uses substrate-assisted catalysis: kinetic analysis and development of highly selective mechanism-inspired inhibitors. *Journal of Biological Chemistry*, 280(27):25313–22, 2005.

[187] A. K. Mackenzie, A. E. Naas, S. K. Kracun, J. Schckel, J. U. Fangel, J. W. Agger, W. G. T. Willats, V. G. H. Eijsink, P. B. Pope, and H. L. Drake. A Polysaccharide Utilization Locus from an Uncultured Bacteroidetes Phylotype Suggests Ecological Adaptation and Substrate Versatility. *Applied and Environmental Microbiology*, 81(1):187–195, 2015.

[188] L. F. Mackenzie, Q. Wang, R. A. J. Warren, and S. G. Withers. Glycosynthases: Mutant Glycosidases for Oligosaccharide Synthesis. *Journal of the American Chemical Society*, 120 (22):5583–5584, 1998.

[189] T. Magoč and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

[190] J. E. Maldonado, G. T. Bergmann, J. M. Craine, M. S. Robeson, and N. Fierer. Seasonal Shifts in Diet and Gut Microbiota of the American Bison (*Bison bison*). *PLOS ONE*, 10(11): e0142409, 2015.

[191] E. C. Martens, N. M. Koropatkin, T. J. Smith, and J. I. Gordon. Complex Glycan Catabolism by the Human Gut Microbiota: The Bacteroidetes Sus-like Paradigm. *Journal of Biological Chemistry*, 284(37):24673–24677, 2009.

[192] E. C. Martens, A. G. Kelly, A. S. Tauzin, and H. Brumer. The devil lies in the details: how variations in polysaccharide fine-structure impact the physiology and evolution of gut microbes. *Journal of Molecular Biology*, 426(23):3851–65, 2014.

[193] M. Martin, S. Biver, S. Steels, T. Barbeyron, M. Jam, D. Portetelle, G. Michel, and M. Vandenbol. Identification and characterization of a halotolerant, cold-active marine endo-beta-1,4-glucanase by using functional metagenomics of seaweed-associated microbiota. *Applied and Environmental Microbiology*, 80(16):4958–4967, 2014.

[194] C. F. Maurice, S. Cl Knowles, J. Ladau, K. S. Pollard, A. Fenton, A. B. Pedersen, and P. J. Turnbaugh. Marked seasonal variation in the wild mouse gut microbiota. *The ISME Journal*, 9(11):2423–2434, 2015.

[195] B. V. McCleary and R. Codd. Measurement of $(1\rightarrow 3),(1\rightarrow 4)-\beta$-D-glucan in barley and oats: A streamlined enzymic procedure. *Journal of the Science of Food and Agriculture*, 55 (2):303–312, 1991.

[196] L. S. McKee, H. Sunner, G. E. Anasontzis, G. Toriz, P. Gatenholm, V. Bulone, F. Vilaplana, and L. Olsson. A GH115 alpha-glucuronidase from Schizophyllum commune contributes to the synergistic enzymatic deconstruction of softwood glucuronoarabinoxylan. *Biotechnology for Biofuels*, 9:2, 2016.

[197] V. J. McKenzie, S. J. Song, F. Delsuc, T. L. Prest, A. M. Oliverio, T. M. Korpita, A. Alexiev, K. R. Amato, J. L. Metcalf, M. Kowalewski, N. L. Avenant, A. Link, A. Di Fiore, A. Seguin-Orlando, C. Feh, L. Orlando, J. R. Mendelson, J. Sanders, and R. Knight. The Effects of Captivity on the Mammalian Gut Microbiome. *Integrative and Comparative Biology*, 57(4): 690–704, 2017.

[198] N. D. Meadow, D. K. Fox, and S. Roseman. The Bacterial Phosphoenol-Pyruvate: Glycose Phosphotransferase System. *Annual Review of Biochemistry*, 59(1):497–542, 1990.

[199] V. Menon and M. Rao. Trends in bioconversion of lignocellulose: Biofuels, platform chemicals biorefinery concept. *Progress in Energy and Combustion Science*, 38(4):522–550, 2012.

[200] K. Mewis, M. Taupp, and S. J. Hallam. A high throughput screen for biomining cellulase activity from metagenomic libraries. *Journal of Visualized Experiments*, (48), 2011.

[201] K. Mewis, Z. Armstrong, Y. C. Song, S. A. Baldwin, S. G. Withers, and S. J. Hallam. Biomining active cellulases from a mining bioremediation system. *Journal of Biotechnology*, 167(4):462–471, 2013.

[202] K. Mewis, N. Lenfant, V. Lombard, and B. Henrissat. Dividing the Large Glycoside Hydrolase Family 43 into Subfamilies: a Motivation for Detailed Enzyme Characterization. *Applied and Environmental Microbiology*, 82(6):1686–92, 2016.

[203] D. Meyer, C. Schneider-Fresenius, R. Horlacher, R. Peist, and W. Boos. Molecular characterization of glucokinase from *Escherichia coli* K-12. *Journal of Bacteriology*, 179(4):1298–306, 1997.

[204] G. Michel, M. Czjzek, E. Rebuffet, J.-H. Hehemann, and F. Thomas. Environmental and Gut Bacteroidetes: The Food Connection. *Frontiers in Microbiology*, 2, 2011.

[205] D. Mohnen. Pectin structure and biosynthesis. *Current Opinion in Plant Biology*, 11(3): 266–277, 2008.

[206] S. Moras, Y. B. David, L. Bensoussan, S. H. Duncan, N. M. Koropatkin, E. C. Martens, H. J. Flint, and E. A. Bayer. Enzymatic profiling of cellulosomal enzymes from the human gut bacterium, R uminococcus champanellensis , reveals a fine-tuned system for cohesin-dockerin recognition. *Environmental Microbiology*, 18(2):542–556, 2016.

[207] L. R. S. Moreira and E. X. F. Filho. An overview of mannan structure and mannan-degrading enzyme systems. *Applied Microbiology and Biotechnology*, 79(2):165–178, 2008.

[208] B. D. Muegge, J. Kuczynski, D. Knights, J. C. Clemente, A. Gonzlez, L. Fontana, B. Henrissat, R. Knight, and J. I. Gordon. Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science*, 332(6032):970–974, 2011.

[209] J. Munoz-Munoz, A. Cartmell, N. Terrapon, B. Henrissat, and H. J. Gilbert. Unusual active site location and catalytic apparatus in a glycoside hydrolase family. *Proc Natl Acad Sci*, 114 (19):4936–4941, 2017.

[210] T. Nagy, K. Emami, C. M. Fontes, L. M. Ferreira, D. R. Humphry, and H. J. Gilbert. The membrane-bound alpha-glucuronidase from Pseudomonas cellulosa hydrolyzes 4-O-methyl-D-glucuronoxylooligosaccharides but not 4-O-methyl-D-glucuronoxylan. *Journal of Bacteriology*, 184(17):4925–9, 2002.

[211] M. Najah, R. Calbrix, I. Mahendra-Wijaya, T. Beneyton, A. Griffiths, and A. Drevelle. Droplet-Based Microfluidics Platform for Ultra-High-Throughput Bioprospecting of Cellulolytic Microorganisms. *Chemistry & Biology*, 21(12):1722–1732, 2014.

[212] K. Nakashima, L. Yamada, Y. Satou, J.-i. Azuma, and N. Satoh. The evolutionary origin of animal cellulose synthase. *Development Genes and Evolution*, 214(2):81–88, 2004.

[213] M. N. Namchuk and S. G. Withers. Mechanism of *Agrobacterium* beta-glucosidase: kinetic analysis of the role of noncovalent enzyme/substrate interactions. *Biochemistry*, 34(49): 16194–202, 1995.

[214] K. Naresh, F. Schumacher, H. S. Hahm, and P. H. Seeberger. Pushing the limits of automated glycan assembly: synthesis of a 50mer polymannoside. *Chemical Communications*, 53(65): 9085–9088, 2017.

[215] D. Ndeh, A. Rogowski, A. Cartmell, A. S. Luis, A. Basle, J. Gray, I. Venditto, J. Briggs, X. Zhang, A. Labourel, N. Terrapon, F. Buffetto, S. Nepogodiev, Y. Xiao, R. A. Field, Y. Zhu, M. A. O'Neill, B. R. Urbanowicz, W. S. York, G. J. Davies, D. W. Abbott, M. C. Ralet, E. C. Martens, B. Henrissat, and H. J. Gilbert. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*, 544(7648):65–70, 2017.

[216] C. E. Nelson, A. Rogowski, C. Morland, J. A. Wilhide, H. J. Gilbert, and J. G. Gardner. Systems analysis in *Cellvibrio japonicus* resolves predicted redundancy of β-glucosidases and determines essential physiological functions. *Molecular Microbiology*, 104(2):294–305, 2017.

[217] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson. Host-gut microbiota metabolic interactions. *Science*, 336(6086):1262–7, 2012.

[218] H. Nielsen. *Predicting Secretory Proteins with SignalP*, volume 1611. Humana Press, 2017.

[219] Y. Nijikken, T. Tsukada, K. Igarashi, M. Samejima, T. Wakagi, H. Shoun, and S. Fushinobu. Crystal structure of intracellular family 1 β-glucosidase BGL1A from the Basidiomycete *Phanerochaete chrysosporium*. *FEBS Letters*, 581(7):1514–1520, 2007.

[220] I. Nobeli, A. D. Favia, and J. M. Thornton. Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, 27(2):157–167, 2009.

[221] D. R. Nobles, D. K. Romanovicz, and R. M. Brown. Cellulose in Cyanobacteria. Origin of Vascular Plant Cellulose Synthase? *Plant Physiology*, 127(2):529–542, 2001.

[222] T. R. Northen, J. C. Lee, L. Hoang, J. Raymond, D. R. Hwang, S. M. Yannone, C. H. Wong, and G. Siuzdak. A nanostructure-initiator mass spectrometry-based enzyme activity assay. *Proc Natl Acad Sci*, 105(10):3678–3683, 2008.

[223] M. Nyyssonen, H. M. Tran, U. Karaoz, C. Weihe, M. Z. Hadi, J. B. Martiny, A. C. Martiny, and E. L. Brodie. Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Frontiers in Microbiology*, 4:282, 2013.

[224] P. J. O'Brien and D. Herschlag. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology*, 6(4):R91–R105, 1999.

[225] H. Ochiai, W. Huang, and L.-X. Wang. Expeditious Chemoenzymatic Synthesis of Homogeneous N-Glycoproteins Carrying Defined Oligosaccharide Ligands. *Journal of the American Chemical Society*, 130(41):13790–13803, 2008.

[226] R. M. O'Connor, J. M. Fung, K. H. Sharp, J. S. Benner, C. McClung, S. Cushing, E. R. Lamkin, A. I. Fomenkov, B. Henrissat, Y. Y. Londer, M. B. Scholz, J. Posfai, S. Malfatti, S. G. Tringe, T. Woyke, R. R. Malmstrom, D. Coleman-Derr, M. A. Altamia, S. Dedrick, S. T. Kaluziak, M. G. Haygood, and D. L. Distel. Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. *Proc Natl Acad Sci*, 111(47):E5096–E5104, 2014.

[227] T. Ohnuma, T. Fukuda, S. Dozen, Y. Honda, M. Kitaoka, and T. Fukamizo. A glycosynthase

derived from an inverting GH19 chitinase from the moss *Bryum coronatum*. *Biochemical Journal*, 444(3):437–43, 2012.

[228] M. A. O'Neill, T. Ishii, P. Albersheim, and A. G. Darvill. RHAMNOGALACTURONAN II: Structure and Function of a Borate Cross-Linked Cell Wall Pectic Polysaccharide. *Annual Review of Plant Biology*, 55(1):109–139, 2004.

[229] H. Pang, P. Zhang, C. J. Duan, X. C. Mo, J. L. Tang, and J. X. Feng. Identification of cellulase genes from the metagenomes of compost soils and functional characterization of one novel endoglucanase. *Current Microbiology*, 58(4):404–8, 2009.

[230] J. S. Papadopoulos and R. Agarwala. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, 23(9):1073–1079, 2007.

[231] H. Parker, P. Nummi, G. Hartman, and F. Rosell. Invasive North American beaver *Castor canadensis* in Eurasia: a review of potential consequences and a strategy for eradication. *Wildlife Biology*, 18(4):354–365, 2012.

[232] L. Paulova, P. Patakova, B. Branska, M. Rychtera, and K. Melzoch. Lignocellulosic ethanol: Technology design and its impact on process efficiency. *Biotechnology Advances*, 33(6):1091–1107, 2015.

[233] M. Pauly and K. Keegstra. Biosynthesis of the Plant Cell Wall Matrix Polysaccharide Xyloglucan. *Annual Review of Plant Biology*, 67(1):235–259, 2016.

[234] M. J. Pea, A. R. Kulkarni, J. Backe, M. Boyd, M. A. ONeill, and W. S. York. Structural diversity of xylans in the cell walls of monocots. *Planta*, 244(3):589–606, 2016.

[235] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11): 1420–1428, 2012.

[236] S. Pengthaisong, C.-F. Chen, S. G. Withers, B. Kuaprasert, and J. R. Ketudat Cairns. Rice BGlu1 glycosynthase and wild type transglycosylation activities distinguished by cyclophellitol inhibition. *Carbohydrate Research*, 352:51–59, 2012.

[237] G. Perugino, A. Trincone, A. Giordano, J. van der Oost, T. Kaper, M. Rossi, and M. Moracci. Activity of Hyperthermophilic Glycosynthases Is Significantly Enhanced at Acidic pH. *Biochemistry*, 42(28):8484–8493, 2003.

[238] D. R. Plichta, A. S. Juncker, M. Bertalan, E. Rettedal, L. Gautier, E. Varela, C. Manichanh, C. Fouqueray, F. Levenez, T. Nielsen, J. Dor, A. M. D. Machado, M. C. R. de Evgrafov, T. Hansen, T. Jrgensen, P. Bork, F. Guarner, O. Pedersen, M. O. A. Sommer, S. D. Ehrlich, T. Sicheritz-Pontn, S. Brunak, and H. B. Nielsen. Transcriptional interactions suggest niche segregation among microorganisms in the human gut. *Nature Microbiology*, 1(11):16152, 2016.

[239] M. L. Polizeli, A. C. Rizzatti, R. Monti, H. F. Terenzi, J. A. Jorge, and D. S. Amorim. Xylanases from fungi: properties and industrial applications. *Applied Microbiology and Biotechnology*, 67(5):577–91, 2005.

[240] P. B. Pope, S. E. Denman, M. Jones, S. G. Tringe, K. Barry, S. A. Malfatti, A. C. McHardy, J.-F. Cheng, P. Hugenholtz, C. S. McSweeney, and M. Morrison. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci*, 107(33):14793–14798, 2010.

[241] P. B. Pope, A. K. Mackenzie, I. Gregor, W. Smith, M. A. Sundset, A. C. McHardy, M. Morrison, and V. G. H. Eijsink. Metagenomics of the Svalbard Reindeer Rumen Microbiome Reveals Abundance of Polysaccharide Utilization Loci. *PLOS ONE*, 7(6):e38571, 2012.

[242] T. Pozzo, J. L. Pasten, E. N. Karlsson, and D. T. Logan. Structural and Functional Analyses of β-Glucosidase 3B from *Thermotoga neapolitana*: A Thermostable Three-Domain Representative of Glycoside Hydrolase 3. *Journal of Molecular Biology*, 397(3):724–739, 2010.

[243] T. Pozzo, M. Plaza, J. Romero-Garca, M. Faijes, E. N. Karlsson, and A. Planas. Glycosynthases from *Thermotoga neapolitana* β-glucosidase 1A: A comparison of α-glucosyl fluoride and in situ-generated α-glycosyl formate donors. *Journal of Molecular Catalysis B: Enzymatic*, 107:132–139, 2014.

[244] T. Pozzo, J. Romero-Garca, M. Faijes, A. Planas, and E. Nordberg Karlsson. Rational de-

sign of a thermostable glycoside hydrolase from family 3 introduces β-glycosynthase activity. *Glycobiology*, 27(2):165–175, 2017.

[245] H. Prade, L. F. MacKenzie, and S. G. Withers. Enzymatic synthesis of disaccharides using *Agrobacterium sp.* beta-glucosidase. *Carbohydrate Research*, 305(3-4):371–381, 1997.

[246] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137–40, 2001.

[247] M. Qi, H. S. Jun, and C. W. Forsberg. Characterization and Synergistic Interactions of Fibrobacter succinogenes Glycoside Hydrolases. *Applied and Environmental Microbiology*, 73 (19):6098–6105, 2007.

[248] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2012.

[249] J. Ravachol, R. Borne, C. Tardif, P. de Philip, and H.-P. Fierobe. Characterization of All Family-9 Glycoside Hydrolases Synthesized by the Cellulosome-producing Bacterium *Clostridium cellulolyticum*. *Journal of Biological Chemistry*, 289(11):7335–7348, 2014.

[250] J. Ravel, M. Martinez-Garcia, D. M. Brazel, B. K. Swan, C. Arnosti, P. S. G. Chain, K. G. Reitenga, G. Xie, N. J. Poulton, M. L. Gomez, D. E. D. Masland, B. Thompson, W. K. Bellows, K. Ziervogel, C.-C. Lo, S. Ahmed, C. D. Gleasner, C. J. Detter, and R. Stepanauskas. Capturing Single Cell Genomes of Active Polysaccharide Degraders: An Unexpected Contribution of Verrucomicrobia. *PLOS ONE*, 7(4):e35314, 2012.

[251] L. J. Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[252] C. Rhn. Chemical composition and gross calorific value of the above-ground biomass components of young *Picea abies*. *Scandinavian Journal of Forest Research*, 19(1):72–81, 2004.

[253] J. R. Rich and S. G. Withers. A chemoenzymatic total synthesis of the neurogenic starfish

ganglioside LLG-3 using an engineered and evolved synthase. *Angewandte Chemie International Edition*, 51(34):8640–3, 2012.

[254] P. M. Richardson, W. Shi, S. Xie, X. Chen, S. Sun, X. Zhou, L. Liu, P. Gao, N. C. Kyrpides, E.-G. No, and J. S. Yuan. Comparative Genomic Analysis of the Endosymbionts of Herbivorous Insects Reveals Eco-Environmental Adaptations: Biotechnology Applications. *PLOS Genetics*, 9(1):e1003131, 2013.

[255] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman. Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, 38(1):525–552, 2004.

[256] C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, 2013.

[257] A. Rogowski, J. A. Briggs, J. C. Mortimer, T. Tryfona, N. Terrapon, E. C. Lowe, A. Basle, C. Morland, A. M. Day, H. Zheng, T. E. Rogers, P. Thompson, A. R. Hawkins, M. P. Yadav, B. Henrissat, E. C. Martens, P. Dupree, H. J. Gilbert, and D. N. Bolam. Glycan complexity dictates microbial resource allocation in the large intestine. *Nature Communications*, 6:7481, 2015.

[258] M. Rother and J. A. Krzycki. Selenocysteine, Pyrrolysine, and the Unique Energy Metabolism of Methanogenic Archaea. *Archaea*, 2010:1–14, 2010.

[259] E. M. Rubin. Genomics of Cellulosic Biofuels. *Nature*, 454(7206):841–845, 2008.

[260] T. L. Ruegg, E.-M. Kim, B. A. Simmons, J. D. Keasling, S. W. Singer, T. Soon Lee, and M. P. Thelen. An auto-inducible mechanism for ionic liquid resistance in microbial biofuel production. *Nature Communications*, 5, 2014.

[261] J. G. Sanders, A. C. Beichman, J. Roman, J. J. Scott, D. Emerson, J. J. McCarthy, and P. R.

Girguis. Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. *Nature Communications*, 6:8285, 2015.

[262] A. Sazci, K. Erenler, and A. Radford. Detection of cellulolytic fungi by using Congo red as an indicator: a comparative study with the dinitrosalicyclic acid reagent method. *Journal of Applied Bacteriology*, 61(6):559–562, 1986.

[263] M. Schallmey, A. Ly, C. Wang, G. Meglei, S. Voget, W. R. Streit, B. T. Driscoll, and T. C. Charles. Harvesting of novel polyhydroxyalkanaote (PHA) synthase encoding genes from a soil metagenome library using phenotypic screening. *FEMS Microbiology Letters*, 321(2): 150–156, 2011.

[264] H. V. Scheller and P. Ulvskov. Hemicelluloses. *Annual Review of Plant Biology*, 61:263–89, 2010.

[265] R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

[266] C. Schröder, S. Blank, and G. Antranikian. First Glycoside Hydrolase Family 2 Enzymes from *Thermus antranikianii* and *Thermus brockianus* with β-Glucosidase Activity. *Frontiers in Bioengineering and Biotechnology*, 3, 2015.

[267] A. Schuster and M. Schmoll. Biology and biotechnology of Trichoderma. *Applied Microbiology and Biotechnology*, 87(3):787–99, 2010.

[268] E. D. Scully, S. M. Geib, K. Hoover, M. Tien, S. G. Tringe, K. W. Barry, T. Glavina del Rio, M. Chovatia, J. R. Herr, and J. E. Carlson. Metagenomic profiling reveals lignocellulose degrading system in a microbial community associated with a wood-feeding beetle. *PLOS ONE*, 8(9):e73827, 2013.

[269] H. F. Seidle, K. McKenzie, I. Marten, O. Shoseyov, and R. E. Huber. Trp-262 is a key residue for the hydrolytic and transglucosidic reactivity of the *Aspergillus niger* family 3 β-glucosidase: Substitution results in enzymes with mainly transglucosidic activity. *Archives of Biochemistry and Biophysics*, 444(1):66–75, 2005.

[270] R. Sender, S. Fuchs, and R. Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8):e1002533, 2016.

[271] Q. She, R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Y. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. K. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. Van der Oost. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proc Natl Acad Sci*, 98 (14):7835–7840, 2001.

[272] J. H. Shim, H. M. Chen, J. R. Rich, E. D. Goddard-Borger, and S. G. Withers. Directed evolution of a -glycosidase from *Agrobacterium sp.* to enhance its glycosynthase activity toward C3-modified donor sugars. *Protein Engineering Design and Selection*, 25(9):465–472, 2012.

[273] T. Shirai, H. Ishida, J. Noda, T. Yamane, K. Ozaki, Y. Hakamada, and S. Ito. Crystal structure of alkaline cellulase K: insight into the alkaline adaptation of an industrial enzyme. *Journal of Molecular Biology*, 310(5):1079–87, 2001.

[274] I. Silman, M. Nakajima, R. Yoshida, A. Miyanaga, K. Abe, Y. Takahashi, N. Sugimoto, H. Toyoizumi, H. Nakai, M. Kitaoka, and H. Taguchi. Functional and Structural Analysis of a β-Glucosidase Involved in β-1,2-Glucan Metabolism in Listeria innocua. *PLOS ONE*, 11 (2):e0148870, 2016.

[275] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, 2009.

[276] H. Smidt, S. J. Noel, G. T. Attwood, J. Rakonjac, C. D. Moon, G. C. Waghorn, and P. H. Janssen. Seasonal changes in the digesta-adherent rumen bacterial communities of dairy cattle grazing pasture. *PLOS ONE*, 12(3):e0173819, 2017.

[277] S. A. Smits, J. Leach, E. D. Sonnenburg, C. G. Gonzalez, J. S. Lichtman, G. Reid, R. Knight, A. Manjurano, J. Changalucha, J. E. Elias, M. G. Dominguez-Bello, and J. L. Sonnenburg.

Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science*, 357(6353):802–806, 2017.

[278] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *Multivariate Statistical Methods, Among-Groups Covariation*, page 269, 1975.

[279] C. Somerville, S. Bauer, G. Brininstool, M. Facette, T. Hamann, J. Milne, E. Osborne, A. Paredez, S. Persson, T. Raab, S. Vorwerk, and H. Youngs. Toward a systems approach to understanding plant cell walls. *Science*, 306(5705):2206–2211, 2004.

[280] S. J. Song, C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, D. Knights, J. C. Clemente, S. Nakielny, J. I. Gordon, N. Fierer, and R. Knight. Cohabiting family members share microbiota with one another and with their dogs. *eLife*, 2, 2013.

[281] J. L. Sonnenburg and F. Bckhed. Dietmicrobiota interactions as moderators of human metabolism. *Nature*, 535(7610):56–64, 2016.

[282] N. A. Spiridonov and D. B. Wilson. Cloning and biochemical characterization of BglC, a beta-glucosidase from the cellulolytic Actinomycete *Thermobifida fusca*. *Current Microbiology*, 42 (4):295–301, 2001.

[283] G. Srinivasan. Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science*, 296(5572):1459–1462, 2002.

[284] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–3, 2014.

[285] F. W. Studier. Protein production by auto-induction in high density shaking cultures. *Protein Expression and Purification*, 41(1):207–34, 2005.

[286] K. S. Swanson, S. E. Dowd, J. S. Suchodolski, I. S. Middelbos, B. M. Vester, K. A. Barry, K. E. Nelson, M. Torralba, B. Henrissat, P. M. Coutinho, I. K. O. Cann, B. A. White, and G. C. Fahey. Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *The ISME Journal*, 5(4):639–649, 2010.

[287] K. Tamura, G. R. Hemsworth, G. Djean, T. E. Rogers, N. A. Pudlo, K. Urs, N. Jain, G. J. Davies, E. C. Martens, and H. Brumer. Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. *Cell Reports*, 21(2):417–430, 2017.

[288] B. Tan, S. J. Fowler, N. Abu Laban, X. Dong, C. W. Sensen, J. Foght, and L. M. Gieg. Comparative analysis of metagenomes from three methanogenic hydrocarbon-degrading enrichment cultures with 41 environmental samples. *The ISME Journal*, 9(9):2028–45, 2015.

[289] E. Tancula, M. J. Feldhaus, L. A. Bedzyk, and A. A. Salyers. Location and characterization of genes involved in binding of starch to the surface of *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*, 174(17):5609–16, 1992.

[290] X. Tang, G. Xie, K. Shao, J. Dai, Y. Chen, Q. Xu, and G. Gao. Bacterial Community Composition in Oligosaline Lake Bosten: Low Overlap of ¡Betaproteobacteria and Bacteroidetes with Freshwater Ecosystems. *Microbes and Environments*, 30(2):180–188, 2015.

[291] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1):22–8, 2001.

[292] M. Taupp, S. Lee, A. Hawley, J. Yang, and S. J. Hallam. Large insert environmental genomic library production. *Journal of Visualized Experiments*, (31), 2009.

[293] M. Taupp, K. Mewis, and S. J. Hallam. The art and design of functional metagenomic screens. *Current Opinion in Biotechnology*, 22(3):465–72, 2011.

[294] M. J. Temple, F. Cuskin, A. Basl, N. Hickey, G. Speciale, S. J. Williams, H. J. Gilbert, and E. C. Lowe. A Bacteroidetes locus dedicated to fungal 1,6-β-glucan degradation: Unique substrate conformation drives specificity of the key endo-1,6-β-glucanase. *Journal of Biological Chemistry*, 292(25):10639–10650, 2017.

[295] N. Terrapon, V. Lombard, H. J. Gilbert, and B. Henrissat. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, 2014.

[296] J. Thompson, F. W. Lichtenthaler, S. Peters, and A. Pikis. β-Glucoside Kinase (BglK) from *Klebsiella pneumoniae*. *Journal of Biological Chemistry*, 277(37):34310–34321, 2002.

[297] R. Toffanin, S. H. Knutsen, C. Bertocchi, R. Rizzo, and E. Murano. Detection of cellulose in the cell wall of some red algae by ^{13}C NMR spectroscopy. *Carbohydrate Research*, 262(1): 167–171, 1994.

[298] H. Togashi, A. Kato, and K. Shimizu. Enzymatically derived aldouronic acids from *Eucalyptus globulus* glucuronoxylan. *Carbohydrate Polymers*, 78(2):247–252, 2009.

[299] A. Trincone, G. Perugino, M. Rossi, and M. Moracci. A novel thermophilic Glycosynthase that effects branching glycosylation. *Bioorganic & Medicinal Chemistry Letters*, 10(4):365–368, 2000.

[300] S. G. Tringe and E. M. Rubin. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11):805–14, 2005.

[301] T. Tsukada, K. Igarashi, M. Yoshida, and M. Samejima. Molecular cloning and characterization of two intracellular beta-glucosidases belonging to glycoside hydrolase family 1 from the basidiomycete *Phanerochaete chrysosporium*. *Applied Microbiology and Biotechnology*, 73(4): 807–14, 2006.

[302] T. Uchiyama and K. Miyazaki. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current Opinion in Biotechnology*, 20(6):616–622, 2009.

[303] M. Umekawa, W. Huang, B. Li, K. Fujita, H. Ashida, L. X. Wang, and K. Yamamoto. Mutants of *Mucor hiemalis* endo-beta-N-acetylglucosaminidase show enhanced transglycosylation and glycosynthase-like activities. *Journal of Biological Chemistry*, 283(8):4469–79, 2008.

[304] R. Vanholme, B. Demedts, K. Morreel, J. Ralph, and W. Boerjan. Lignin biosynthesis and structure. *Plant Physiology*, 153(3):895–905, 2010.

[305] M. Vega-Sanchez, Y. Verhertbruggen, H. V. Scheller, and P. Ronald. Abundance of mixed linkage glucan in mature tissues and secondary cell walls of grasses. *Plant Signaling & Behavior*, 8(2):e23143, 2014.

[306] J. K. Vester, M. A. Glaring, and P. Stougaard. Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. *Microbial Cell Factories*, 13:72, 2014.

[307] A. H. Viborg, T. Katayama, T. Arakawa, M. Abou Hachem, L. Lo Leggio, M. Kitaoka, B. Svensson, and S. Fushinobu. Discovery of α-L-arabinopyranosidases from human gut microbiome expands the diversity within glycoside hydrolase family 42. *Journal of Biological Chemistry*, 292(51):21092–21101, 2017.

[308] C. Vispo and I. D. Hume. The digestive tract and digestive function in the North American porcupine and beaver. *Canadian Journal of Zoology*, 73(5):967–974, 1995.

[309] J. Vogel. Unique aspects of the grass cell wall. *Current Opinion in Plant Biology*, 11(3): 301–307, 2008.

[310] H. Wang, R. Wang, K. Cai, H. He, Y. Liu, J. Yen, Z. Wang, M. Xu, Y. Sun, X. Zhou, Q. Yin, L. Tang, I. T. Dobrucki, L. W. Dobrucki, E. J. Chaney, S. A. Boppart, T. M. Fan, S. Lezmi, X. Chen, L. Yin, and J. Cheng. Selective in vivo metabolic cell-labeling-mediated cancer targeting. *Nature Chemical Biology*, 13(4):415–424, 2017.

[311] K. Wang, G. V. Pereira, J. J. V. Cavalcante, M. Zhang, R. Mackie, and I. Cann. *Bacteroides intestinalis* DSM 17393, a member of the human colonic microbiome, upregulates multiple endoxylanases during growth on xylan. *Scientific Reports*, 6(1), 2016.

[312] Q. Wang and S. G. Withers. Substrate-assisted catalysis in glycosidases. *Journal of the American Chemical Society*, 117(40):10137–10138, 1995.

[313] Q. Wang, R. W. Graham, D. Trimbur, R. A. J. Warren, and S. G. Withers. Changing Enzymic Reaction Mechanisms by Mutagenesis: Conversion of a Retaining Glucosidase to an Inverting Enzyme. *Journal of the American Chemical Society*, 116(25):11594–11595, 1994.

[314] B. B. Ward. How many species of prokaryotes are there? *Proc Natl Acad Sci*, 99(16): 10234–10236, 2002.

[315] F. Warnecke, P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450 (7169):560–565, 2007.

[316] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–9, 2001.

[317] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: the unseen majority. *Proc Natl Acad Sci*, 95(12):6578–83, 1998.

[318] A. Wierzbicka-Wos, P. Bartasun, H. Cieslinski, and J. Kur. Cloning and characterization of a novel cold-active glycoside hydrolase family 1 enzyme with beta-glucosidase, beta-fucosidase and beta-galactosidase activities. *BMC Biotechnology*, 13:22, 2013.

[319] S. Willför, A. Sundberg, A. Pranovich, and B. Holmbom. Polysaccharides in some industrially important hardwood species. *Wood Science and Technology*, 39(8):601–617, 2005.

[320] S. G. Withers, I. P. Street, P. Bird, and D. H. Dolphin. 2-Deoxy-2-fluoroglucosides: a novel class of mechanism-based glucosidase inhibitors. *Journal of the American Chemical Society*, 109(24):7530–7531, 1987.

[321] C. R. Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–71, 1987.

[322] D. W. Wong. Structure and action mechanism of ligninolytic enzymes. *Applied Biochemistry and Biotechnology*, 157(2):174–209, 2009.

[323] M. T. Wong, W. Wang, M. Lacourt, M. Couturier, E. A. Edwards, and E. R. Master. Substrate-Driven Convergence of the Microbial Community in Lignocellulose-Amended Enrichments of Gut Microflora from the Canadian Beaver (*Castor canadensis*) and North American Moose (*Alces americanus*). *Frontiers in Microbiology*, 7, 2016.

[324] M. T. Wong, W. Wang, M. Couturier, F. M. Razeq, V. Lombard, P. Lapebie, E. A. Edwards, N. Terrapon, B. Henrissat, and E. R. Master. Comparative Metagenomics of Cellulose- and Poplar Hydrolysate-Degrading Microcosms from Gut Microflora of the Canadian Beaver (Castor canadensis) and North American Moose (Alces americanus) after Long-Term Enrichment. *Frontiers in Microbiology*, 8, 2017.

[325] J. J. Wright, S. Lee, E. Zaikova, D. A. Walsh, and S. J. Hallam. DNA Extraction from 0.22 mu;M Sterivex Filters and Cesium Chloride Density Gradient Centrifugation. *Journal of Visualized Experiments*, (31), 2009.

[326] J. J. Wright, K. Mewis, N. W. Hanson, K. M. Konwar, K. R. Maas, and S. J. Hallam. Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME Journal*, 8(2):455–468, 2013.

[327] Y. Q. Xu, C. J. Duan, Q. N. Zhou, J. L. Tang, and J. X. Feng. Cloning and identification of cellulase genes from uncultured microorganisms in pulp sediments from paper mill effluent. *Wei Sheng Wu Xue Bao*, 46(5):783–8, 2006.

[328] C. Yang, Y. Niu, C. Li, D. Zhu, W. Wang, X. Liu, B. Cheng, C. Ma, and P. Xu. Characterization of a novel metagenome-derived 6-phospho-beta-glucosidase from black liquor sediment. *Applied and Environmental Microbiology*, 79(7):2121–2127, 2013.

[329] G.-Y. Yang, C. Li, M. Fischer, C. W. Cairo, Y. Feng, and S. G. Withers. A FRET Probe for Cell-Based Imaging of Ganglioside-Processing Enzyme Activity and High-Throughput Screening. *Angewandte Chemie*, 127(18):5479–5483, 2015.

[330] M. Yang, S. M. Luoh, A. Goddard, D. Reilly, W. Henzel, and S. Bass. The bgIX gene located at 47.8 min on the *Escherichia coli* chromosome encodes a periplasmic β-glucosidase. *Microbiology*, 142(7):1659–1665, 1996.

[331] P. Yarza, P. Yilmaz, E. Pruesse, F. O. Glöckner, W. Ludwig, K.-H. Schleifer, W. B. Whitman, J. Euzby, R. Amann, and R. Rossell-Mra. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 12(9): 635–645, 2014.

[332] Y. F. Yeh, S. C. Chang, H. W. Kuo, C. G. Tong, S. M. Yu, and T. H. Ho. A metagenomic approach for the identification and cloning of an endoglucanase from rice straw compost. *Gene*, 519(2):360–366, 2013.

[333] V. L. Y. Yip, A. Varrot, G. J. Davies, S. S. Rajan, X. Yang, J. Thompson, W. F. Anderson, and S. G. Withers. An Unusual Mechanism of Glycoside Hydrolysis Involving Redox and Elimination Steps by a Family 4 β-Glycosidase from *Thermotoga maritima*. *Journal of the American Chemical Society*, 126(27):8354–8355, 2004.

[334] W. S. York. The composition and structure of plant primary cell walls. *The Plant Cell Wall*, pages 1–54, 2003.

[335] W. S. York, A. G. Darvill, and P. Albersheim. Inhibition of 2,4-dichlorophenoxyacetic Acid-stimulated elongation of pea stem segments by a xyloglucan oligosaccharide. *Plant Physiology*, 75(2):295–7, 1984.

[336] D. L. Zechel and S. G. Withers. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Accounts of Chemical Research*, 33(1):11–8, 2000.

[337] C. Zhang, Q. Fu, C. Albermann, L. Li, and J. S. Thorson. The In Vitro Characterization of the Erythronolide Mycarosyltransferase EryBV and Its Utility in Macrolide Diversification. *ChemBioChem*, 8(4):385–390, 2007.

[338] F. Zhang and J. Keasling. Biosensors and their applications in microbial metabolic engineering. *Trends in Microbiology*, 19(7):323–329, 2011.

[339] F. Zhang, J. M. Carothers, and J. D. Keasling. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nature Biotechnology*, 30(4): 354–359, 2012.

[340] M. Zhang, N. Liu, C. Qian, Q. Wang, Q. Wang, Y. Long, Y. Huang, Z. Zhou, and X. Yan. Phylogenetic and Functional Analysis of Gut Microbiota of a Fungus-Growing Higher Termite: Bacteroidetes from Higher Termites Are a Rich Source of β-Glucosidase Genes. *Microbial Ecology*, 68(2):416–425, 2014.

[341] X. Zhang, D. E. Green, V. L. Schultz, L. Lin, X. Han, R. Wang, A. Yaksic, S. Y. Kim, P. L. DeAngelis, and R. J. Linhardt. Synthesis of 4-Azido-N-acetylhexosamine Uridine Diphosphate Donors: Clickable Glycosaminoglycans. *The Journal of Organic Chemistry*, 82(18):9910–9915, 2017.

[342] L. Zhu, Q. Wu, J. Dai, S. Zhang, and F. Wei. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proc Natl Acad Sci*, 108(43):17714–17719, 2011.

[343] Y. Zhu and X. Chen. Expanding the Scope of Metabolic Glycan Labeling in *Arabidopsis thaliana*. *ChemBioChem*, 18(13):1286–1296, 2017.

Appendix A

Chapter 2 Supplemental Material

A.1 Supplemental Tables

Table A.1: Relative Initial Rates of Hydrolysis by Fosmids Clones. Rates are given as a percentage of the maximum rate on MU cellobioside (MU-C), MU lactoside (MU-Lac), MU β-D-mannoside (MU-Man), MU β-D-galactoside (MU-Gal), MU β-D-xyloside (MU-X), MU β-D-glucoside (MU-Glc) , MU *N*-acetyl-β-D-glucosaminide (MU-GlcNAc) or MU α-L-arabinoside (MU-Ara).

Fosmid	MU-C	MU-Lac	MU-Man	MU-Gal	MU-X	MU-Glc	MU-GlcNAc	MU-Ara
12200_16_F10	100	67	0	6	21	3	0	9
12500_09_F02	100	71	0	1	0	0	0	1
40500_12_L11	55	14	6	0	11	17	1	100
CB002_04_H07	5	3	0	35	0	100	1	0
CB003_08_B11	11	3	0	2	1	100	0	3
CB004_07_C21	2	0	100	3	3	15	18	8
CB004_10_B20	8	0	1	1	16	100	58	5
CB005_08_O01	9	3	0	1	2	3	100	6
CB006_04_L11	50	8	0	28	25	100	56	26
CB006_08_D19	2	2	0	18	5	100	2	3
CG23A_01_C20	21	0	1	0	1	100	4	3
CG23A_09_O05	27	18	0	12	0	100	0	11
CG23A_23_H23	9	12	0	16	0	14	100	7
CO002_07_L07	100	86	0	7	0	0	0	0
CO003_01_D22	100	68	0	0	1	1	0	1
CO003_10_H14	65	100	1	6	0	84	0	5
CO004_05_B17	2	2	0	53	100	26	0	2
CO004_10_P05	12	0	1	30	7	100	14	7
CO182_11_I14	16	1	0	5	12	51	0	100

Table A.1 – *Continued from previous page*

Fosmid	MU-Cel	MU-Lac	MU-Man	MU-Gal	MU-Xyl	MU-Glc	MU-GlcNAc	MU-Ara
CO182_24_J12	1	1	0	100	1	6	0	4
CO182_36_O01	12	0	0	7	2	100	6	3
CO182_36_O04	100	61	7	80	12	46	0	0
CO183_09_B08	100	56	0	4	0	1	0	0
CO183_11_O01	83	85	0	59	0	100	0	19
FOS62_08_C22	2	1	0	2	100	3	6	0
FOS62_08_D12	100	93	1	12	64	7	0	6
FOS62_08_G04	14	18	0	4	0	100	1	9
FOS62_08_J18	24	38	0	2	0	100	10	24
FOS62_10_O15	0	0	0	0	0	11	0	100
FOS62_10_P15	10	7	0	0	2	100	2	97
FOS62_21_B24	3	2	0	100	1	6	1	6
FOS62_21_D16	5	3	0	28	0	2	16	100
FOS62_21_J05	11	0	0	6	2	100	7	3
FOS62_22_C08	18	0	0	71	36	53	0	100
FOS62_23_B24	19	0	0	46	9	100	9	5
FOS62_23_F03	100	75	0	0	0	37	0	22
FOS62_23_J07	100	88	0	0	0	0	0	0
FOS62_24_J23	100	89	0	0	0	0	0	0
FOS62_24_L18	50	31	0	4	22	100	19	1
FOS62_24_P09	22	4	0	1	16	100	1	0
FOS62_25_H06	38	9	0	10	10	100	7	25
FOS62_25_L08	18	0	0	33	10	100	0	1
FOS62_25_O06	2	2	1	100	0	3	9	28
FOS62_26_C23	15	5	2	100	1	33	1	17
FOS62_26_C24	14	1	0	14	62	74	100	31
FOS62_26_K06	100	55	1	2	0	5	1	0
FOS62_26_K16	100	46	2	1	1	45	0	0
FOS62_26_L14	8	0	0	10	65	100	0	0
FOS62_26_M02	100	73	0	0	0	86	7	53
FOS62_27_M17	44	0	0	38	21	100	61	47
FOS62_27_N22	12	13	0	32	0	100	8	9
FOS62_27_P24	100	95	0	0	0	0	0	0

Table A.1 – *Continued from previous page*

Fosmid	MU-Cel	MU-Lac	MU-Man	MU-Gal	MU-Xyl	MU-Glc	MU-GlcNAc	MU-Ara
FOS62_28_A14	100	90	0	0	0	0	0	0
FOS62_28_K23	17	1	0	2	1	100	1	88
FOS62_29_C04	5	2	0	15	0	100	1	2
FOS62_29_F15	8	0	0	0	0	17	0	100
FOS62_30_E15	69	74	0	19	0	100	68	52
FOS62_30_E20	6	0	0	13	85	100	0	0
FOS62_30_H03	30	0	0	23	21	100	47	16
FOS62_30_J11	100	73	1	13	0	1	0	2
FOS62_30_L24	100	78	2	0	2	5	0	0
FOS62_30_N01	2	0	0	1	1	7	100	1
FOS62_34_D13	100	48	6	7	3	6	0	0
FOS62_34_J06	19	0	0	65	2	100	5	2
FOS62_34_K14	4	0	0	2	1	100	7	4
FOS62_34_O23	100	79	0	7	0	68	38	30
FOS62_35_C14	11	1	0	28	1	100	1	2
FOS62_36_J17	73	29	0	0	67	100	90	0
FOS62_36_K01	100	89	0	0	0	1	0	6
FOS62_37_C18	18	10	3	75	1	3	25	100
FOS62_37_N04	21	0	0	68	3	100	17	2
FOS62_37_N12	100	43	6	2	3	5	2	0
FOS62_38_A06	3	0	0	0	0	7	0	100
FOS62_38_C16	5	5	0	20	79	100	0	3
FOS62_38_D22	54	23	0	1	46	100	46	0
FOS62_38_G18	100	47	14	0	2	11	0	0
FOS62_38_N16	11	0	0	26	0	100	3	1
FOS62_40_E07	100	82	0	0	0	1	0	0
FOS62_40_G22	0	0	0	0	0	3	0	100
FOS62_41_A23	1	0	0	1	0	10	0	100
FOS62_41_C11	0	0	0	0	0	100	0	0
FOS62_41_D24	13	2	0	28	34	100	0	1
FOS62_41_I01	45	100	0	1	0	1	0	0
FOS62_41_K10	100	99	0	0	0	0	0	0
FOS62_41_K19	4	0	0	12	42	100	0	0

Table A.1 – *Continued from previous page*

Fosmid	MU-Cel	MU-Lac	MU-Man	MU-Gal	MU-Xyl	MU-Glc	MU-GlcNAc	MU-Ara
FOS62_41_L01	66	100	3	81	11	22	1	0
FOS62_41_N11	11	2	1	1	100	1	3	0
FOS62_42_D11	100	65	0	0	0	0	0	0
FOS62_42_K13	95	58	0	1	61	100	63	0
FOS62_43_C07	100	63	6	11	1	10	0	0
FOS62_43_F03	32	0	0	3	100	19	3	0
FOS62_43_J20	100	39	2	2	2	2	7	2
FOS62_43_J23	100	92	0	3	0	0	0	0
FOS62_43_O18	100	83	0	0	0	0	0	0
FOS62_44_A15	100	87	0	5	0	0	0	11
FOS62_44_E09	100	81	0	0	0	0	0	0
FOS62_44_F23	13	0	0	30	5	100	0	1
FOS62_44_J10	51	22	0	100	0	87	4	15
FOS62_45_J16	100	77	0	1	0	41	18	25
FOS62_46_D05	0	2	0	100	0	3	0	5
FOS62_46_E02	8	2	1	6	100	1	3	0
FOS62_46_L17	100	32	10	2	7	11	24	37
FOS62_47_B05	100	76	0	2	3	9	0	0
FOS62_47_F04	100	50	2	1	1	2	0	0
FOS62_47_H05	2	1	0	100	6	87	0	1
FOS62_47_J09	17	0	0	54	1	100	2	2
FOS62_47_P19	11	0	0	6	6	100	63	19
NA001_01_P12	1	1	0	1	1	11	0	100
NA001_02_B17	2	3	0	26	0	100	0	1
NA001_07_E13	100	67	0	1	0	6	0	0
NA001_07_F24	0	0	0	0	0	0	0	100
NA001_11_K24	4	1	0	22	1	100	0	0
NA001_16_B03	12	7	0	10	12	39	25	100
NA002_01_B04	28	7	0	41	37	85	60	100
NA004_04_B18	100	49	19	2	4	21	0	0
NapDC_20_D21	100	10	1	1	2	6	0	0
NapDC_21_E17	7	3	0	6	2	26	20	100
NapDC_52_E10	14	2	0	91	34	100	0	1

Table A.1 – *Continued from previous page*

Fosmid	MU-Cel	MU-Lac	MU-Man	MU-Gal	MU-Xyl	MU-Glc	MU-GlcNAc	MU-Ara
NapDC_53_D04	18	11	0	17	10	100	10	7
NB001_03_I24	5	3	0	2	1	7	1	100
NB001_12_A01	100	55	0	35	13	75	0	12
NB001_13_B14	100	48	1	32	13	42	0	6
NB001_14_K20	1	1	0	10	2	100	0	2
NB001_23_D20	4	2	0	1	5	8	0	100
NO001_01_G23	82	0	0	8	53	100	0	0
NO001_01_I19	38	32	40	94	40	100	3	0
NO001_03_P09	16	0	1	3	11	100	4	0
NO001_04_B04	0	0	0	100	0	1	0	3
NO001_06_D04	100	71	0	0	0	77	0	9
NO001_07_A13	39	18	38	1	32	50	2	100
NO001_08_K19	36	12	3	2	15	100	25	0
NO001_08_N01	42	0	0	3	23	100	0	0
NO001_10_L12	77	58	0	6	3	10	100	1
NO001_13_N07	100	73	0	17	2	1	5	2
NO002_01_J07	0	1	0	54	0	100	0	0
NO002_04_P09	4	0	0	0	10	13	6	100
NO002_07_F01	37	24	0	9	24	100	51	6
NO002_11_N21	56	52	0	90	0	100	10	19
NR003_03_D21	67	16	0	17	73	100	96	0
NR003_09_O07	100	55	0	0	19	36	14	0
NR003_36_K13	100	54	0	1	0	17	0	0
PWCG7_19_I21	2	1	0	2	1	100	1	1
PWCG7_19_J20	2	2	0	7	0	100	0	1
PWCG7_33_K24	37	6	6	0	100	5	2	0
PWCG7_49_G20	100	66	0	0	0	0	1	1
SCR03_04_B15	6	0	3	1	4	100	1	0
SCR03_01_L21	6	5	0	100	0	4	0	4
TolDC_06_L02	7	0	0	0	2	47	0	100
TolDC_08_I17	2	0	0	0	2	12	0	100
TolDC_10_A11	12	0	0	1	8	57	1	100
TolDC_13_D14	74	69	0	8	0	100	3	58

Table A.1 – *Continued from previous page*

Fosmid	MU-Cel	MU-Lac	MU-Man	MU-Gal	MU-Xyl	MU-Glc	MU-GlcNAc	MU-Ara
TolDC_15_C08	21	1	0	70	14	100	0	1
TolDC_15_D05	5	0	14	12	9	100	46	0
TolDC_15_E19	9	0	0	0	3	24	0	100
TolDC_15_G15	24	2	0	59	11	100	0	1
TolDC_20_J14	31	15	0	0	100	0	1	0
TolDC_22_A01	0	1	0	1	0	11	1	100
TolDC_22_J01	23	0	0	58	15	100	0	1
TolDC_25_I24	0	0	0	5	7	100	0	0
TolDC_30_A19	49	12	6	0	4	100	0	9
TolDC_30_J10	4	0	0	1	4	31	0	100
TolDC_31_E21	24	11	1	0	100	7	1	0
TolDC_31_L02	56	60	0	0	0	100	6	7
TolDC_32_D22	2	1	0	0	0	2	0	100
TolDC_35_I03	19	0	1	26	0	3	21	100
TolDC_38_E11	18	1	0	62	12	100	0	3
TolDC_39_M03	0	0	0	0	0	43	17	100
TolDC_41_A17	53	29	0	15	20	100	34	26
TolDC_46_B16	20	1	0	2	15	81	0	100
TolDC_50_B06	16	4	0	2	13	23	0	100
TolDC_50_P08	18	9	0	4	100	10	0	2
TolDC_55_H19	73	33	0	27	100	66	2	0
TolDC_56_H11	16	0	0	0	1	72	100	2
TolDC_56_L15	9	8	0	53	0	19	23	100
TolDC_59_E21	100	64	0	46	4	96	4	3
TolDC_59_J01	9	3	0	3	4	9	0	100
TolDC_59_J06	8	2	0	1	15	56	0	100
TolDC_59_K14	100	62	0	2	0	0	0	1

Appendix B

Chapter 3 Supplemental Material

B.1 Supplemental Figures

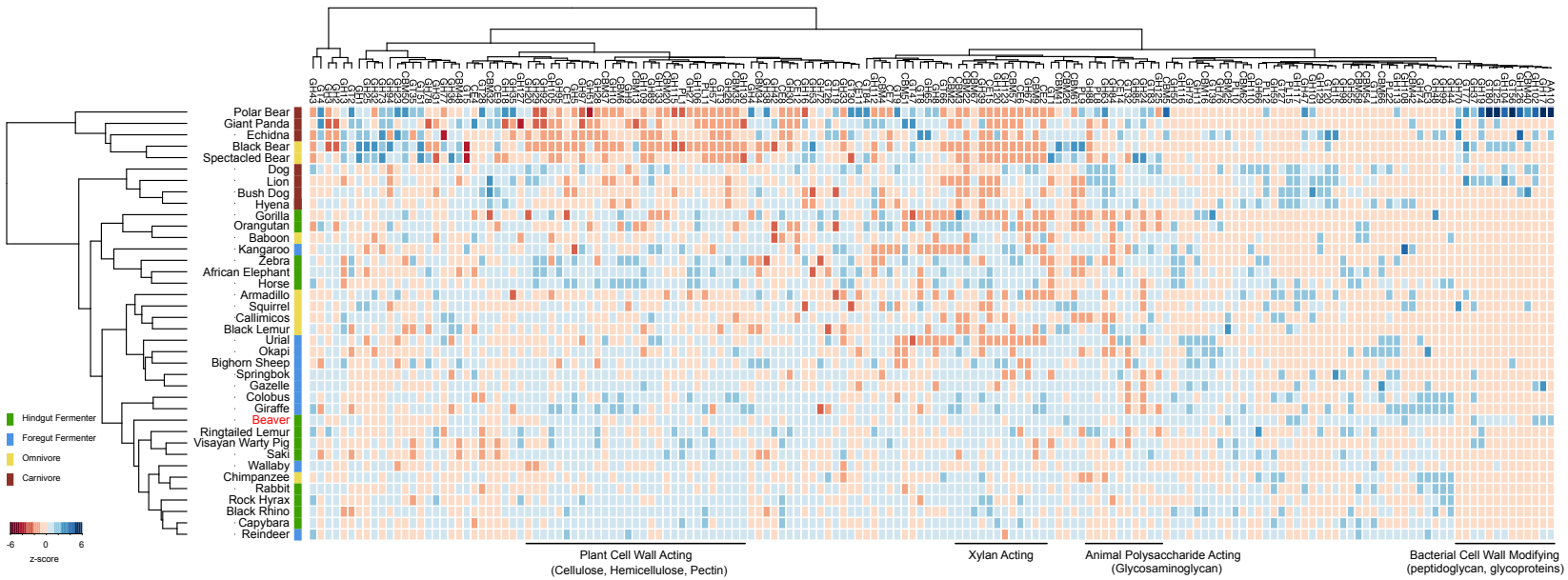

Figure B.1: *Unabridged Comparison of Beaver Fecal Metagenome with Other Sequenced Mammal Microbiomes.* Heatmap shows enrichment (blue) or depletion (red) of all families of CAZymes for each mammal. Clustering of mammals shows CAZyme abundance correlates with host digestive strategy. Clusters of genes enriched in herbivores include: 1) families active on plant polysaccharides including cellulose, hemicellulose and pectin; 2) families active on xylan. Clusters of genes enriched in carnivores include: 3) families active on animal polysaccharides such as glycosaminoglycans. Figure generated and analysed by Dr. Keith Mewis. Metagenomic data from previous studies [208, 241, 286, 342] was downloaded from the RAST online database and used for comparison. Counts of each GH family were normalized for library size using a variance stabilizing transformation provided in the DESeq2 R package [9], and the z-score for each GH family was calculated on a per sample basis. Samples and GH families were independently clustered using the Manhattan distance metric, and z-scores were plotted as a heatmap.

Appendix C

Chapter 4 Supplemental Material

C.0.1 NMR Assignments of Glycosynthase Products

Glc-β-1,3-Glc-β-pNP (pNP Laminaribiose)

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.28 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.26 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.31 (d, $J_{1,2}$ = 7.6 Hz, 1H, H-1), 4.81 (d, $J_{1',2'}$ = 8.2 Hz, 1H, H-1), 3.97 (dd, $J_{5,6a'}$= 2 Hz, $J_{6a,6b}$= 12.3, 1H, H-6a), 3.95 (dd, $J_{5',6'a}$= 2.1 Hz, $J_{6'a,6'b}$= 12.4 Hz, 1H, H-6a), 3.91 (dd, $J_{2,3'}$= 9.3 Hz, $J_{3,4}$= 8.5 Hz, 1H, H-3), 3.85 (dd, $J_{1,2}$ = 7.6 Hz, $J_{2,3'}$= 9.3 Hz, 1H, H-2), 3.80 (dd, $J_{5,6a'}$= 5.3 Hz, $J_{6a,6b}$= 12.3, 1H, H-6b), 3.74 (dd, $J_{5',6b'}$= 5.9 Hz, $J_{6'a,6'b}$= 12.3 Hz, 1H, H-6b), 3.76-3.73 (m, 1H, H-5), 3.65 (dd, $J_{3,4}$= 8.5 Hz, $J_{4,5}$= 9.7 Hz, 1H, H-4), 3.55 (dd, $J_{2',3'}$= 9.4 Hz, $J_{3',4'}$= 8.9 Hz, 1H, H-3), 3.51 (ddd, $J_{4',5'}$= 9.7 Hz, $J_{5',6'a}$= 2.2 Hz, $J_{5',6b'}$= 6.0 Hz, 1H, H-5), 3.43 (dd, $J_{3',4'}$= 8.9 Hz, $J_{4',5'}$= 9.8 Hz, 1H, H-4), 3.40 (dd, $J_{1',2'}$ = 7.9 Hz, $J_{2',3'}$= 9.4 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.78 (pNP-C-1), 142.75 (pNP-C-4), 126.21(2C, pNP-C-3 and C-5), 116.60 (2C, pNP-C-2 and C-6), 102.92(C-1), 99.33 (C-1), 83.95(C-3), 76.13(C-5), 76.01(C-5), 75.65(C-3), 73.56(C-2), 72.64(C-2), 69.70(C-4), 67.93(C-4), 60.82(C-6), 60.51(C-6).

^{13}C NMR (101 MHz, MeOD) δ 163.78, 143.95, 126.61 (2xC), 117.75(2xC), 105.24, 101.21, 87.48, 78.22, 78.04, 77.83, 75.52, 74.08, 71.57, 69.64, 62.64, 62.30.

Glc-β-1,4-Glc-β-pNP (pNP cellobioside)

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.29 (d,J= 9.2 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.2 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.32 (d, $J_{1,2}$ = 7.8 Hz, 1H, H-1), 4.56 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.02 (dd, $J_{6a,6b}$= 10.4, $J_{5,6a'}$= 3.6 Hz, 1H, H-6a), 3.95 (dd, $J_{6'a,6'b}$= 12.5 Hz, $J_{5',6'a}$= 2.1 Hz, 1H, H-6a), 3.87-3.81 (m, 2H, H-5 and H-6b), 3.82-3.74 (m, 3H, H-3, H-4, H-6b), 3.70 (dd, $J_{1,2}$ =

7.8 Hz, $J_{2,3'}$= 9.3 Hz, 1H, H-2), 3.54 (dd, $J_{2',3'}$= 9.2 Hz, $J_{3',4'}$= 9.2 Hz, 1H, H-3), 3.54-3.50 (m, 1H, H-5), 3.44 (dd, $J_{3',4'}$= 9.2 Hz, $J_{4',5'}$= 9.2 Hz, 1H, H-4), 3.35 (dd, $J_{1',2'}$ = 8.0 Hz, $J_{2',3'}$= 9.2 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.64 (pNP-C-1), 142.63 (pNP-C-4), 126.09(2C, pNP-C-3 and C-5), 116.46 (2C, pNP-C-2 and C-6), 102.56(C-1), 99.22 (C-1), 78.14 (C-4), 75.99(C-5), 75.48 (C-3), 75.08(C-5), 73.97(C-3), 73.15(C-2), 72.51(C-2), 69.44(C-4), 60.56 (C-6), 59.73(C-6).

Gal-β-1,2-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.33 8.26 (m, 2H), 7.32 7.23 (m, 2H), 5.49 (d, J = 7.5 Hz, 1H, H-1), 4.80 (d, J = 7.5 Hz, 1H, H-1), 3.98 3.89 (m, 3H, H-6, H-4, H-2), 3.84 (pt, J = 9.1 Hz, 1H, H-3), 3.80 3.68 (m, 3H, H-6, H-5, H-3), 3.66 3.61 (m, 1H, H-5), 3.60 3.53 (m, 3H, H-5, H-4, H-6a), 3.25 (dd, J = 11.2, 6.4 Hz, 1H, H-6b).

Linkage was determined by ^{1}H-^{13}C HMBC experiment (Heteronuclear Multiple Bond Correlation) showing correlation between H-1 (4.80 ppm) and C-2 (81.19 ppm), COSY (homonuclear COrrelation SpectroscopY) experiment showing correlation between H-1 (5.49 ppm) and H-2 (3.93 ppm), as well as 1H-13C HSQC experiment (Heteronuclear Single Quantum Correlation) showing correlation between H-2 (3.93 ppm) and C-2 (81.19 ppm).

Gal-β-1,3-Glc-β-pNP

1H NMR was shown to be identical to that data previously recorded by Faijes et al [87].

Glc-β-1,4-Glc-β-1,3-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide)1,2 δ 8.29 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.26 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.32 (d, $J_{1,2}$ = 7.7 Hz, 1H, H-1), 4.84 (d, $J_{1',2'}$ = 8.1 Hz, 1H, H-1), 4.53 (d, $J_{1'',2''}$ = 8.0 Hz, 1H, H-1), 4.02 (dd, $J_{5,6a'}$= 2 Hz, $J_{6a,6b}$= 12.3, 1H, H-6a), 3.95 (dd, $J_{5',6'a}$= 1.8 Hz, $J_{6'a,6'b}$= 12.5 Hz, 1H, H-6a), 3.94 (dd, $J_{5'',6''a}$= 2.1 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6a), 3.92 (dd, $J_{2,3'}$= 9.2 Hz, $J_{3,4}$= 8.5 Hz, 1H, H-3), 3.87 (dd, $J_{1,2}$ = 7.7 Hz, $J_{2,3'}$= 9.1 Hz, 1H, H-2), 3.84 (dd, $J_{5',6'b}$= 5.2 Hz, $J_{6a,6b}$= 12.5 Hz, 1H, H-6b), 3.80 (dd, $J_{5',6b'}$= 5.4 Hz, $J_{6'a,6'b}$= 12.4 Hz, 1H, H-6b), 3.78-3.72 (m, 2H, H-5 and H-6b), 3.72-3.61 (m, 4H, H-4, H-3, H-4 and H-5),

3.57-3.47 (m, 2H, H-3 and H-5), 3.44 (dd, $J_{1',2'}$ = 7.8 Hz, $J_{2',3'}$= 9.5 Hz, 1H, H-2), 3.43 (dd, $J_{3'',4''}$= 8.9 Hz, $J_{4'',5''}$= 9.8 Hz, 1H, H-4), 3.33 (dd, $J_{1'',2''}$ = 7.9 Hz, $J_{2'',3''}$= 9.4 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.78 (pNP-C-1), 142.76 (pNP-C-4), 126.21(2C, pNP-C-3 and C-5), 116.60 (2C, pNP-C-2 and C-6), 102.66 and 102.51 (C-1 and C-1), 99.35 (C-1), 83.76 (C-3), 78.69 (C-4), 76.09(C-5), 76.01(C-5), 75.59(C-5), 74.95(C-3), 74.23 (C-3), 73.35(C-2), 73.25(C-2), 72.67(C-2), 69.55(C-4), 67.88 (C-4), 60.67 (C-6), 60.50 (C-6), 60.12(C-6).

Glc-β-1,4-Glc-β-1,4-Glc-β-pNP (pNP cellotriose)

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.27 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.25 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.30 (d, $J_{1,2}$ = 7.8 Hz, 1H, H-1), 4.58 (d, $J_{1',2'}$ = 8.0 Hz, 1H, H-1), 4.53 (d, $J_{1'',2''}$ = 7.9 Hz, 1H, H-1), 4.05-3.97 (m, 2H, H-6a and H-6a), 3.93 (dd, $J_{5'',6''a}$= 2.1 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6a), 3.90-3.81 (m, 3H, H-5, H-6b and H-6b), 3.80-3.75 (m, 2H, H-3 and H-4), 3.72 (dd, $J_{5'',6''b}$= 6.2 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6b), 3.70-3.61 (m, 4H, H-2, H-3, H-4 and H-5), 3.55-3.48 (m, 2H, H-3 and H-5), 3.47 (dd, $J_{3'',4''}$= 8.8 Hz, $J_{4'',5''}$= 9.7 Hz, 1H, H-4), 3.40 (dd, $J_{1',2'}$ = 8.1 Hz, $J_{2',3'}$= 9.1 Hz, 1H, H-2), 3.33 (dd, $J_{1'',2''}$ = 8.1 Hz, $J_{2'',3''}$= 9.2 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.76 (pNP-C-1), 142.72 (pNP-C-4), 126.21(2C, pNP-C-3 and C-5), 116.58 (2C, pNP-C-2 and C-6), 102.67(C-1), 102.47(C-1), 99.37(C-1), 78.51(C-4), 78.16(C-4), 76.09(C-5), 75.58(C-3), 75.20(C-5), 74.94(C-5), 74.16(C-3), 74.05(C-3), 73.25(C-2), 73.04(C-2), 72.63(C-2), 69.56(C-4), 60.68 (C-6), 60.01(C-6), 59.82(C-6).

Glc-β-1,3-Xyl-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.29 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.32-5.24 (m, 1H, H-1), 4.81 (d, $J_{1',2'}$ = 8.7 Hz, 1H, H-1), 4.14-4.09 (m, H-5a), 3.95 (dd, $J_{5',6'a}$= 2.2 Hz, $J_{6'a,6'b}$= 12.3 Hz, 1H, H-6a), 3.90-3.82 (m, 3H, H-2, H-3, H-4), 3.74 (dd, $J_{5',6b'}$= 6.1 Hz, $J_{6'a,6'b}$= 12.3 Hz, 1H, H-6b), 3.64-3.47 (m, 3H, H-5b, H-3 and H-5), 3.43 (dd, $J_{3',4'}$= 9.1 Hz, $J_{4',5'}$= 9.9 Hz, 1H, H-4), 3.39 (dd, $J_{1',2'}$ = 8.0 Hz, $J_{2',3'}$= 9.3 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.51(pNP-C-1), 142.67 (pNP-C-4), 126.09(2C, pNP-C-3 and C-5), 116.47 (2C, pNP-C-2 and C-6), 102.68 (C-1), 99.78(C-1), 83.34(C-3), 75.99(C-5), 75.54(C-3), 73.42(C-2), 72.26(C-2), 69.62(C-4), 67.59(C-4), 64.93(C-5), 60.73(C-6).

^{13}C NMR (101 MHz, Methanol-d4) δ 163.63(pNP-C-1), 143.96(pNP-C-4), 126.61(2C, pNP-C-3 and C-5), 117.66(2C, pNP-C-2 and C-6), 105.07(C-1), 101.76(C-1), 86.87(C-3), 78.20(C-5), 77.84(C-3), 75.49(C-2), 73.76(C-2), 71.62(C-4), 69.69(C-4), 66.65(C-5), 62.67(C-6).

Glc-β-1,2-Xyl-α-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.35 8.20 (m, 2H), 7.35 7.24 (m, 2H), 6.09 (d, J = 3.5 Hz, 1H, H-1), 4.63 (d, J = 7.7 Hz, 1H, H-1), 4.05 (t, J = 9.3 Hz, 1H), 3.91 (dd, J = 3.6 Hz, 9.1 Hz, 1H, H-2), 3.91 3.88 (m, 1H), 3.84 3.73 (m, 2H), 3.70 3.63 (m, 2H), 3.61 3.53 (m, 3H), 3.45 (dd, J = 11.4, 7.3 Hz, 1H, H-6b).

Linkage was determined by 1H-13C HMBC experiment showing correlation between H-1 (4.63 ppm) and C-2 (80.3 ppm), COSY experiment showing correlation between H-1(6.09 ppm) and H-2 (3.91 ppm), as well as ^{1}H-^{13}C HSQC experiment showing correlation between H-2 (3.91 ppm) and C-2 (80.3 ppm).

Glc-β-1,2-Xyl-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.35 8.28 (m, 2H), 7.33 7.25 (m, 2H), 5.51 (d, J = 6.8 Hz, 1H, H-1), 4.79 (d, J = 7.8 Hz, 1H, H-1), 4.10 (dd, J = 11.6, 4.0 Hz, 1H), 3.99 3.91 (m, 2H), 3.87 3.79 (m, 2H), 3.72 (dd, J = 10.0, 3.4 Hz, 1H), 3.67 3.56 (m, 3H), 3.29 (dd,J= 11.2, 6.4 Hz, 1H H-6b).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.79 ppm) and C-2 (80.72 ppm), COSY experiment showing correlation between H-1(5.51 ppm) and H-2 (3.94 ppm), as well as ^{1}H-^{13}C HSQC experiment showing correlation between H-2 (3.94 ppm) and C-2 (80.72 ppm).

Glc-β-1,3-Glc-β-1,3-Xyl-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide)1,2 δ 8.29 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.30-5.26 (m, 1H, H-1), 4.86 (d, $J_{1',2'}$ = 8.0 Hz, 1H, H-1), 4.77 (d, $J_{1',2'}$ = 7.8 Hz, 1H, H-1), 4.15-4.09 (m, 1H, H-5a), 3.95 (dd, J5,6a= 1.9 Hz, $J_{6'a,6'b}$= 12.5, 1H, H-6a), 3.93 (dd, $J_{5'',6''a}$= 2.1 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6a), 3.89-3.83 (m, 3H, H-2, H-3 and

H-4), 3.80 (dd, $J_{2',3'}$= 9.1 Hz, $J_{3',4'}$= 9.1 Hz, 1H, H-3), 3.76 (dd, J5,6b= 5.6 Hz, $J_{6'a,6'b}$= 12.5, 1H, H-6b), 3.73 (dd, $J_{5'',6''b}$= 6.0 Hz, $J_{6'''a,6'''b}$= 12.2 Hz, 1H, H-6b), 3.63-3.56 (m, 1H, H-5b), 3.59 (dd, $J_{1',2'}$ = 7.7 Hz, $J_{2',3'}$= 9.3 Hz, 1H, H-2), 3.57-3.53 (m, 1H, H-4), 3.54 (dd, $J_{2'',3''}$= 9.4 Hz, $J_{3'',4''}$= 8.9 Hz, 1H, H-3), 3.57-3.47 (m, 2H, H-5 and H-5), 3.42 (dd, $J_{3'',4''}$= 8.9 Hz, $J_{4'',5''}$= 7.5 Hz, 1H, H-4), 3.38 (dd, $J_{1'',2''}$ = 7.9 Hz, $J_{2'',3''}$= 9.4 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Methanol-d4) δ 161.73(pNP-C-1), 142.91(pNP-C-4), 126.31(2C, pNP-C-3 and C-5), 116.70 (2C, pNP-C-2 and C-6), 103.01(C-1), 102.59(C-1), 100.01 (C-1), 84.43(C-3), 83.35 (C-3), 76.21(C-5), 75.82(C-3), 75.76(C-5), 73.65(C-2), 73.44(C-2), 72.53(C-2), 69.78(C-4), 68.38 (C-4), 67.77(C-4), 66.15(C-5), 60.95(C-6), 60.90(C-6).

Glc-β-1,4-Glc-β-1,3-Xyl-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide)1,2 δ 8.27 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.25 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.28-5.25 (m, 1H, H-1), 4.84 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.52 (d, $J_{1'',2''}$ = 7.9 Hz, 1H, H-1), 4.14-4.09 (m, 1H, H-5a), 4.02 (dd, $J_{5',6'a}$= 2.0 Hz, $J_{6'a,6'b}$= 12.3, 1H, H-6a), 3.93 (dd, $J_{5'',6''a}$= 2.1 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6a), 3.88-3.83 (m, 3H, H-2, H-3 and H-4), 3.83 (dd, $J_{5',6b'}$= 4.9 Hz, $J_{6'a,6'b}$= 12.4, 1H, H-6b), 3.75 (dd, $J_{5'',6''b}$= 6.9 Hz, $J_{6'''a,6'''b}$= 12.4 Hz, 1H, H-6b), 3.71-3.66 (m, 2H, H-3 and H-4), 3.66-3.55 (m, 2H, H-5b and H-5), 3.52 (dd, $J_{2'',3''}$= 9.1 Hz, $J_{3'',4''}$= 9.1 Hz, 1H, H-3), 3.53-3.47 (m, 1H, H-5), 3.44 (dd, $J_{1',2'}$ = 7.8 Hz, $J_{2',3'}$= 9.5 Hz, 1H, H-2), 3.43 (dd, $J_{3'',4''}$= 9.0 Hz, $J_{4'',5''}$= 9.7 Hz, 1H, H-4), 3.33 (dd, $J_{1'',2''}$ = 7.9 Hz, $J_{2'',3''}$= 9.3 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Methanol-d4) δ 161.50(pNP-C-1), 142.63(pNP-C-4), 126.06(2C, pNP-C-3 and C-5), 116.43 (2C, pNP-C-2 and C-6), 102.55(C-1), 102.41(C-1), 99.77 (C-1), 83.12(C-3), 78.59 (C-4), 75.94, 75.44, 74.80, 74.09, 73.19, 73.11, 72.28, 69.41 (C-4), 67.52(C-4), 64.91(C-5), 60.53(C-6), 60.00(C-6).

3-NH_2-Glc-β-1,3-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.30 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.31 (d, $J_{1,2}$ = 7.6 Hz, 1H, H-1), 4.82 (d, $J_{1',2'}$ = 8.0 Hz, 1H, H-1), 3.97 (dd, $J_{5,6a'}$= 2.2 Hz, $J_{6a,6b}$= 9.6, 1H, H-6a), 3.94 (dd, $J_{5',6'a}$= 2.3 Hz, $J_{6'a,6'b}$= 9.7 Hz,

1H, H-6a), 3.91 (dd, $J_{2,3'}$= 8.9 Hz, $J_{3,4}$= 8.9 Hz, 1H, H-3), 3.85 (dd, $J_{1,2}$ = 7.8 Hz, $J_{2,3'}$= 9.1 Hz, 1H, H-2), 3.80 (dd, $J_{5,6a'}$= 5.3 Hz, $J_{6a,6b}$= 12.3, 1H, H-6b), 3.78-3.71 (m, 2H, H-5 and H-6b), 3.65 (dd, $J_{3,4}$= 8.6 Hz, $J_{4,5}$= 9.7 Hz, 1H, H-4), 3.55 (ddd, $J_{4',5'}$= 9.8 Hz, $J_{5',6'a}$= 2.1 Hz, $J_{5',6b'}$= 6 Hz, 1H, H-5), 3.37 (dd, $J_{2',3'}$= 9.7 Hz, $J_{3',4'}$= 9.7 Hz, 1H, H-3), 3.33 (dd, $J_{1',2'}$ = 7.7 Hz, $J_{2',3'}$= 9.8 Hz, 1H, H-2), 2.90 (dd, $J_{3',4'}$= 9.8 Hz, $J_{4',5'}$= 9.8 Hz, 1H, H-4).

3-NH_2-Glc-β-1,4-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.29 (d,J= 9.2 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.2 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.32 (d, $J_{1,2}$ = 7.8 Hz, 1H, H-1), 4.57 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.05-3.99 (m, 1H, H-6a), 3.94 (dd, $J_{6'a,6'b}$= 12.0 Hz, $J_{5',6'a}$= 1.9 Hz, 1H, H-6a), 3.90-3.84 (m, 2H, H-5, H-6b), 3.82-3.77 (m, 2H, H-3 and H-4), 3.76 (dd, $J_{6'a,6'b}$= 12.7 Hz, $J_{5',6b'}$= 5.9 Hz, 1H, H-6b), 3.70 (dd, $J_{1,2}$ = 7.8 Hz, $J_{2,3'}$= 9.5 Hz, 1H, H-2), 3.55 (ddd, $J_{4',5'}$= 9.8 Hz, $J_{5',6'a}$= 2.3 Hz, $J_{5',6b'}$=5.9 Hz, 1H, H-5), 3.36 (dd, $J_{2',3'}$= 9.7 Hz, $J_{3',4'}$= 9.7 Hz, 1H, H-4), 3.27 (dd, $J_{1',2'}$ = 7.8 Hz, $J_{2',3'}$= 10 Hz, 1H, H-2), 2.85 (dd, $J_{3',4'}$= 9.8 Hz, $J_{4',5'}$= 9.8 Hz, 1H, H-3),

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.73 (pNP-C-1), 142.74 (pNP-C-4), 126.18(2C, pNP-C-3 and C-5), 116.55 (2C, pNP-C-2 and C-6), 103.11(C-1), 99.30(C-1), 78.33(C-4), 77.26(C-5), 75.19(C-5), 74.08(C-3), 72.91(C-2), 72.59(C-2), 69.31(C-4), 60.71(C-6), 59.82(C-6), 57.58(C-3).

4-NH_2-Glc-β-1,4-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.29 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.32 (d, $J_{1,2}$ = 7.8 Hz, 1H, H-1), 4.54 (d, $J_{1',2'}$ = 7.5 Hz, 1H, H-1), 4.06-3.99 (m, 1H, H-6a), 3.93 (dd, $J_{6'a,6'b}$= 12.5 Hz, $J_{5',6'a}$= 2.5 Hz, 1H, H-6a), 3.90-3.84 (m, 2H, H-5 and H-6b), 3.82-3.78 (m, 2H, H-3 and H-4), 3.75 (dd, $J_{6'a,6'b}$= 12.6 Hz, $J_{5',6b'}$= 6.0 Hz, 1H, H-6b), 3.69 (dd, $J_{1,2}$ = 7.8 Hz, $J_{2,3'}$= 9.5 Hz, 1H, H-2), 3.55 (ddd, $J_{4',5'}$= 9.9 Hz, $J_{5',6'a}$= 2.4 Hz, $J_{5',6b'}$= 5.8 Hz, 1H, H-5), 3.40 (dd, $J_{2',3'}$= 9.3 Hz, $J_{3',4'}$= 9.3 Hz, 1H, H-3), 3.35 (dd, $J_{1',2'}$ = 7.5 Hz, $J_{2',3'}$= 9.3 Hz, 1H, H-2), 2.75 (dd, $J_{3',4'}$= 9.7 Hz, $J_{4',5'}$= 9.7 Hz, 1H, H-4).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.73 (pNP-C-1), 142.73 (pNP-C-4), 126.19(2C, pNP-C-3 and C-5), 116.55 (2C, pNP-C-2 and C-6), 102.82 (C-1), 99.30 (C-1), 78.35 (C-4), 76.78(C-5), 75.70(C-3), 75.15(C-5), 74.10(C-3), 73.65(C-2), 72.59(C-2), 60.92(C-6), 59.85(C-6), 52.46 (C-4).

6-NH_2-Glc-β-1,4-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.30 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.27 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.31 (d, $J_{1,2}$ = 7.7 Hz, 1H, H-1), 4.58 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.03 (dd, $J_{6a,6b}$ = 12.2 Hz, 1H, H-6a), 3.89 (dd, $J_{5',6'b}$= 3.5 Hz, $J_{6a,6b}$= 12.4 Hz, 1H, H-6b), 3.86-3.75 (m, 4H, H-3, H-4 and H-5), 3.69 (dd, $J_{1,2}$ = 8.0 Hz, $J_{2,3'}$= 8.9 Hz, 1H, H-2), 3.59-3.5 (m, 2H, H-3 and H-5), 3.40-3.33 (m, 2H, H-2 and H-4), 3.30 (dd, $J_{5',6'a}$= 1.7 Hz, $J_{6'a,6'b}$= 13.2 Hz, 1H, H-6a), 3.01 (dd, $J_{5',6b'}$= 8.4 Hz, $J_{6'a,6'b}$= 13.4 Hz, 1H, H-6b).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 163.29(pNP-C-1), 144.29(pNP-C-4), 127.75(2C, pNP-C-3 and C-5), 118.11(2C, pNP-C-2 and C-6), 103.98(C-1), 100.95(C-1), 78.68(C-4), 76.94(C-3), 76.85(C-5), 75.70(C-5), 75.50(C-3), 74.86(C-2), 74.30(C-2), 72.64(C-4), 61.20(C-6), 42.54(C-6).

6-N_3-Glc-β-1,3-Glc-β-pNP

^{1}H NMR (600 MHz, Deuterium Oxide) δ 8.28 (d, J = 9.3 Hz, 1H), 7.26 (d, J = 9.3 Hz, 1H), 5.30 (d, J = 7.7 Hz, 1H, H-1), 4.82 (d, J = 8.0 Hz, 1H, H-1), 3.95 (dd, J = 12.4, 2.2 Hz, 1H), 3.90 (pt, J = 9.0 Hz, 1H), 3.85 (dd, J = 9.3, 7.7 Hz, 1H, H-2), 3.78 (dd, J = 12.4, 5.5 Hz, 1H), 3.77 3.72 (m, 2H), 3.66 3.59 (m, 2H), 3.56 3.51 (m, 2H), 3.46 (pt, J = 9.4 Hz, 1H, H-4), 3.40 (dd, J = 9.3, 8.0 Hz, 1H, H-2).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.82 ppm) and C-3 (83.74 ppm), H-2 (3.85 ppm) and C-1 (99.32 ppm), as well as H-2 (3.85 ppm) and C-3 (83.74 ppm).

6-N_3-Glc-β-1,4-Glc-β-pNP

^{1}H NMR (600 MHz, Deuterium Oxide) δ 8.34 8.24 (m, 1H), 7.30 7.21 (m, 2H), 5.30 (d, J = 7.9 Hz, 1H, H-1), 4.56 (d, J = 7.9 Hz, 1H, H-1), 4.04 3.98 (m, 1H, H-6a), 3.88 3.83 (m, 2H, H-6b and H-5), 3.82 3.76 (m, 3H, H-6a, H-3, H-4), 3.72 3.65 (m, 1H, H-2), 3.61 (ddd, J = 9.4, 5.6, 2.5 Hz, 1H, H-5), 3.55 (dd, J = 13.4, 5.5 Hz, 1H, H-6b), 3.51 (pt, J = 9.2 Hz, 1H, H-3), 3.47 (pt, J = 9.2 Hz, 1H, H-4), 3.36 (dd, J = 9.1, 8.0 Hz, 1H, H-2).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.56

ppm) and C-4 (78.05 ppm), ^{1}H-^{13}C HMBC correlation of H-4 (3.78 ppm) and C-2 (72.54 ppm), 1H-13C HSQC experiment showing correlation between H-2 (3.69 ppm) and C-2 (72.54 ppm), as well as HSQC experiment showing correlation between H-4 (3.78 ppm) and C-2 (78.05 ppm).

6-N_3-Gal-β-1,2-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.34 8.27 (m, 2H), 7.33 7.27 (m, 2H), 5.50 (d,J= 7.5 Hz, 1H, H-1), 4.85 (d,J= 7.9 Hz, 1H H-1), 3.98 3.90 (m, 2H), 3.88 3.87 (m, 1H), 3.84 (t,J= 9.2 Hz, 1H), 3.80 3.68 (m, 3H), 3.61 3.54 (m, 2H), 3.25 3.21 (m, 2H, H-6a and H-6b).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.85 ppm) and C-2 (81.24 ppm), COSY experiment showing correlation between H-1(5.50 ppm) and H-2 (3.92 ppm), as well as ^{1}H-^{13}C HSQC experiment showing correlation between H-2 (3.92 ppm) and C-2 (81.24 ppm).

6-N_3-Gal-β-1,3-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.32 8.24 (m, 2H), 7.31 7.22 (m, 2H), 5.31 (d,J= 7.7 Hz, 1H), 4.77 (s, 1H), 3.99 3.83 (m, 5H), 3.83 3.77 (m, 1H), 3.77 3.60 (m, 5H), 3.51 (dd,J= 13.1, 4.0 Hz, 1H, H-6b).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.77 ppm) and C-2 (83.94 ppm), H-2 (3.85 ppm) and C-1 (99.37 ppm), as well as H-2 (3.85 ppm) and C-3 (83.92 ppm).

6-N_3-Gal-β-1,4-Glc-β-pNP

^{1}H NMR (600 MHz, Deuterium Oxide) δ 8.33 8.23 (m, 2H), 7.26 (d,J= 9.1 Hz, 2H), 5.31 (d,J= 7.8 Hz, 1H, H-1), 4.52 (d,J= 7.8 Hz, 1H, H-1), 4.04 3.99 (m, 1H, H-6a), 3.93 (d,J= 3.4 Hz, 1H, H-4,), 3.88 3.84 (m, 3H), 3.82 3.79 (m, 2H), 3.72 3.67 (m, 2H), 3.63 (dd,J= 13.1, 8.5 Hz, 1H, H-6a), 3.59 3.55 (m, 2H).

Linkage was determined by ^{1}H-^{13}C HMBC experiment showing correlation between H-1 (4.52 ppm) and C-4 (78.2 ppm), H-6a (4.00 ppm) and C-4 (78.2 ppm), as well as H-6b (3.86 ppm) and C-4 (78.2 ppm).

6-N_3-Glc-β-1,4-Glc-β-1,4-Glc-β-pNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.28 (d,J= 9.3 Hz, 2H, 2x pNP -O-C-C-H), 7.26 (d,J= 9.3 Hz, 2H, 2x pNP -O_2N-C-C-H), 5.31 (d, $J_{1,2}$ = 7.8 Hz, 1H, H-1), 4.58 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.55 (d, J1,2 = 7.9 Hz, 1H, H-1), 4.03 (dd, $J_{5,6a'}$= 1 Hz, $J_{6a,6b}$= 11.8, 1H, H-6a), 4.02 (dd, $J_{5',6'a}$= 1.7 Hz, $J_{6'a,6'b}$= 12.1 Hz, 1H, H-6a), 3.91-3.84 (m, 3H, H-5, H-6b, H-6b), 3.84-3.78 (m, 3H, H-6a, H-3 and H-4), 3.72-3.65 (m, 3H, H-2, H-3, H-4 and H-5), 3.62 (ddd, $J_{4',5'}$= 8.9 Hz, $J_{5'',6''a}$= 2.3 Hz, $J_{5'',6''b}$=5.7 Hz, 1H, H-5), 3.55 (dd, $J_{5'',6''b}$= 5.7 Hz, $J_{6'''a,6'''b}$= 13.4 Hz, 1H, H-6b), 3.52 (dd, $J_{2'',3''}$= 8.9 Hz, $J_{3'',4''}$= 9.0 Hz, 1H, H-3), 3.47 (dd, $J_{3'',4''}$= 9.0 Hz, $J_{4'',5''}$= 9.0 Hz, 1H, H-4), 3.41 (dd, $J_{1',2'}$ = 7.8 Hz, $J_{2',3'}$= 9.1 Hz, 1H, H-2), 3.35 (dd, $J_{1'',2''}$ = 7.9 Hz, $J_{2'',3''}$= 9.0 Hz, 1H, H-2).

^{13}C NMR (101 MHz, Deuterium Oxide) δ 161.64 (pNP-C-1), 142.62 (pNP-C-4), 126.09 (2C, pNP-C-3 and C-5), 116.46 (2C, pNP-C-2 and C-6), 102.50 (C-1), 102.39 (C-1), 99.24 (C-1), 78.27 (C-4), 78.05 (C-4), 75.22, 75.09, 74.79, 74.20, 73.94, 73.93, 73.10, 72.92, 72.52, 70.03 (C-4), 59.85 (C-6), 59.69 (C-6), 50.85 (C-6).

4-NH_2-Glc-β-1,4-2FGlc-β-DNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.93 (d,J= 2.8 Hz, 1H, DNP-H-3), 8.58 (dd,J= 9.3, 2.8 Hz, 1H, DNP-H-5), 7.66 (d,J= 9.4 Hz, 1H, DNP-H-6), 5.78 (dd, $J_{1,2}$ = 7.6 Hz, $J_{1,F}$ = 2.9 Hz, 1H, H-1), 4.64 (ddd, $J_{1,2}$ = 7.6 Hz, $J_{2,F}$ = 51.2 Hz, $J_{2,3'}$ = 9.1Hz, 1H, H-2), 4.55 (d, $J_{1',2'}$ = 7.6 Hz, 1H, H-1), 4.11 (ddd, $J_{2,3'}$= 8.8 Hz, $J_{3,4}$= 8.8 Hz, $J_{3,F}$ = 15.7 Hz, 1H, H-3), 4.04 (dd, $J_{6a,6b}$= 12.2 Hz, $J_{5,6a'}$= 1.4 Hz, 1H, H-6a), 3.96-3.84 (m, 4H, H-4, H-5, H-6b and H-6a), 3.75 (dd, $J_{6'a,6'b}$= 12.4 Hz, $J_{5',6b'}$= 5.9 Hz, 1H, H-6b), 3.48 (ddd, $J_{4',5'}$= 9.9 Hz, $J_{5',6'a}$= 2.5 Hz, $J_{5',6b'}$= 5.8 Hz, 1H, H-5), 3.41 (dd, $J_{2',3'}$= 9.7 Hz, $J_{3',4'}$= 9.7 Hz, 1H, H-3), 3.35 (dd, $J_{1',2'}$ = 7.6 Hz, $J_{2',3'}$= 9.3 Hz, 1H, H-2), 2.76 (dd, $J_{3',4'}$= 9.7 Hz, $J_{4',5'}$= 9.7 Hz, 1H, H-4)

^{13}C NMR (101 MHz, D_2O+DMSO-d6) δ 155.35 (DNP C-1), 143.35(DNP C-2) 140.63(DNP C-4), 131.41(DNP C-5), 123.71(DNP C-6), 119.45(DNP C-3), 104.24(C-1), 99.20(d, C-1), 92.54 (d, C-2), 78.87 (d, C-4), 78.15 (C-5), 77.06, 77.01(C-3 and C-5), 75.08(C-2), 74.09(d, C-3), 62.45(C-6), 61.04(C-6), 54.03 (C-4).

Glc-β-1,4-Glc-β-1,4-2F-Glc-β-DNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.92 (d,J= 2.8 Hz, 1H, DNP-H-3), 8.56 (dd,J= 9.4, 2.8 Hz, 1H, DNP-H-5), 7.63 (d,J= 9.4 Hz, 1H, DNP-H-6), 5.78 (dd, $J_{1,2}$ = 7.6 Hz, $J_{1,F}$ = 3.0 Hz, 1H, H-1), 4.64 (ddd, $J_{1,2}$ = 7.6 Hz, $J_{2,F}$ = 51.1 Hz, $J_{2,3'}$ = 9.0Hz, 1H, H-2), 4.59 (d, $J_{1',2'}$ = 7.9 Hz, 1H, H-1), 4.53 (d, $J_{1'',2''}$ = 7.9 Hz, 1H, H-1), 4.11 (ddd, $J_{2,3'}$= 8.6 Hz, $J_{3,4}$= 8.6 Hz, $J_{3,F}$ = 15.8 Hz, 1H, H-3), 4.06-3.99 (m, 2H, H-6a and H-6a), 3.96-3.87 (m, 4H, H-4, H-5, H-6b and H-6a), 3.85 (dd, $J_{6'a,6'b}$= 12.3 Hz, $J_{5',6b'}$= 4.7 Hz, 1H, H-6b), 3.75 (dd, $J_{6'''a,6'''b}$= 12.3 Hz, $J_{5'',6''b}$= 5.7 Hz, 1H, H-6b), 3.72-3.62 (m, 3H, H-3, H-4 and H-5), 3.53 (dd, $J_{2'',3''}$= 9.1 Hz, $J_{3'',4''}$= 9.1 Hz, 1H, H-3), 3.52-3.48 (m, 1H, H-5), 3.43 (dd, $J_{3'',4''}$= 9.0 Hz, $J_{4'',5''}$= 9.7 Hz, 1H, H-4), 3.40 (dd, $J_{1',2'}$ = 7.9 Hz, $J_{2',3'}$= 8.9 Hz, 1H, H-2), 3.33 (dd, $J_{1'',2''}$ = 7.9 Hz, $J_{2'',3''}$= 9.3 Hz, 1H, H-2)

^{13}C NMR (101 MHz, D_2O+DMSO-d6) δ 154.28(DNP C-1), 142.18(DNP C-2) 139.43(DNP C-4), 130.25(DNP C-5), 122.63(DNP C-6), 118.29(DNP C-3), 102.99(C-1), 102.69(C-1), 98.07(d, C-1), 91.37 (d, C-2), 78.83 (C-4), 77.40(d, C-4), 76.40(C-5), 75.90, 75.88(C-3 and C-5), 75.28(C-5), 74.45(C-3), 73.57(C-2), 73.32(C-2), 72.88 (d, C-3), 69.87(C-4), 60.99(C-6), 60.34(C-6), 59.84 (C-6).

Glc-β-1,4-Glc-β-1,4-Glc-β-1,4-2F-Glc-β-DNP

^{1}H NMR (400 MHz, Deuterium Oxide) δ 8.93 (d,J= 2.8 Hz, 1H, DNP-H-3), 8.57 (dd,J= 9.4, 2.8 Hz, 1H, DNP-H-5), 7.64 (d,J= 9.4 Hz, 1H, DNP-H-6), 5.79 (dd, $J_{1,2}$ = 7.5 Hz, $J_{1,F}$ = 2.9 Hz, 1H, H-1), 4.64 (ddd, $J_{1,2}$ = 7.6 Hz, $J_{2,F}$ = 51.1 Hz, $J_{2,3'}$ = 9.0Hz, 1H, H-2), 4.60 (d, $J_{1',2'}$ = 8.0 Hz, 1H, H-1), 4.56 (d, $J_{1'',2''}$ = 7.9 Hz, 1H, H-1), 4.53 (d, $J_{1''',2'''}$ = 7.9 Hz, 1H, H-1), 4.12 (ddd, $J_{2,3'}$= 8.5 Hz, $J_{3,4}$= 8.5 Hz, $J_{3,F}$ = 16.2 Hz, 1H, H-3), 4.07-3.97 (m, 3H, H-6a, H-6a and H-6a), 3.97-3.88 (m, 4H, H-4, H-5, H-6b, H-6a), 3.86 (dd, $J_{6'a,6'b}$= 12.3 Hz, $J_{5',6b'}$= 4.5 Hz, 1H, H-6b), 3.85 (dd, $J_{6'''a,6'''b}$= 12.5 Hz, $J_{5'',6''b}$= 5.7 Hz, 1H, H-6b), 3.75 (dd, $J_{6'''a,6'''b}$= 12.5 Hz, $J_{5''',6'''b}$= 5.7 Hz, 1H, H-6b), 3.73-3.61 (m, 6H, H-3, H-4, H-5, H-3, H-4 and H-5), 3.56-3.48 (m, 2H, H-3 and H-5), 3.46-3.36 (m, 3H, H-2, H-2, H-4), 3.33 (dd, $J_{1''',2'''}$ = 7.9 Hz, $J_{2''',3'''}$= 9.3 Hz, 1H, H-2).

Glc-β-1,4-Glc-β-1,3-Glc-β-octyl

^{1}H NMR (400 MHz, Deuterium Oxide) δ 4.77 (d, $J_{1',2'}$ = 7.6 Hz, 1H, H-1), 4.52 (d, $J_{1'',2''}$ = 7.9 Hz, 1H, H-1), 4.49 (d, $J_{1,2}$ = 8.1 Hz, 1H, H-1), 4.00 (dd, $J_{5,6a'}$= 2.1 Hz, $J_{6a,6b}$= 12.3, 1H, H-6a), 3.96-3.89 (m, 3H, Octyl-H-1a, H-6a and H-6a), 3.82 (dd, $J_{5',6'b}$= 5.0 Hz, $J_{6a,6b}$= 12.3, 1H, H-6b),3.78-3.69 (m, 3H, H-3, H-6b and H-6b), 3.69-3.60 (m, 5H, Octyl-H-1b, H-5, H-3, H-4 and H-5), 3.55-3.49 (m, 3H, H-4, H-3 and H-5), 3.49-3.38 (m, 3H, H-2, H-2 and H-4), 3.33 (dd, $J_{1'',2''}$ = 7.8 Hz, $J_{2'',3''}$= 9.2 Hz, 1H, H-2), 1.69-1.58 (m, 2H, Octyl-H-2), 1.41-1.25 (m, 10H, Octyl H-3, H-4, H-5, H-6 and H-7), 0.91-0.84 (m, 3H, Octyl-H-8).