Targeted enrichment of cellulase genes with stable-isotope probing and metagenomics

by

Lee Pinnell

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2011

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis although several sections contain text from publications arising from my thesis research and some results were generated as part of a collaboration. The two month DNA-SIP incubations were conducted by Eric Dunford, the cellulose degrading enrichment culture was provided by Patrick Ronan and Dr. Martina Hausner at Ryerson University, the Illumina sample preparation was completed by Dr. Stefan Green at the University of Illinois at Chicago, and the generation of unrooted neighbor-joining trees in R was completed by Dr. Michael Lynch. Parts of this thesis were previously published in Stable-Isotope Probing and Related Technologies (Pinnell *et al.*, 2011), and in the Journal for Microbiological Methods (Engel *et al.*, 2011). This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

Cellulose is the most abundant organic compound on the planet, and is found in nearly every ecosystem. Cellulose is also the most abundant waste product produced by human activity. These enormous stores of natural cellulose and cellulose-containing wastes are a potential renewable energy source. The hydrolysis of cellulose is referred to as cellulolysis and is carried out by cellulase enzymes, which are members of certain glycoside hydrolase families. For most of its history, the microbiology of organisms like those that hydrolyze cellulose was based solely on the testing of physiological and biochemical behaviour of isolated organisms in pure cultures. Despite having gained an important foundation of knowledge in the characterization of microorganisms, cultivation-based techniques introduce major bias into understanding the role that specific microorganisms play because the majority of microorganisms are not readily cultured. Two of the most powerful culture-independent approaches for accessing microbial communities are DNA stable-isotope probing (DNA-SIP) and metagenomics. Though each methodology has been used on its own, it is a combination of these two approaches that has enormous potential to generate results for industrial applications and to help characterize biogeochemical cycling. This thesis presents the first research combining DNA-SIP and metagenomics using cellulose, and the first to target glycoside hydrolase genes from Arctic tundra. For this research, two-month DNA-SIP incubations were carried out with 200 mg of ^{13}C-labelled cellulose as a substrate. Denaturing gradient gel electrophoresis (DGGE) provided evidence indicating the successful enrichment of microorganisms able to metabolize cellulose. Multiple displacement amplification (MDA) was applied to both the bulk-soil samples and

DNA-SIP samples. Following MDA, all DNA samples were subjected to Illumina sequencing, including DNA from a cellulose-degrading enrichment. Functional annotation for each Illumina library was done using the SwissProt database within MG-RAST. The DNA-SIP enrichment resulted in a ~3 fold increase in the relative abundance of glycoside hydrolases and cellulase enzymes in relation to bulk soil samples. A cellulose degrading enrichment contained the highest relative abundance of glycoside hydrolases and cellulase enzymes, with a five fold increase relative to the DNA-SIP enrichment. The enrichment culture had a much lower relative diversity, which was measured using the Shannon Index. An unrooted neighbor-joining tree constructed using Bray-Curtis similarity coefficients for each sample demonstrated that as a result of a considerably higher proportion of cellulase gene sequences and a lower diversity the enrichment culture was the most distinct library, with the DNA-SIP library most closely related to it. DGGE provided initial evidence that MDA introduced bias into the amplification of DNA from the DNA-SIP sample. This was confirmed following sequencing and annotation as the proportion of glycoside hydrolase enzymes sequences decreased 67% following MDA of DNA-SIP enriched DNA and the mean G+ C content of libraries decreased. This research provides evidence indicating that DNA-SIP enrichment prior to the construction of metagenomic libraries increases the abundance of targeted gene sequences, which should enable greater access to functional genes of active microorganisms for potential industrial applications.

Acknowledgements

I would like to thank my supervisor Dr. Josh Neufeld for giving me the opportunity to perform this research, for his encouragement, advice, and dedication. I would also like to thank Dr. Trevor Charles and Dr. Barb Butler for their advice and support.

I'd also like to thank members of the Neufeld Lab for helping me along the way. I'd especially like to thank Andrea Bartram, Tara Moore, and Eric Dunford for teaching me the majority of research techniques. Without them I would still be making gels using water. Special thanks to Eric Dunford for doing preliminary research and showing me the DNA-SIP ropes. Long live the Dunford-Pinnell theory! Another special thanks to Michael Lynch for his assistance with all things computers and to Jiujun Cheng for his advice and assistance along the way, especially regarding cloning.

Finally, I'd like to thank my family and friends for their support and encouragement throughout the last two years.

Support for this study was provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Greenfield Ethanol Inc.

And now for something completely different.

Table of Contents

AUTHOR'S DECLARATION ii
Abstract iii
Acknowledgements v
Table of Contents vi
List of Figures viii
List of Tables ix
List of Abbreviations x
Chapter 1 Literature Review 1
1.1 Cellulose and Related Processes 1
1.1.1 Introduction 1
1.1.2 Glycoside hydrolases and their industrial potential 2
1.1.3 The Arctic tundra: cold and loaded with cellulose 4
1.2 Culture independent studies 6
1.2.1 A change in approach 6
1.2.2 DNA-SIP 8
1.2.3 Metagenomics 9
1.2.4 Challenges and solutions 11
1.3 Research Objectives and Hypotheses 16
Chapter 2 Methods 17
2.1 Introduction 17
2.2 DNA Extraction/Purification Optimization 17
2.2.1 FastDNA Isolation 18
2.2.2 Modified Gentle Lysis Extraction 19
2.2.3 Wizard DNA Clean-Up Column Purification 19
2.2.4 SCODA Purification 20
2.2.5 Pulsed-Field Gel Electrophoresis (PFGE) 25
2.3 DNA-SIP and Metagenomics 26
2.3.1 DNA-SIP 28
2.3.2 Multiple Displacement Amplification (MDA) 29
2.3.3 Polymerase Chain Reaction Amplification (PCR) 30
2.3.4 Denaturing Gradient Gel Electrophoresis (DGGE) 30

2.3.5 Illumina Sequencing 31
2.3.6 MG-RAST and Functional Annotation 31
Chapter 3 Results and Discussion 34
3.1 Introduction 34
3.2 Optimization of DNA Extraction and Purification 34
3.3 DNA-SIP and metagenomics 39
3.3.1 Bulk soil DNA extraction, and DNA-SIP microcosm incubation 40
3.3.2 DGGE fingerprinting of fractions 41
3.3.3 MDA and DGGE fingerprinting of all samples 43
3.3.4 Shot-gun metagenomics using Illumina and MG-RAST 46
3.3.5 Bray-Curtis dissimilarity unrooted, neighbor-joining trees 59
Chapter 4 Conclusions and Future Considerations 62
4.1 Conclusions 62
4.2 Future Considerations 65
Appendix A Total GH Tables 70
References 86

List of Figures

Figure 2.1 Schematic of the three DNA extraction and purification techniques tested…..…18

Figure 2.2 Overview of the SCODA gel boat arrangement…….…21

Figure 2.3 Schematic of sample processing for DNA-SIP and metagenomics……...……26

Figure 3.1 DNA from each of the three extraction and purification methods…..…….……..…36

Figure 3.2 1% agarose gel demonstrating the DNA content of gradient fractions……..……41

Figure 3.3 DGGE with DNA fingerprint from all DNA-SIP fractions……...…..…..…….....42

Figure 3.4 Quantification of DNA from MDA reactions……..……..……44

Figure 3.5 DGGE with the DNA fingerprint of all samples sequenced……..46

Figure 3.6 Source hits distribution for an Illumina library using MG-RAST…….48

Figure 3.7 Taxonomic distribution of domains for each Illumina library……...…53

Figure 3.8 Distribution of functional categories for each Illumina library…….….54

Figure 3.9 Unrooted neighbor-joining trees for all sequenced samples……..61

List of Tables

Table 1.1 Summary of all studies to date coupling DNA SIP and metagenomics……7

Table 2.1 List of the SCODA gel boat components……20

Table 2.2 List of SCODA components and their functions……21

Table 3.1 DNA yield and size following extraction using the three methods tested……..37

Table 3.2 Names and description of the eight samples sequenced by Illumina……....……..39

Table 3.3 Characteristics of the Arctic tundra soil used…….……....40

Table 3.4 Number of sequences, G+C content, and MG-RAST accession numbers for each library…….……47

Table 3.5 Five most abundant GH enzyme sequences for the DNA-SIP, DNA-SIP MDA, and bulk soil libraries…….……..50

Table 3.6 The number of enzymes annotated as glycoside hydrolases (GH) and cellulose degrading GH per one million sequences)…… ……..…….....51

Table 3.7 Proportion of GH enzymes comprising the ten most abundant GH families.....56

Table 3.8 Shannon index values for each library…….56

Table 3.9 Number of sequences annotated as cellulases for aech library……58

List of Abbreviations

°C	degrees Celsius
µg	microgram
µL	microliter
µm	micrometer
µmol	micromolar
APE	analyses of phylogenetics and evolution software
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
C	carbon atom
CAZy	carbohydrate-active enzymes database
cm	centimeter(s)
CMC	carboxymethyl cellulose
CsCl	cesium chloride
ddH_2O	distilled and deionized water
DGGE	denaturing gradient gel electropohresis
DNA	deoxyribonucleic acid
DNA-SIP	deoxyribonucleic acid stable-isotope probing
dNTP	deoxyribonucleic triphosphate
EDTA	ethylenediamine tetraacetate
EtBr	ethidium bromide
g	gram(s)
GC	guanine-cytosine
g/cm^3	grams per cubic centimeter
GH	glycoside hydrolase family
g/L	grams per litre
g/mL	grams per milliliter
H_2O	water
HCl	hydrogen chloride
kb	kilobase pair(s)
km^2	square kilometres
L	litre(s)
M	molar
Mbp	megabase pair(s)
MDA	multiple displacement amplification
mg/mL	milligrams per milliliter
MG-RAST	metagenomics rapid annotation using subsystem technology server
mL	milliliter
mm	millimetre
mM	millimolar
MS	Microsoft
N	nitrogen atom

ng	nanogram(s)
ng/µL	nanograms per microlitre
NaCl	sodium chloride
O	oxygen atom
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulsed-field gel electrophoresis
pH	power of hydrogen
PLFA-SIP	phospholipid fatty acid stable-isotope probing
pmol	picomoles
R	programming language for statistical computing and graphics
RNA-SIP	ribonucleic acid stable-isotope probing
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid
s	second(s)
SCODA	synchronous coefficient of drag alteration
SDS	sodium dodecyl sulfate
SwissProt	manually annotated universal protein resource database
TE	10mM Tris HCl (pH 8.0), 1 mM EDTA
Tris	tris (hydroxymethyl) amino methane
UV	ultraviolet
V	volt(s)
V/cm	volts per centimetre
w/v	weight per volume
WGA	whole genome amplification

Chapter 1 Literature Review

1.1 Cellulose and Related Processes

1.1.1 Introduction

Cellulose is the most abundant organic compound on Earth (Lynd *et al.*, 2002, O'Sullivan *et al.*, 2007) and is a polysaccharide consisting of β-1-4 linked glucose molecules. As a major component in the cell walls of plants and algae, cellulose represents 35-50% of their dry weight (Schellenberger *et al.*, 2010). As a result, cellulose is found in nearly all of the planet's ecosystems, but it is also the most abundant waste product produced by human society (Bayer *et al.*, 2007). These cellulose-containing wastes associated with sewage and leftover agriculture biomass are a potential renewable energy source (Bayer *et al.*, 2007).

The hydrolysis of cellulose is referred to as cellulolysis. Cellulolysis is a multi-step metabolic pathway that can be carried out either aerobically and anaerobically. The ability of microorganisms to hydrolyze cellulose depends on the production of extracellular enzymes that are either free (aerobic hydrolysis) or cell associated (anaerobic hydrolysis) (Lynd *et al.*, 2002). Aerobically, microorganisms secrete several endoglucanases, exoglucanases, and supplementary enzymes that can act together to attack and break down cellulose (Doi & Kosugi, 2004b), while anaerobic microorganisms rely on the formation of a large, extracellular enzyme complex called a cellulosome, which is made up of a scaffolding protein and numerous membrane-bound enzymes (Bayer *et al.*, 1985). Hydrolysis enzymes can be placed into three classes. The first class contains endoglucanases, which randomly cut

internal bonds creating oligosaccharides of various lengths from the polysaccharide cellulose chain. The second class is comprised of cellobiohydrolases, which are exoglucanases that cleave 2-4 linked glucose units from the end of larger chains of glucose. The third class is made up of β-glucosidases that hydrolyze the oligosaccharides into single glucose monomers (Lynd *et al.*, 2002, Pang *et al.*, 2009). Together, the previously described enzymes are referred to as cellulases, which make up part of the large family of glycoside hydrolase enzymes (Doi & Kosugi, 2004a).

1.1.2 Glycoside hydrolases and their industrial potential

Glycoside hydrolases are defined as enzymes with the ability to cleave the glycosidic bond between two carbohydrates, or between a carbohydrate and another functional group (Coker *et al.*, 2003). Traditionally, glycoside hydrolases were classified using functional similarity and substrate specificity (Coker *et al.*, 2003, Lynd *et al.*, 2002), but recently this diverse group of enzymes has been organized into 130 glycoside hydrolase families based on amino acid sequence similarity and reaction mechanisms (Coker *et al.*, 2003, Cantarel *et al.*, 2009). Cellulases, for example, are all classified as glycoside hydrolase enzymes (Duan & Feng, 2010).

Cellulases are used for a wide variety of industrial applications including denim stonewashing, detergent composition, beer brewing, and increasing the nutritional value of animal feed (Bhat, 2000, Pang *et al.*, 2009, Wilson, 2009). Currently, it is estimated that, by dollar value, cellulases represent the third largest group of industrial enzymes, and account for approximately 20% of global industrial enzyme use (Wilson, 2009). Cellulases also have

potential to become a sustainable energy source because they contribute to the conversion of biomass to biofuel (Bayer *et al.*, 2007, Pang *et al.*, 2009). The fermentation of glucose derived from cellulolytic sources has shown promise for the generation of large volumes of ethanol and other hydrocarbons. This process is attractive to industry due to the vast amount of cellulose found in natural ecosystems, and the excessive amount of cellulolytic waste generated by human activities. The development of an efficient and economical means to harness its decomposition would have substantial impact and utility. Microbial cellulolysis could potentially be used to generate large volumes of ethanol in a much more efficient manner than previously possible, allowing for the profitable production of biofuels. If an efficient method for converting cellulose to ethanol and other biofuels can be found, cellulases will become the most profitable group of enzymes (Wilson, 2009). Efforts have already begun to harness the bioconversion of organic waste to methane through anaerobic fermentation as an energy source through a biogas reactor at landfill sites (Krause *et al.*, 2008), and the generation of hydrogen gas from cellulolytic sources using anaerobic bacteria (Levin *et al.*, 2006, Yokoi *et al.*, 2002).

Numerous cellulases have been discovered and used in industrial applications, but the demand for novel cellulases with unique properties remains high. The majority of previously discovered cellulases are thermophilic enzymes with high reaction rates and optimal temperatures that are far above standard room temperature (Andrews *et al.*, 2004). However, there are situations where an enzyme with a lower optimal temperature is favourable for industry. Cold-adapted enzymes possess high specific activity at both low and moderate temperatures, and can easily be deactivated by moderate temperature increases (Gerday *et*

al., 2000). These traits are attractive to industry because cold-adapted enzymes will save energy costs by lowering the required reaction temperatures. However, cold-adapted enzymes have disadvantages as well. These include an increased risk of contamination due to lowered reaction temperatures, and the possibility of slower reaction rate (Georlette *et al.*, 2002).

1.1.3 The Arctic tundra: cold and loaded with cellulose

The global carbon cycle includes important biogeochemical processes at risk of imbalance due to anthropogenic inputs. The majority of these inputs come from the burning of fossil fuels and the modification of landscape (Falkowski *et al.*, 2000, Schuur *et al.*, 2008). Since the end of the pre-industrial era, there has been a 30% rise in total atmospheric carbon from 560 gigatons to 730 gigatons today (Zimov *et al.*, 2006). Representing one of the largest carbon reservoirs on Earth, soil contains 1500 gigatons of carbon (Davidson & Janssens, 2006, Zimov *et al.*, 2006). Considering the vast size of this reservoir, it is not surprising that one of the largest fluxes in the global carbon cycle comes from the emission of carbon dioxide from soil environments (Schlesinger & Andrews, 2000). The emission of carbon dioxide from soil is largely attributed to microbial respiration (Hanson *et al.*, 2000). Unsurprisingly, microbial communities are responsible for the majority of global carbon cycling, containing upwards of ten times the carbon stored in plants (Whitman *et al.*, 1998a). The importance of understanding the microbial role in the terrestrial carbon cycle is underlined by the fact that positive feedback from microbial carbon decomposition may be a major contributor to the future acceleration of global warming (Melillo *et al.*, 2002).

Northern ecosystems are important for climate change because the Arctic is expected to experience the most extreme changes in temperature due to climate change (Rahmstorf & Ganopolski, 1999, Epstein *et al.*, 2000). Northern ecosystems are also important because they store a disproportionally large amount of carbon in the soil. The tundra represents only 6.8% of the world's soil, but contains 13.7% of the planet's total soil carbon pool in its active layer (Post *et al.*, 1982), which is the layer that freezes and thaws each winter and summer (Schuur *et al.*, 2008). Permafrost, which is soil that remains frozen for a period greater than one year, covers 23.9% of the exposed land in the northern hemisphere (Schuur *et al.*, 2008), and contains up to 90% of total tundra carbon, which stored within the organic horizons and frozen mineral soils of permafrost (Mack *et al.*, 2004).

Because the majority of carbon is stored in frozen soils, northern ecosystems have traditionally been a carbon sink (Houghton *et al.*, 1998). However, with a warming climate the terrestrial biosphere's role is changing and the tundra is becoming a net carbon source (Houghton *et al.*, 1998, Melillo *et al.*, 2002, Schuur *et al.*, 2009). Traditionally, the reservoir of carbon in the Arctic increases each winter because of a vertical increase in soil surface (Schuur *et al.*, 2008), but with warming soils the carbon becomes available for microbial decomposition, and this increase in the carbon reservoir no longer occurs. The input of organic carbon to tundra soils via plant detritus has been overtaken by the output of carbon dioxide and methane due to thawing and subsequent microbial decomposition (Davidson & Janssens, 2006). Thawing of permafrost is a concern because the Arctic is warmer than it has been in the past 400 years (Chapin *et al.*, 2005) and models predict that the Arctic will

experience the most extreme changes in temperature due to climate change in the future (Rahmstorf & Ganopolski, 1999, Epstein *et al.*, 2000).

Understanding microorganisms of the carbon cycle is important for characterizing how carbon moves through the cycle. However, soil microbial communities are misrepresented by cultivation-dependent approaches, because only a small proportion of microorganisms in soil are estimated to be readily cultivable (Amann *et al.*, 1995). An alternative approach with less bias is required for microbial community analyses and accessing the organisms and enzymes involved in carbon cycling.

1.2 Culture independent studies

1.2.1 A change in approach

For most of its history, microbiology was based solely on the testing of physiological and biochemical behaviour of isolated organisms in pure cultures (O'Sullivan *et al.*, 2007). Despite having garnered an important foundation of knowledge in the characterization of microorganisms, cultivation-based techniques introduce bias into understanding the role that specific microorganisms play because the majority of microorganisms are not readily cultured. For any given habitat, less than 15% of all microorganisms are readily cultured, and for most environments less than 1% are cultured (Amann *et al.*, 1995). Of the estimated 52 bacterial phyla, half of them do not have cultured members (Rappe & Giovannoni, 2003). Because of our inability to culture the vast majority of microorganisms, it is impossible to deduce what functional role many of them play. Although this inability is a major challenge

in the field of molecular microbial ecology, recent advances in culture-independent approaches have begun to shed more light on the role of microorganisms *in situ*.

Two promising culture-independent approaches are DNA stable-isotope probing (DNA-SIP) and metagenomics. Though each methodology has been used on its own (Lorenz & Eck, 2005, Neufeld *et al.*, 2007a, Kim *et al.*, 2008, Berlemont *et al.*, 2009, Chen *et al.*, 2009, Han *et al.*, 2009), it is a combination of these two approaches that has enormous potential. Presently, only a few studies have combined DNA-SIP and metagenomics (Table 1.1).

Table 1.1 Summary of all studies to date coupling DNA SIP and metagenomics.

Researchers	Labelled substrate	Target	Library Type	Significance	Screening method
Schwarz *et al.* 2005	^{13}C-glycerol	Glycerol dehydratases	Small-insert	Industrial polymers	Sequence & function-based
Dumont *et al.* 2006	^{13}C-methane	Forest soil methylotrophs	BAC	methane monooxygenase operon ID	Sequence-based
Kalyuzhnaya *et al.* 2008	^{13}C-single carbon compounds	Sediment methyltrophs	Small-insert	Proof-of-principle/C_1 cycling	Sequence-based
Chen *et al.* 2008	^{13}C-methane	Peatland methanotrophs	Fosmid	*Methylocystis* involvement in methanotrophy	Sequence-based
Neufeld *et al.* 2008	^{13}C-methanol	Marine methylotrophs	Fosmid	Proof-of-principle/C_1 cycling	Sequence-based
Sul *et al.* 2009	^{13}C-biphenyl	Biphenyl dioxygenases	Cosmid	PCB degradation	Sequence & function-based

With the success of these studies, the combination of DNA-stable isotope probing and metagenomics promises to be a heavily used methodology in the future.

1.2.2 DNA-SIP

One of the major problems with the traditional screening of metagenomic libraries is that a very large number of clones need to be analyzed due to the low frequency of clones containing the desired genes (Schwarz *et al.*, 2006). The reason for this is that the targeted genes are commonly not contained in the microorganisms found in high abundance in the environment. This is especially true in complex environments such as soil and sediments, where there may be up to two billion cells per gram of substrate, representing many thousands of species (Whitman *et al.*, 1998b). In order to solve this problem, enrichment cultures can provide active microorganisms a competitive advantage, increasing their abundance. Traditionally, enrichment cultures were established by inoculating a sample with substrate known to be used by the targeted microorganisms (Schwarz *et al.*, 2006). Like with any biological technique, there are drawbacks to enrichment. The major drawback is that fast-growing community members became more prevalent. In spite of this, studies have shown that the combination of traditional enrichment and metagenomics is an effective way to increase the proportion of positive clones in metagenomic library screening, and to isolate new biomolecules from complex environmental communities (Entcheva *et al.*, 2001, Knietsch *et al.*, 2003a, Knietsch *et al.*, 2003b, Voget *et al.*, 2003).

The development of DNA-SIP vastly improved the ability to enrich the community members harbouring the desired genes. DNA-SIP functions by incorporating a stable-isotope labeled substrate, such as ^{13}C or ^{15}N, into the nucleic acids of microorganisms (Radajewski *et al.*, 2003). During this incubation, the environmental sample is exposed to the stable-isotope

labeled substrate as the only source of that substrate. This is a means to select for microorganisms containing the desired genes, as the target microorganisms metabolize the heavy substrate during enrichment. The increase in DNA density that results from this incubation is used to separate the "heavy" and "light" DNA via ultracentrifugation (Radajewski *et al.*, 2003). Following the recovery of heavy DNA, the nucleic acids should represent community members with the ability to metabolize labeled substrate, which are also the members containing the genes of interest. DNA-SIP could also be used as a "pre-enrichment" prior to metagenomic analysis, allowing for the detection of low-abundance but active species within metagenomic libraries, and facilitating the detection of novel enzymes (Chen & Murrell, 2010).

1.2.3 Metagenomics

Metagenomics refers to the direct retrieval and storage of DNA from microbial communities to screen for genes or enzymes of interest, without the prerequisite of cultivation. The term "metagenome" was coined by Jo Handelsman to describe the collective genomes of microbial communities that could be treated as one genomic unit, analogous to the genome of a single organism (Handelsman *et al.*, 1998). The first direct capture of "large-insert" environmental DNA came from a marine environment via research conducted by Stein and coworkers (Stein *et al.*, 1996), and many subsequent studies have since reported the capture and characterization of community DNA from both terrestrial and aquatic environments. These approaches have involved screening libraries directly for a variety of gene products (Handelsman, 2004) or directly sequencing bulk DNA to identify the

metabolic potential of microbial communities (Tringe & Rubin, 2005).

Since the use of metagenomics as a methodological approach began, studies have targeted a wide variety of genes and biocatalysts for "green" industrial processes and applications (Schmeisser *et al.*, 2007). Studies using metagenomics have discovered numerous industrially useful genes and enzymes. For example, over 80 novel lipolytic genes were found using a metagenomics-based approach from numerous environments (Schmeisser *et al.*, 2007). Metagenomic studies have also reported novel proteases (primarily used in detergents; (Gupta *et al.*, 2002), nitrilases (DeSantis *et al.*, 2002), genes exhibiting antibiotic activity (Brady *et al.*, 2009), and cellulases (Kim *et al.*, 2008).

In addition to industrial applications, metagenomics has helped with the understanding of microbial ecology and the potential biogeochemical roles of specific microorganisms. A good example of this was the discovery of proteorhodopsin-encoding genes through the usage of a bacterial artificial chromosome (BAC) library derived from western Pacific Ocean samples (Béjà *et al.*, 2000). This newly discovered class of photosynthetic genes within the rhodopsin family was transferred into an *Escherichia coli* host and expressed, resulting in a protein product that functioned as a light-driven proton pump. These findings were important because prior to this study, rhodopsins were only known to occur in extremely halophilic archaea. The biogeochemical significance of this discovery is that the light-driven proton pump potentially plays a large role in marine ATP production and light sensing. It was because of metagenomics that the novel discovery of the function, host, and habitat of this process took place. A second example includes the discovery of ammonia-oxidizing archaea, which was achieved through the high-throughput

metagenomic sequencing approach of Craig Venter within the marine environment (Venter *et al.*, 2004) and was subsequently confirmed, again by metagenomics, in terrestrial environments (Treusch *et al.*, 2005). Although metagenomic analysis has proven to be an extremely valuable technique, it is not without drawbacks.

1.2.4 Challenges and solutions

A problem that arises through the coupling DNA-SIP and metagenomic approaches is the low amounts of heavy DNA available for cloning following DNA-SIP. In some cases, particularly low-biomass environments, genome amplification must be applied after DNA-SIP, but prior to metagenomic cloning. Recently, multiple displacement amplification (MDA) has provided a means to overcome the low levels of SIP DNA recovered via ultracentrifugation. A whole genome amplification, which is a high-sensitivity PCR procedure, provides access to the genomic DNA of these low-biomass communities (Binga *et al.*, 2008). Multiple displacement amplification (MDA) and whole community genome amplification (WGA) have been applied to metagenomic studies. For example, a 2006 study concerned with subsurface-soil samples successfully applied MDA to the heavily contaminated samples (Abulencia *et al.*, 2006). Using WGA, the researchers were able to access genomic information that would have otherwise been inaccessible. Initial calculations estimated that a minimum of 11-88 kilograms of sample would be needed to extract the required amount of DNA for library construction, which is a completely unachievable amount of subsurface soil. Prior to MDA, only one of the three samples produced libraries, but whole-genome amplification overcame this (Abulencia *et al.*, 2006). Multiple

displacement amplification is the first whole genome amplification method based on an isothermal reaction and it entails denaturing a double-stranded DNA template, followed by an incubation. It is catalyzed by a phi29 DNA polymerase, which is derived from the *Bacillus subtilis* bacteriophage phi29, which possesses high strand displacement activity and processivity.

However, as is the case with many biological techniques, there are limitations associated with MDA. The three main limitations include nonspecific amplification as a result of primer-dimer formation or contaminating DNA template, chimeric DNA rearrangements, and representation bias (Binga *et al.*, 2008). With respect to MDA, representation bias refers to the alteration of gene relative abundances in a genomic DNA sample during the MDA reaction. The GenomiPhi Kit, which was the MDA protocol used in this research, under represented template DNA with a high G+C content in a recent study (Yilmas *et al.*, 2010) and thus the potential for representation bias is precedented. With proper enzymatic pretreatment, the errors introduced via these three main limitations can be reduced. For example, sequential treatment of amplified DNA with three enzyme reactions (a "debranching" step, digestion of single-stranded DNA, and nick translation) can reduce chimera formation by up to 80% (Zhang *et al.*, 2006). Regardless of its limitations, MDA provides access to DNA for metagenomic analysis, regardless of biomass in a sampled environment.

Once a sufficient amount of DNA has been extracted, purified, and amplified successfully from an environmental sample, it must be transferred to a clone library. As is usually the case with microbial ecology, there is more than one method to carry out this

process. A cloning strategy needs to be selected depending on what the study intends to achieve. Clone libraries fall into two general size groups: small-insert libraries in plasmid vectors (less than 15 kb), or large insert libraries in cosmid, fosmid (30-50 kb), or BAC (greater than 40kb) vectors (Daniel, 2005). To capture DNA fragments of this size into a vector, a suitable and gentle DNA isolation procedure must be selected. Most mechanical isolation procedures result in DNA shearing, yielding fragments too small to create large-insert libraries. DNA isolation itself involves two important processes: extraction and purification. To successfully prepare a metagenomic library, the nucleic acid extraction method must be chosen carefully based on sample characteristics and the desired size and purity of retrieved DNA (Tringe & Rubin, 2005). Extractions from soils or sediments frequently contain enzyme inhibitors (e.g., humic acids) that must be removed during DNA purification (Tringe & Rubin, 2005).

An alternative to traditional DNA purification methods for metagenomic library preparation (Tringe & Rubin, 2005) includes the recent availability of the synchronous coefficient of drag alteration (SCODA) electrophoresis system. It is not based on the chemical affinity of DNA and thus it allows the removal of contaminants that have similar chemical properties to DNA (Pel *et al*., 2009). The SCODA system exploits the physics of electrophoresis in response to alternating fields, similar to pulsed-field gel electrophoresis (PFGE). Nucleic acids exhibit a complex electrophoretic behavior when migrating through an agarose gel, and they display a nonlinear response in their drift velocity to changes in field magnitude (Pel *et al*., 2009). Because of this, under the influence of rotating electric fields, only molecules that respond in a nonlinear fashion have a net drift. This allows for the

purification of small amounts of high-molecular-weight DNA, even in the presence of large amounts of contaminants. This technology could potentially be the solution to SIP-related issues such as persistent contamination of extracts, low DNA yields from low biomass samples, or excessive shearing of DNA during extraction.

A suitable library type depends on whether low copy and low level gene expression or high level expression is preferred. If the study is targeting individual genes or gene products, or if the study is targeting operons and gene clusters encoding biosynthetic or degrading pathways (Schmeisser *et al.*, 2007) usually determines the suitable library type. After clone libraries have been effectively created, the vectors containing the environmental DNA must be packaged and transformed into a host. The host used for construction and maintenance is almost exclusively *Escherichia coli* in published libraries (Daniel, 2005). Once packaged into a host, the clone libraries need to be screened.

There are two different approaches to screening: sequence-based and function-based. Sequence-based screening encompasses a few different screening methods. PCR-based and hybridization-based approaches can be used to screen individual inserts for genes of interest. Both of these approaches require primers and probes that are obtained from known genes and gene products, limiting these methods to identifying previously known genes. A common use of this approach is the identification of 16S rRNA genes, and other genes with highly conserved regions (Daniel, 2005). Functional screens are based on the metabolic activity of the clone library. This method relies on identifying active clones that, commonly, have been plated on media that are selective for the activity of interest. Both sequence and functional-

based screenings have different advantages that, depending on the nature of the study, are beneficial.

Functional screening has been employed for the discovery of a wide range of novel enzymatic genes, including cellulases (Berlemont *et al.*, 2009), xylanases (Warnecke *et al.*, 2007), and genes with antibiotic activity (Allen *et al.*, 2008, Torres-Cortés *et al*., 2011). An example of a functional screen used to discover cellulases is the Congo Red overlay method. First described by Teather and Wood (1982), the method has been used by numerous metagenomic studies in the past and resulted in the detection of novel cellulases (Healy *et al.*, 1995, Kim *et al*., 2008, Pang *et al*., 2009**)**. It involves growing the metagenomic libraries on Luria-Bertani agar plates containing carboxymethyl cellulose (CMC) as the sole carbon source. Following a growth period, the colonies are overlaid with the same medium and incubated further. The plates are flooded with Congo Red indicator dye, which interacts with intact β-glucans, a group of molecules that includes cellulose, to produce a reddish-orange colour (Teather & Wood 1982**)**. In areas where the colony has metabolized the CMC in the media, there are no intact β-glucans and the Congo Red does not interact, causing a zone without a red colouration.

Gaining an understanding of microbial communities and their interactions is important for both industrial purposes and to advance our knowledge of our planet's environment. Cellulolysis is an important degradation process that has enormous potential for industrial use, and is an integral part of the global carbon cycle. Although recent studies have shown that the coupling of DNA-SIP and metagenomics has the potential to discover genes

of industrial significance, prior to this thesis these techniques had yet to be applied to genes coding for cellulose degradation.

1.3 Research Objectives and Hypotheses

The primary objective of this research was to determine whether the inclusion of a DNA-SIP pre-enrichment with labeled cellulose would increase the abundance of glycoside hydrolases, and more specifically cellulose degrading glycoside hydrolases, in the resulting metagenomic libraries generated from a Canadian Arctic tundra soil. I hypothesized that the addition of a DNA-SIP enrichment step prior to metagenomic screening would increase the abundance of the targeted cellulose degrading enzymes. My research marks the first attempt to combine DNA-SIP and metagenomics using cellulose as a substrate, as well as the first time this methodological approach was applied to tundra soil. The study also assessed the effect of multiple displacement amplification on metagenomic libraries with regards to bias in the representation of metagenomic libraries. Preliminary research in this thesis also aimed to determine an optimal DNA isolation and purification procedure for functional metagenomics with large-insert libraries.

Chapter 2 Methods

2.1 Introduction

The primary objective of this thesis was to determine whether the addition of a DNA-SIP pre-enrichment would increase the efficiency of recovering targeted gene sequences through metagenomics. Preliminary research also involved designing a protocol for DNA extraction and purification to produce high-molecular weight, high-quality DNA suitable for the generation of fosmid or cosmid metagenomic libraries. Once the optimal techniques were identified, they were used to extract DNA from the bulk soil samples, a cellulose enrichment culture, and the soil samples that had undergone a DNA-SIP incubation. All soil samples were collected from Resolute Bay, Nunavut, Canada on August 17, 2009 from the top 10 cm of surface soil. The soil was sieved (4.75 mm) and frozen at -80°C for further use.

2.2 DNA Extraction/Purification Optimization

To determine the optimal extraction and purification method for the production of high quality and high molecular weight DNA, three extraction/purification procedures were assessed. The first was the FastDNA Soil Extraction Kit (MP Biomedicals), which is a bead-beating based procedure. The second used a protocol modified from Zhou *et al*. (1996) combined with Wizard DNA Clean-Up Kit (Promega). The modified Zhou extraction utilizes a high-salt buffer and sodium dodecyl sulfate (SDS) and the Wizard kit uses resin-column

purification. The third method was the same modified Zhou extraction followed by purification using the Synchronous Coefficient of Drag Alteration (SCODA) instrument (Boreal Genomics). Following the extraction and purification, DNA from each technique was visualized using pulsed-field gel electrophoresis (PFGE) to determine DNA fragment lengths.

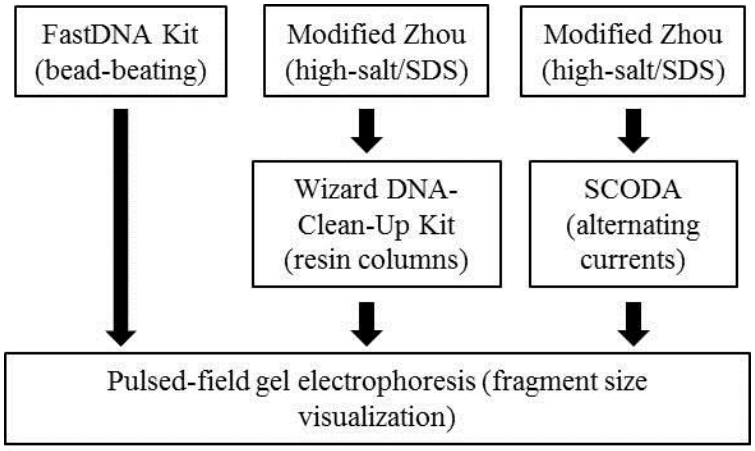


Figure 2.1 Diagram of the three DNA extraction and purification techniques tested.

2.2.1 FastDNA Isolation

DNA was isolated using the FastDNA Soil Extraction kit (MP Biomedicals) according to the manufacturer's instructions. Briefly, this isolation technique agitates 12 sample tubes at very high speeds to lyse cells. During this process, sample tubes contain polymer beads, a lysing matrix, and a mixture of detergent and salts to remove membrane lipids. Following isolation, the DNA was visualized and quantified using 1% (w/v) agarose gels stained with ethidium bromide dye (1 mg/mL). Gel electrophoresis was performed using a Mini-Sub-Cell GT system (Bio-Rad) at 85V for 40 minutes. Gels were imaged under ultraviolet light using

an AlphaImager® HP gel documentation system (Cell Biosciences) using 1Kb Plus ladder (Invitrogen) as a marker.

2.2.2 Modified Gentle Lysis Extraction

DNA was extracted according to the protocol published by Zhou *et al*. (1996). Briefly, this extraction utilizes a high-salt buffer and SDS detergent to lyse cells. The benefit of this extraction is it results in less shearing of the DNA than mechanical lysing. Two modifications were made to the protocol. A phenol:chloroform step was added before the chloroform:isoamyl alcohol purification step in the original procedure. The DNA precipitation method was also altered to increase the amount of DNA recovered. The modified precipitation added ammonium acetate (final concentration of 2.5M) in addition to isopropanol, instead of using just isopropanol as suggested in the original method. Following extraction, DNA was visualized using agarose gel electrophoresis (see 2.2.1) and was subjected to one of two purification methods (see 2.2.3 and 2.2.4).

2.2.3 Wizard DNA Clean-Up Column Purification

Following the modified gentle lysis extraction, the DNA must be purified to remove contaminants, such as humic acids, to produce high-quality DNA for metagenomic library construction. The Wizard DNA Clean-up column (Promega) was used to purify the DNA according to the manufacturer's instructions. In brief, the DNA is bound to a resin, which in turn, is bound to a binding matrix within a spin column. Following a wash step, the DNA is

eluted from the spin column's binding matrix using centrifugation. Samples were visualized and quantified using agarose gel electrophoresis (see 2.1.1). Samples were stored frozen at -20°C until needed for PFGE.

2.2.4 SCODA Purification

The second method used to purify DNA utilized the SCODA instrument from Boreal Genomics. There are three steps that occur when using the SCODA instrument to purify DNA: gel boat preparation, DNA injection, and DNA wash/concentration. However, before attempting to purify DNA using the SCODA instrument, one must first have a working knowledge of SCODA nomenclature. Please refer to Table 2.1, Table 2.2, and Figure 2.2 when attempting to use the SCODA procedure.

Table 2.1 List of the SCODA gel boat components. Refer to Figure 2.2 for each component's location on the gel boat.

Letter	Part
A	Buffer reservoir
B	Buffer reservoir
C	Buffer reservoir
D	Buffer reservoir
E	Sample chamber
F	SCODA gel chamber
G	Injection barrier gel chamber

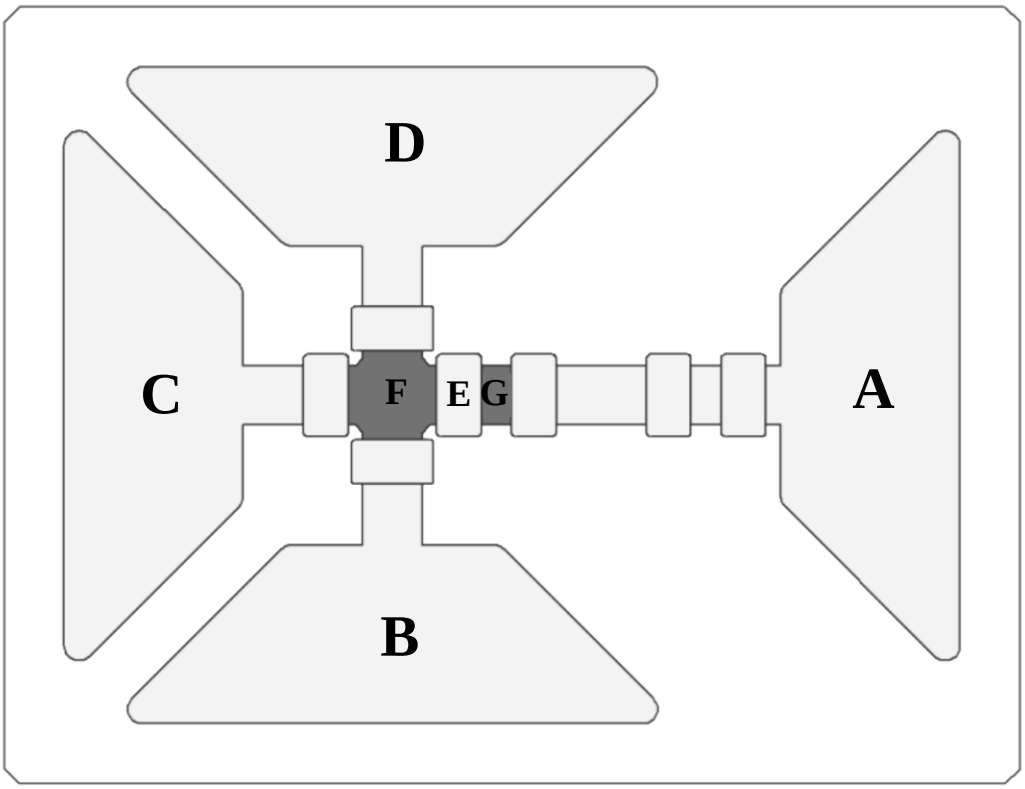


Figure 2.2 Overview of the SCODA gel boat arrangement.

Table 2.2 List of SCODA components and their functions.

Component	Function
gel caps (4mm, 1mm)	Create enclosed space for thin (1 mm) or thick (4 mm) SCODA gel formation; diffusive extraction caps have hole in center for TBE/DNA extraction
Gel plug	Creates 100 µL well for TBE buffer
Solid rubber dam	Barrier to form SCODA gel and injection barrier gel
Cutout rubber dam	Barrier to form SCODA gel, but cutout on the inside to allow the instrument's electrodes to be lowered

In this study, a 4 mm SCODA gel was used with a diffusive extraction gel cap. This allows for the concentrated DNA to be pipetted from center well created by the gel plug,

rather than be extracted from the gel itself. The procedure for the preparation of the gel boat is as follows:

1. Place the 4 mm diffusive extraction gel cap over the SCODA gel chamber.
2. Place a solid rubber dam in the chamber on each side of the injection barrier gel chamber (one of these is the sample chamber).
3. Place a cutout rubber dam in each of the chambers between the SCODA gel chamber and buffer reservoirs B and D.
4. Raise the buffer reservoir C end of the gel boat by approximately one inch.
5. Prepare a 1% agarose gel by dissolving 0.1 g of agarose in 10mL of 0.5X TBE (add 10 µL of 1000X SYBR Green I stain if DNA is to be visualized in the SCODA gel).
6. Pipette approximately 4 mL of gel into the SCODA gel chamber and approximately 0.5 mL of gel into the injection barrier gel chamber.
7. Allow the gel to solidify for 20 minutes at room temperature.
8. Remove the gel plug from the gel, being careful to not disrupt the seal between the bottom of the SCODA gel and the gel boat.
9. Remove the rubber barriers from both sides of the injection barrier gel, being careful not to disrupt the gel's seal with the gel boat.
10. Leave the rubber barriers between the SCODA gel and buffer reservoirs B and D in place.
11. Carefully cut away any gel that overflowed into buffer reservoir C and discard.

12. Place the gel boat inside the SCODA, and add 100 µL of 0.5X TBE to the diffusive extraction well in the centre of the gel cap, and 25 mL of 0.5X TBE to buffer reservoirs A and C (make sure there is no leakage of buffer from reservoir A into the sample chamber or the DNA will not stack correctly in the SCODA gel).
13. Add 1 mL of sample to the sample chamber.
14. Lower the electrode board so that the steel pins lock the gel boat into place, and place the electrodes surrounding the SCODA gel in the raised position.
15. Close the door to the SCODA, turn the power on, and turn on the water cooling system tap.
16. Open the SCODA Control software on the SCODA computer.
17. Set injection time to 15 minutes, and voltage to 90 V/cm. Click “start run”.

Following the injection, the DNA will be stacked within the SCODA gel. It must now be washed to remove co-contaminants like humic acids and concentrated to centre the DNA into diffusive extraction well buffer. The procedure for the wash and concentration steps is as follows:

1. Turn off the power and lift the electrode board, freeing the gel boat.
2. Aspirate all the remaining sample from the sample chamber, rinse, and aspirate again.
3. Remove the injection barrier gel.

4. Remove the cutout rubber dams between the SCODA gel chamber and reservoirs B and D.
5. Add 25 mL of 0.5X TBE to buffer reservoirs B and D.
6. Lower the electrode board as in the injection step, but place the electrodes surrounding the SCODA gel in the lowered position.
7. Close the door, turn the power back on and open SCODA Control software.
8. Set parameters for the wash step as 30 V/cm, 2 hours total run time, 4 second rotation period with a 1.5V bias on the electrodes in reservoir B and D and a 3V bias on the electrode in reservoir A (this will wash the contaminants into reservoir A).
9. Also set the parameters for the concentration step. It will run continuously after the wash. The parameters are the same, but without an electrode bias.
10. After the 4 hour run, lift the electrode board and pipette the 100 µL of 0.5X TBE (containing the purified DNA) into a sterile 1.5 mL microfuge tube.
11. Precipitate DNA using ammonium acetate (final concentration 2.5M) and 1.0X volume of isopropanol for 2 hours at -20°C.
12. Wash twice with cold 70% ethanol, dry pellet, and resuspend in 0.5X TBE.
13. Wash the gel boat, gel cap, gel plug, and and rubber dams with 10% bleach and sterile water. Air dry.

Following purification DNA was visualized and quantified using agarose gel electrophoresis (see 2.2.1). DNA quality was measured with A_{260}/A_{280} using a NanoDrop 2000 Spectrophotometer (Thermo-Scientific) according to the manufacturer's instructions. Samples were stored at -20°C until needed for PFGE.

2.2.5 Pulsed-Field Gel Electrophoresis (PFGE)

The CHEF Mapper XA Pulsed-Field System (Bio-Rad) was used to compare the size of DNA fragments following the extraction and purification according to the manufacturer's instructions. PFGE utilizes a long run time and alternating electric currents to visualize DNA with a very large size range (from 100 bp up to 10 Mbp). PFGE gels were stained in 400 mL of 1X TAE for 1 hour using 10 µL of 10,000X SYBR Green I. DNA was visualized using a Dark Reader (Mo Bio Tec), which uses visible blue light as an excitation source for the SYBR Green I stain. Using a UV imager is permissible as long as the DNA is not going to be used to construct metagenomic libraries.

2.3 DNA-SIP and Metagenomics

Following the optimization of a DNA extraction and purification method, the method was used to prepare DNA for shotgun metagenomic sequencing.

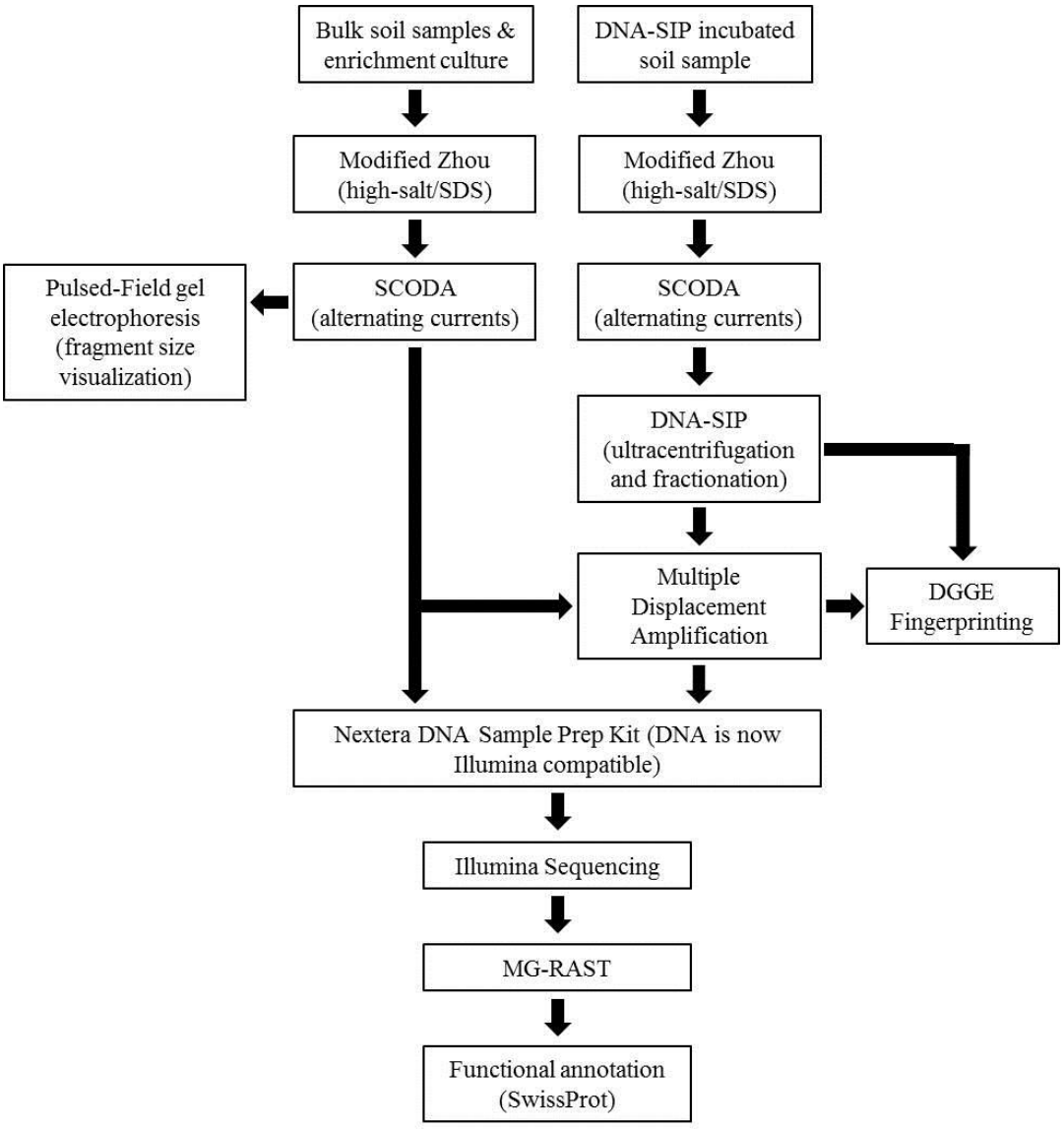


Figure 2.3 Schematic of sample processing for DNA-SIP and metagenomics. Please note duplicate bulk soil samples were used, as well as duplicate bulk soil and DNA-SIP multiple displacement amplification samples.

The following subsections outline methods for preparing metagenomic libraries from DNA-SIP DNA. The order will only reflect the DNA-SIP pipeline (the pipeline on the right side of Figure 2.3), but the methods used for the bulk soil samples were done using the same procedures. The modified gentle extraction (Zhou *et al*., 1996) and SCODA purification will not be detailed as they were outlined in section 2.2. Soil samples were collected from Resolute Bay, Nunavut on August 17, 2009. Surface soil from the top 10 cm of the organic horizon below a vegetated area was collected. Soil was sieved (4.75 mm) and stored at -80°C. Soil pH, total carbon content, total nitrate content, and soil texture were analyzed by the Soil Nutrient Laboratory at the University of Guelph (see Table 3.3 in the Results and Discussion section).

For bulk soil samples, duplicate modified Zhou extractions were done, and subsequently each extraction had multiple SCODA purifications applied. This is a result of only injecting 1 mL of sample per SCODA run. For each pooled extraction, three separate MDA reactions were performed. The same MDA procedure was applied twice to the DNA from fraction 7 from the DNA-SIP gradient. These procedures resulted in seven samples: two duplicate bulk soil samples, an MDA sample for each duplicate bulk soil, fraction 7 from the DNA-SIP gradient, and two duplicate MDA samples of fraction 7 from the DNA-SIP gradient. In addition to these seven samples, a cellulose enrichment culture from collaborators at Ryerson University was also analyzed. The enrichment culture is a stable mixed culture that was enriched from compost, and it exhibited the ability to rapidly degrade cellulose at 60°C under static aerobic conditions (data not shown).

2.3.1 DNA-SIP

For the DNA-SIP sample incubation, soil was incubated with a heavy-isotope carbon (^{13}C) substrate. In this case the heavy-isotope carbon was incorporated into cellulose, and cellulose was used as a substrate for the DNA-SIP incubation. To obtain heavy-isotope labeled cellulose, *Gluconacetobacter xylinus* was cultured with $^{13}C_6$-labelled glucose. *G. xylinus* produced cellulose pellicles that were recovered for purification. Briefly, the cellulose purification involved the removal of the pellicle from the culture, followed by a wash with ddH_2O, treating with sodium hydroxide, followed by a second treatment with acetic acid, and a second ddH_2O wash step. The cellulose was then frozen and lyophilized to remove moisture and ground into small pieces using a mortar and pestle with liquid nitrogen. For the complete protocol for cellulose production refer to Eric Dunford's dissertation (Dunford, 2011).An important control to include for DNA-SIP experiments is an identical incubation that uses native substrate (^{12}C). This control incubation provides a comparison to ensure that any labeling of nucleic acid is in fact a result of the heavy-isotope substrate being metabolized, and not an artifact of ultracentrifugation or the result of G+C content-based density different in DNA (Neufeld *et al.*, 2007b).

Soil microcosms were incubated aerobically for two months in 100 mL crimp-top vials. 200 mg of ^{13}C-labeled cellulose was added to the SIP incubation. The vials were incubated at a temperature of 15°C, to simulate air temperatures in the Canadian Arctic during the summer months. It has been demonstrated that tundra surface soils reach temperatures between 20 and 26°C during the summer (Chapin *et al.*, 2005). This incubation

temperature also was intended to select for the growth of cold-adapted microorganisms that could possess enzymes capable of functioning at lower temperatures, which have excellent industrial potential (Gerday *et al.*, 2000).

Following the two month incubation the soil was stored at -80°C for further use. DNA was isolated using the modified Zhou extraction and SCODA purification (see 2.2.2 and 2.2.4). Following DNA extraction, DNA-SIP was completed as previously described (Neufeld *et al.*, 2007b). Briefly, extracted DNA is added to a cesium chloride (CsCl) gradient and ultracentrifuged for 40 hours. Based on density, the gradient was then fractionated into 12 fractions. The DNA in each fraction was visualized using agarose gel electrophoresis (see 2.2.1) and 16S rRNA gene PCR coupled with denaturing gradient gel electrophoresis (DGGE). The resulting heavy DNA was used for metagenomic library construction.

2.3.2 Multiple Displacement Amplification (MDA)

MDA was applied to both the heavy DNA and bulk soil samples using the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) according to the manufacturer's instructions. With each MDA reaction, both positive and negative controls were included with control genomic DNA and no added DNA, respectively. All reactions yielded 4-7 µg of DNA from the positive control and no DNA synthesis from the negative controls. Sample reactions were pooled and stored at -20°C.

2.3.3 Polymerase Chain Reaction (PCR) Amplification)

Before using DGGE to generate bacterial community fingerprints, samples were amplified using PCR. In this case, the general bacterial-specific primers 341f-GC clamp (5'-CGC CCG CCG CGC GCG GCG GGC GGG GCG GGG GCA CGG GGGG CCT ACG GGA GGC ACG AG- 3' [GC-clamp underlined]) and 518r (5'-ATT ACC CGC GCT GCT GG- 3') were used to amplify the 16S rRNA V3 region's gene sequence (Muyzer *et al.*, 1993). Each reaction mixture contained 19.6 µL of UV-treated PCR H_2O, 2.5 µL of ThermoPol buffer (New England BioLabs), 15 µg of 10 ng/µL bovine serum albumin, 5 pmol of 341f-GC clamp primer, 5 pmol of 518r primer, 0.05 µL of 100 nM dNTP mixture (New England BioLabs), 0.25 µL *Taq* DNA polymerase, and 1.0 µL of template DNA. Amplification conditions were 95°C for 5 minutes followed by 30 cycles of 1 minute at 95°C, 1 minute at 55°C, and 1 minute at 72°C. Final extension occurred at 72°C for 7 minutes. The PCR resulted in fragments approximately 180 bp in length. DNA was visualized using agarose gel electrophoresis (see 2.2.1), and stored at 4°C for short term storage (one week or less), or at -20°C for the longer term.

2.3.4 Denaturing Gradient Gel Electrophoresis (DGGE)

DGGE was used to generate a bacterial fingerprint for the DNA-SIP gradient fractionation, and to generate bacterial community fingerprints of all samples. Using the PCR products from above, general bacterial 16S rRNA gene profiles were generated. In all cases a 30% to 70% denaturing gradient in a 10% acrylamide gel was used (C.B.S. Scientific). Gels

were run for 14 hours at 85V according to a previously published procedure (Green *et al*. 2010). Gels were stained with SYBR Green I nucleic acid gel stain (Invitrogen) and visualized using the Pharos FX™ Plus Molecular Imager (Bio-Rad).

2.3.5 Illumina Sequencing

Following community profiling, Illumina libraries were constructed for all samples. The metagenomic samples were pre-processed with the Nextera DNA Sample Prep Kit (Epicentre). In brief, the Nextera process uses in vitro transposition to tag and fragment DNA, then purifies and concentrates the tagged and fragmented DNA using the DNA Clean and Concentrator Kit (Zymo Research), resulting in DNA 100 bp DNA fragments This is followed by the addition of Illumina-compatible primers and the amplification of the library using a limited-cycle PCR. Again, the DNA is purified using the DNA Clean and Concentrator Kit. This procedure results in template DNA that is Illumina-compatible. Following the production of shot-gun fragment libraries, samples underwent paired-end (PE) Illumina sequencing using the Illumina HiSeq2000 sequencing system.

2.3.6 MG-RAST and Functional Annotation

Each read from each Illumina library was uploaded to MG-RAST separately so an estimate of the variation within the annotation of duplicate libraries could be attained. After the completion of uploading, each of the 16 PE libraries (8 duplicate samples) were processed by the MG-RAST quality control pipeline, omitting the “demultiplexing and

model organism" screening steps. Libraries were annotated using SwissProt with no maximum e-value cutoff, a 54% minimum percentage identity cutoff, and a 30 bp minimum alignment length cutoff. All of the genes annotated as glycoside hydrolases were exported to MS Excel and sorted by sample into their CAZy families (www.cazy.org). Using MS Excel, a Bray-Curtis dissimilarity matrix was constructed for all samples using the number of gene sequences in each CAZy family per one million sequences. This provided a measure of the inter-sample relationships with regards to glycoside hydrolases. The Bray-Curtis dissimilarity equation is:

$$BC_{ab} = \frac{S_a + S_b - 2C_{ab}}{S_a + S_b}$$

Where C_{ab} represents the sum of the minimum proportions of the GH enzyme gene sequences and S_a and S_b represent the total number of GH gene sequences found in each sample. The matrix was constructed by subtracting the calculated the Bray-Curtis value, and dividing it by 100 for each sample in relation to every other sample. From the Bray-Curtis matrices unrooted trees were constructed using the neighbor-joining method (Saitou & Nei, 1987) with the APE package in R (Paradis *et al.*, 2004). Two trees were constructed, one for the matrix relating to all glycoside hydrolase families, and the other with only the cellulose-degrading glycoside hydrolase families (GH 5-12, 26, 44, 48). In addition to the Bray-Curtis dissimilarity value, the Shannon index was calculated for each sample using MS Excel. The Shannon index is a measure of diversity, and it is calculated using the following formula:

$$H' = \sum p_i \ln(p_i)$$

Where H’ represents the Shannon index, and p_i represents the relative abundance of each group of enzymes, which is calculated as the proportion of enzymes in a specific class to the total number of enzymes (Spellerberg & Fedor, 2003).

Chapter 3 Results and Discussion

3.1 Introduction

The research described in this chapter was focused on determining whether inclusion of a DNA-SIP pre-enrichment with labeled cellulose would increase the abundance of glycoside hydrolases, and more specifically cellulose-degrading glycoside hydrolases, in the resulting metagenomic libraries generated from a Canadian Arctic tundra soil. The described research also assessed the effect of multiple displacement amplification on metagenomic libraries with regards to bias in the representation of metagenomic libraries. Preliminary research presented in this chapter also determined an optimal DNA isolation procedure for functional metagenomics with large-insert libraries. Note that preliminary research was also done using functional metagenomic screening to discover novel cellulase genes. Soil from an Elora corn-field was used to produce a cosmid library from purified DNA and clones were screened using the Congo Red Overlay method (see 1.2.4). No positive clones were found. Although this initial research led to the work presented in this thesis, these negative results are not presented in this thesis.

3.2 Optimization of DNA Extraction and Purification

In order to identify a suitable DNA extraction method for retrieving DNA fragments suitable for either cloning or next-generation sequencing, a DNA extraction/purification

method was tested for the production of high-quality and high-molecular-weight DNA. Three methods were compared for yield, purity, and amount of DNA shearing in my project (see Figure 2.1). The first method tested was the FastDNA Soil Extraction Kit (MP Biomedicals). This method involved a bead-beating extraction and purification. The second method involved a high-salt and SDS-based gentle lysis extraction (Zhou *et al.*, 1996) (see section 2.2.2) followed by purification using the Wizard DNA Clean-Up Kit (Promega). This kit uses a resin-binding spin-column to wash away any co-contaminants from the DNA (see section 2.2.3). The third method used the same gentle lysis (Zhou *et al.*, 1996) extraction, but was followed by purification using the SCODA (Boreal Genomics) instrument. The SCODA instrument exploits the physics of electrophoresis in response to alternating fields, which results in only molecules that respond in a nonlinear fashion (i.e. nucleic acids) exhibiting a net drift (see sections 1.2.4 & 2.2.4).

The FastDNA Kit (MP Biomedicals) resulted in 25 µg DNA/ one g soil, which was the highest yield (Table 3.1), but DNA fragment sizes were under 20 Kb in length (Figure 3.1b). The small size of DNA fragments makes this method unsuitable to isolate DNA for use in cosmid, fosmid, or BAC metagenomic library construction. The modified gentle lysis extraction followed by purification with the Wizard DNA Clean-Up Kit (Promega) produced ~3 µg DNA/ one g soil, which was the lowest yield (Table 3.1). It also resulted in DNA fragments with a maximum size of 35 Kb (Figure 3.1d), making this method unsuitable for large-insert library construction. Lastly, a modified gentle lysis extraction coupled with SCODA (Boreal Genomics) purification produced ~10 µg DNA/ one g soil (Table 3.1), and

DNA fragment sizes of at least 30 Kb (Figure 3.1b), which is a suitable range for cosmid, fosmid, or BAC metagenomic library construction.

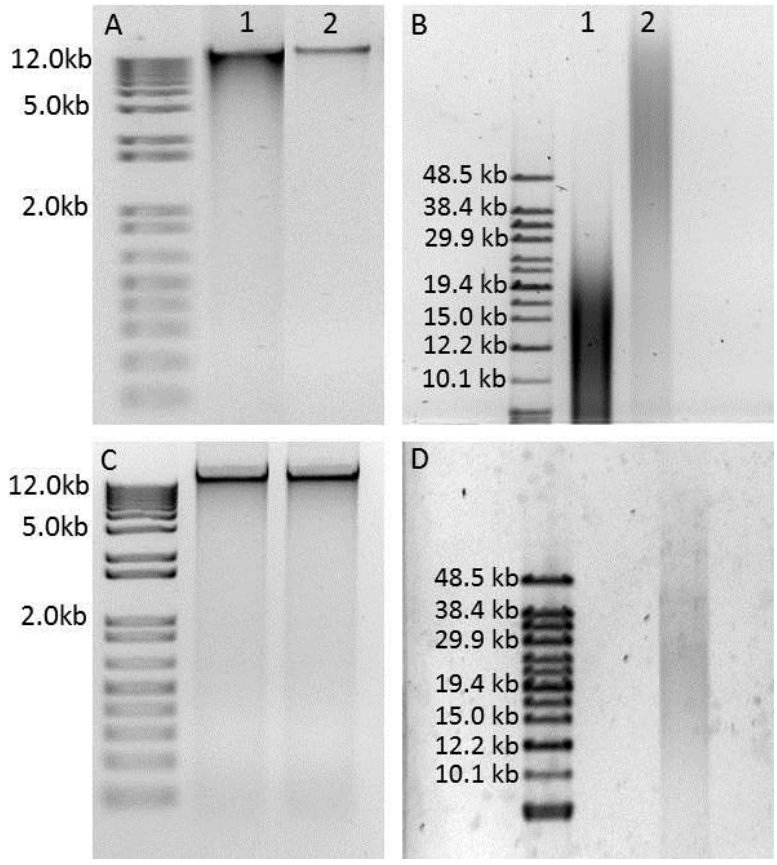


Figure 3.1 Gels with DNA from each of the three extraction and purification methods. A) 1% agarose gel showing 50 ng of DNA extracted and purified using the FastDNA Kit (Lane 1) and the modified gentle lysis method and SCODA (Lane 2). B) 1% PFGE gel showing DNA fragment sizes of DNA extracted and purified using the FastDNA Kit (Lane 1) and the modified gentla lysis extraction and SCODA (Lane 2). C) 1% agarose gel showing ~70 ng of DNA extracted and purified using the gentle lysis extraction coupled with the Wizard DNA Clean-Up Kit (both lanes). D) 1% PFGE gel showing DNA fragment size of extracted and purified DNA using a modified gentle lysis extraction and a Wizard DNA-Clean-Up Kit purification. For the 1% agarose gels (A,C) the marker is the 1 kb Plus marker (Invitrogen), and for PFGE gels (B,D) the marker is the High Molecular Weight marker (Invitrogen).

Table 3.1 DNA yield and size following extraction using the three methods tested. Concentration was determined using gel quantification on a 1% agarose gel (see 2.2.1), and fragment sizes were determined using PFGE (see 2.2.5). Yield was calculated per 1 g of soil used in each extraction.

	Yield	Fragment Size
FastDNA Kit	~25 ug	<20 kb
Gentle lysis + Wizard DNA Clean-Up Kit	~3 ug	<35 kb
Gentle lysis + SCODA	~10 µg	>30 kb

The results demonstrated that a modified gentle lysis extraction followed by SCODA purification led to an adequate yield of high molecular weight DNA suitable for the construction of large-insert, functional metagenomic libraries. This method also produced high quality DNA, with a A_{260}/A_{280} of ~1.8. Note that this A_{260}/A_{280} measurement followed ammonium acetate precipitation and suspension of DNA recovered from the SCODA purification. The other two procedures failed to produce DNA with sufficiently high molecular weight for possible fosmid or cosmid cloning (i.e. 30-50 kb). Bead-beating extractions have been widely used for sequence-based metagenomic approaches (Riesenfeld *et al.*, 2004, Turnbaugh *et al.*, 2009, Lanzen *et al.*, 2011), due to its very high yield (Table 3.1) of purified DNA. However, this study demonstrated that because of the mechanical nature of the extraction, excessive shearing resulted in DNA fragments too small (Fig. 3.1b) for large-insert metagenomic libraries. A gentler extraction method was necessary to produce these types of libraries.

This research used the modified gentle lysis extraction method in contrast with bead-beating, to determine whether it produces an adequate yield of high-molecular-weight DNA.

This extraction procedure has been used by numerous researchers (Neveu *et al.*, 2011, Chen *et al.*, 2008, Sul *et al.*, 2009) utilizing large-insert, functional metagenomic libraries, despite the fact that this procedure is insufficient to yield high amounts of purified DNA due to the co-precipitation of humic acids, which commonly occurs when extracting from soils (Tringe & Rubin, 2005). To circumvent this problem, the "Zhou method" suggests using a Wizard DNA Clean-Up (Promega) column to remove humic acids. However, the synchronous co-efficient of drag alteration (SCODA) electrophoresis system provides a new system for DNA purification. Both of these purification methods were tested following Zhou extraction. This study demonstrated that the yield of DNA was relatively low (Table 3.1) following the use of resin-binding columns for purification, and fragments sizes too small (Figure 3.1d) for large-insert metagenomic libraries. The SCODA purification method is not based on the chemical affinity of DNA and allows the removal of contaminants that have similar chemical properties to DNA (Pel *et al.*, 2009). Instead, the SCODA system exploits the physics of electrophoresis in response to alternating fields. Only molecules that respond in a nonlinear fashion (i.e. nucleic acids) have a net drift. This allows for the purification of small amounts of high-molecular-weight DNA in the presence of large amounts of co-contaminants (Pel *et al.*, 2009). This study demonstrated that using the SCODA instrument for DNA purification following a Zhou extraction provided adequate yields of DNA in a suitable size range for large-insert libraries (Table 3.1, Figure 3.1b). This DNA isolation method is novel in itself, as SCODA technology is in its infancy and not widely used. The ability to isolate high-quality, high molecular weight DNA effectively and efficiently using this method should

allow for the future researchers to successfully construct large-insert metagenomic libraries. This method was also used in the primary research of this thesis (see section 3.3).

3.3 DNA-SIP and metagenomics

This research was conducted on eight samples. The description and names of these samples is provided in Table 3.2. The cellulose degrading enrichment culture (CDE) was provided by Patrick Ronan of Dr. Martina Hausner's laboratory at Ryerson University in Toronto, Ontario. It was a stable mixed culture able to rapidly degrade cellulose at 60°C under static aerobic conditions, that was enriched from a compost sample.

Table 3.2 Names and description of the eight samples used in this research.

Sample Name(s)	Description
RB1A, RB1B	Bulk tundra soil samples; duplicate extractions
RB1A-MDA, RB1B-MDA	Multiple displacement amplification of RB1A and RB1B
SIPf7	"heavy" DNA (fraction 7) resulting from ultracentrifugation and gradient fractionation of cellulose SIP incubation
SIPf7-MDA1, SIPf7-MDA2	Duplicate multiple displacement amplifications of heavy DNA (fraction 7; SIPf7)
CDE	Cellulose degrading enrichment culture (positive control)

3.3.1 Bulk soil DNA extraction, and DNA-SIP microcosm incubation

The soil used for this research was collected in Resolute Bay, Nunavut (82°29.7' N, 62°20.1'W). Soil bulk density, nitrate, organic carbon content, pH and texture are shown in Table 3.3.

Table 3.3 Characteristics of the Arctic tundra soil used for both bulk soil samples and the DNA-SIP incubation.

Test	Result
Bulk Density	0.714 g/cm^3
Nitrate	2.07 mg/kg
Organic carbon content	43.6% of dry weight
pH	6.7
Texture (Sand, Silt, Clay)	16.6%, 32.6%, 50.8%

Duplicate modified Zhou extractions were performed on the tundra soil samples. Each extraction used 2 g of soil, and was followed by SCODA purifications. For bulk soil sample 1 (RB1A), approximately 21 µg of DNA was recovered, and approximately 23 µg of DNA was recovered for bulk soil sample 2 (RB1B). These samples were frozen at -20°C until needed for multiple displacement amplification (see 3.3.3).

The same soil sample was used for the DNA-SIP enrichment incubation. Following the two-month incubation, soil was stored at -80°C until needed. The modified Zhou extraction was applied to 0.5 g of thawed soil from the two month incubation. This method uses a high-salt buffer and SDS (see 2.2.3 and 3.2) to extract high molecular weight DNA (>30 kb). Following SCODA purification, the yield of total DNA was estimated to be 3.3 µg

of DNA. This estimate was obtained using agarose gel electrophoresis quantification. All 3.33 µg of DNA was added to the CsCl gradient tube, and following ultracentrifugation and density gradient fractionation, the majority of extracted DNA was associated with fractions 7 through 12 using agarose gel electrophoresis (Figure 3.2).

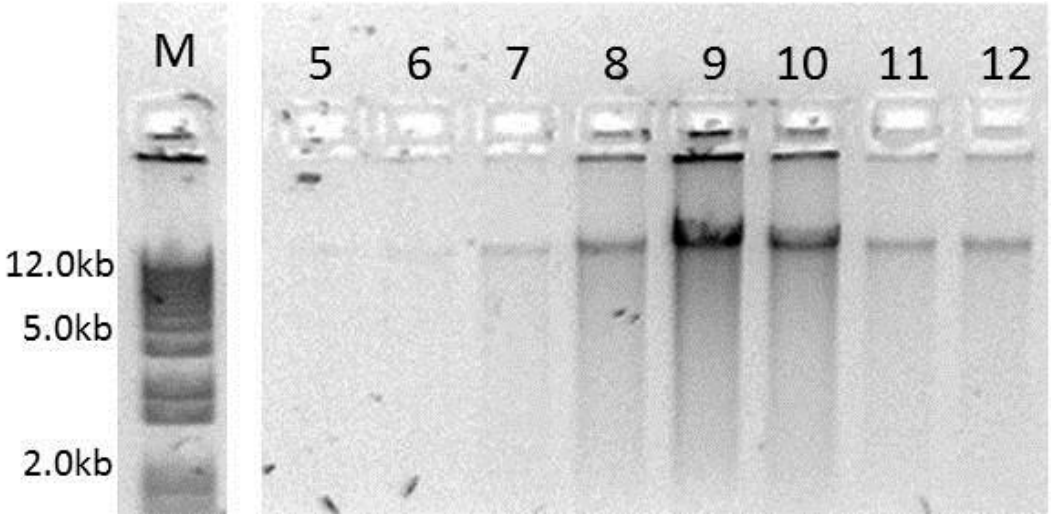


Figure 3.2 1% agarose gel electrophoresis demonstrating the DNA content of gradient fractions. Fractions are shown in order of decreasing density from left to right, beginning with fraction 5. The 1 kb Plus DNA marker (Invitrogen) is shown for comparison.

3.3.2 DGGE fingerprinting of fractions

Before conducting MDA on the heavy DNA, DGGE was used to visualize and assess the general structure of the bacterial community involved in the metabolism of cellulose (Figure 3.3).

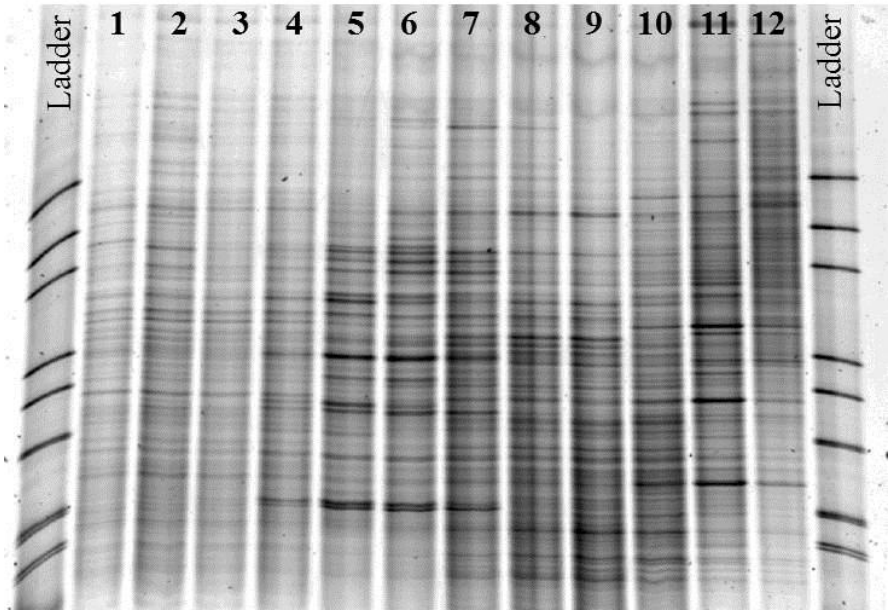


Figure 3.3 A 10% DGGE polyacrylamide gel with a 30-70% denaturing gradient containing the DNA fingerprint from all DNA-SIP fractions. The fractions are in order of decreasing density from left to right. DGGE ladders of cloned PCR products are shown in the outermost lanes for comparison purposes.

The presence of cellulose-degrading bacteria was confirmed using DGGE, as unique community fingerprints were apparent in fractions 5-7; most predominant bands in fraction 5-7 do not appear in fractions 9-12. Previous research has demonstrated that the fractionation of the DNA-SIP density gradient should result in the observation of labeled DNA in fractions 4-7 (Neufeld et al., 2007b). This is a result of microorganisms incorporating the heavy ^{13}C-carbon from the cellulose provided as a substrate in the incubations, and is an indication of successful DNA-SIP enrichment. Unlabelled community DNA has previously been shown to appear in fractions 9-12 (Neufeld et al., 2007b). In this case, the unlabeled DNA was extracted from community members unable to metabolize the heavy ^{13}C-carbon labeled cellulose.

3.3.3 MDA and DGGE fingerprinting

To perform functional metagenomics, MDA is necessary to produce the microgram quantities of DNA required to construct large-insert libraries. However, there are limitations associated with MDA. Namely, nonspecific amplification as a result of primer-dimer formation or contaminating DNA template, chimeric DNA rearrangement formation, and representation bias (Binga *et al.*, 2008). Due to the low amounts of DNA recovered in the heavy fractions, MDA was used to amplify the DNA in fraction 7. Fraction 7 was selected for amplification using MDA because it contained more DNA than fractions 5 and 6 (Figure 3.2). Fraction seven contained ~120 ng of DNA in 30 µL of TE. This resulted in 4.1 ng of DNA being used as template in the MDA reaction. Selecting the fraction with largest quantity of DNA aimed to lower representation bias introduced by MDA; increasing the amount of template DNA has been shown to decrease representation bias (Detter *et al.*, 2002, Bergen *et al.*, 2005, Neufeld *et al.*, 2008).

MDA was also applied to both bulk soil samples. The amount of template DNA provided from these samples was considerable larger (~58 ng). The inclusion of a MDA step for these samples should provide a measure how the amount of template DNA affects representation bias. For each of the bulk soil samples (RB1A, RB1B), three MDA reactions were performed and pooled upon completion. However, for the DNA-SIP fraction 7 (SIPf7) sample, six MDA reactions were performed, which were then pooled into duplicate samples (SIPf7 MDA1, SIPf7 MDA2). Following the MDA reactions, the DNA was quantified using gel electrophoresis (Figure 3.4). Following MDA, there was 4.7 µg of DNA from RB1-A, 2.1

µg of DNA from RB1-B, 11.5 µg of DNA from SIPf7-MDA1, and 11.7 µg of DNA from SIPf7-MDA2.

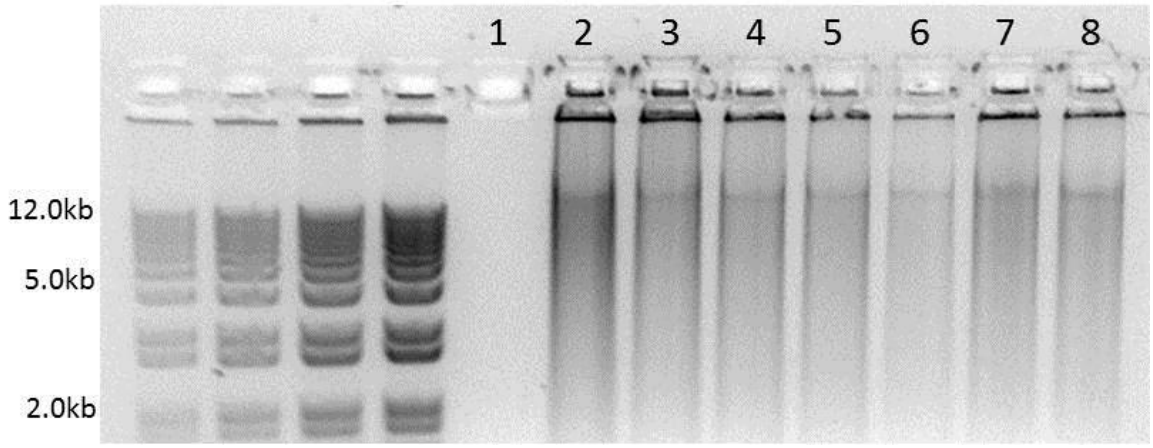


Figure 3.4 Quantification of DNA from MDA reactions using 1% agarose gel electrophoresis. The gel consists of DNA generated by MDA reactions on RB1A (Lanes 3-5) and RB1B (Lanes 6-8). Lane 1 represents the negative control, and lane 2 represents the positive control. The 1 kb Plus DNA marker (Invitrogen) used for quantification is in the four unlabeled lanes on the far left of the image for comparison. These ladder lanes have 50 ng, 100 ng, 200 ng, and 300 ng of DNA in them, from left to right. The same quantification process was used for the SIPf7-MDA samples.

After the completion of MDA, PCR products from all samples were run by DGGE to demonstrate bacterial community fingerprints for each sample, the differences in community composition between non-SIP and SIP-enriched samples, and to provide a basic measure of the bias introduced through MDA. Unique bacterial community fingerprints were observed between the bulk soil samples, DNA-SIP samples, and the cellulose degrading enrichment (Figure 3.5). The predominant bands in SIPf7 did not appear in the bulk soil samples, indicating the enrichment of cellulose degrading bacteria through DNA-SIP. The cellulose degrading enrichment possessed a few prominent bands indicating the sample had a very

high abundance of a select few organisms able to metabolize cellulose. This was expected from an enrichment culture where microorganisms able to metabolize the provided substrate (in this case cellulose) would outcompete all other organisms for the provided substrate. Previous studies using DGGE-generated bacterial community fingerprints, have demonstrated the enrichment cultures typically have a few very predominant community members that account for a high proportion of the community for this reason (Ueno *et al.*, 2001, Shiratori *et al.*, 2006).

DGGE was also used to demonstrate the representation bias introduced into the bacterial community's fingerprint as a result of MDA. These results demonstrate that some representation bias is introduced into SIPf7 samples following amplification (Figure 3.5 Lanes 5-7) because the bacterial fingerprints of the SIPf7-MDA samples are different from the SIPf7 sample fingerprint. The bias introduced after MDA on the bulk soil samples appears to be considerably less. The likely explanation for this is that the bulk soil samples contain higher levels of template DNA, and it has been previously shown that with decreasing template copy number, representation bias increases (Bergen *et al.*, 2005, Detter *et al.*, 2002, Neufeld *et al.*, 2008).

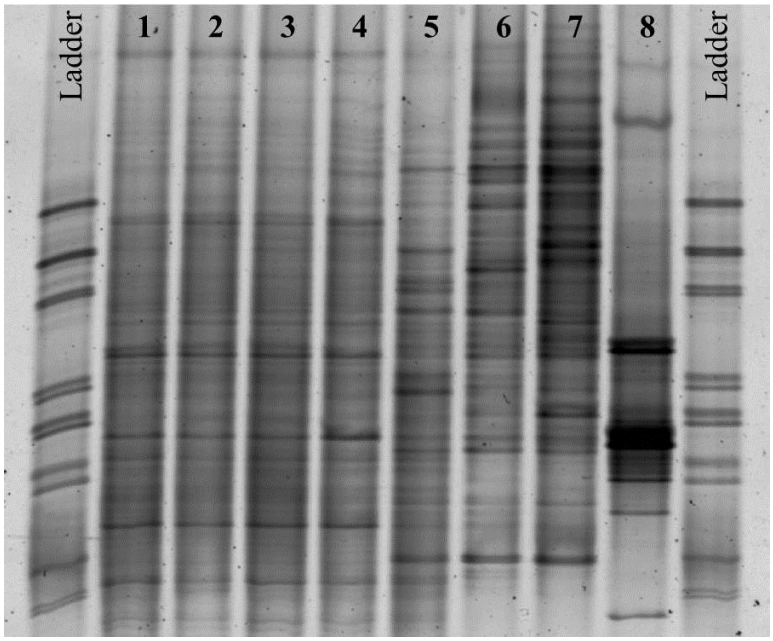


Figure 3.5 A 10% DGGE polyacrylamide gel with a 30-70% denaturing gradient containing the DNA fingerprint of all samples sent for sequencing. Lanes 1-8 represent the following samples: 1-RB1A, 2-RB1A-MDA, 3-RB1B, 4-RB1B-MDA, 5-SIPf7, 6-SIPf7-MDA1, 7-SIPf7-MDA2, 8-CDE. DGGE ladders of cloned PCR products are shown in the outermost lanes for comparison purposes.

3.3.4 Shot-gun metagenomics using Illumina and MG-RAST

Samples were sent to Dr. Stefan Green at the University of Illinois at Chicago where they were processed by the Nextera kit (Epicenter), making them Illumina compatible. Shot-gun metagenomic libraries were generated using Illumina paired-end sequencing, which resulted in a forward and reverse read for each sample. The forward and reverse Illumina reads were uploaded to MG-RAST individually to provide a duplicate measure of annotation.

Table 3.4 The number of sequences, mean G+C content, and MG-RAST accession numbers following the MG-RAST quality control pipeline for each library.

Sample	Sequences	G+C content	MG-RAST accession
RB1A (forward/reverse)	5,162,349/ 5,155,105	56 ± 10/ 55 ± 10	4474948.3/ 4474947.3
RB1B (forward/reverse)	7,321,404/ 7,330,334	56 ± 11/ 56 ± 10	4474945.3/ 4474944.3
RB1AMDA (forward/reverse)	9,994,629/ 10,298,687	45 ± 11/ 44 ± 11	4474980.3/ 4474985.3
RB1BMDA (forward/reverse)	9,839,797/ 10,129/910	44 ± 11/ 44 ± 11	4474984.3/ 4474946.3
SIPf7 (forward/reverse)	6,244,208/ 6,252,902	58 ± 12/ 58 ± 12	4474943.3/ 4474942.3
SIPf7MDA1 (forward/reverse)	12,439,712/ 12,610,469	37 ± 9/ 37 ± 8	4474941.2/ 4475899.3
SIPf7MDA2 (forward/reverse)	12,965,451/ 13,278,690	36 ± 8/ 36 ± 8	4474939.3/ 4474940.3
CDE (forward/reverse)	6,495,502/ 7,023,755	41 ± 8/ 41 ± 8	4474983.3/ 4474949.3

Functional annotations for each sample were performed using SwissProt, which is the manually annotated and reviewed version of Uniprot KnowledgeBase (www.uniprot.org). SwissProt is recognized for high-quality annotations, containing extremely well annotated protein sequences and specific links to detailed databases (Boeckmann *et al.*, 2003, Overbeek *et al.*, 2007). Although SwissProt had one of the lowest proportions of annotation hits per sequences analyzed (7-10% of sequences) (Figure 3.6), it was used for annotation because it is highly curated, provides high-quality predictions, and because it is cross-referenced with the Carbohydrate-Active enzymes (CAZy) database (www.cazy.org).

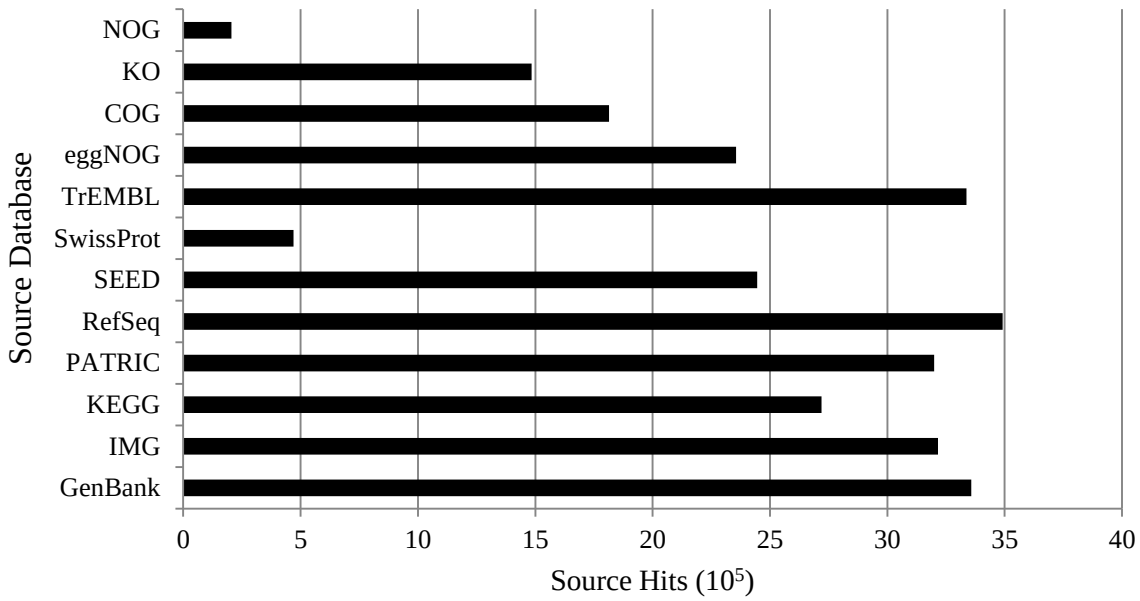


Figure 3.6 An example of the source hits distribution for an Illumina library using MG-RAST. The graph shown represents the source hits distribution for the SIPf7 forward read library. Although the specific number of hits changed for other libraries, the distribution pattern between databases remained similar. NOG: Non-supervised orthologous groups database. KO: Kyoto Encyclopedia of Genes and Genomes Orthology. COG: Clusters of orthologous groups. eggNOG: Evolutionary genealogy of genes- non-supervised orthologous group. TrEMBL: automatically annotated and non-reviewed UniProt protein database. SwissProt: manually annotated and reviewed UniProt protein database. SEED: The SEED project database. RefSeq: The National Center for Biotechnology Information reference sequences database. PATRIC: Pathosystems resource integration center. KEGG: Kyoto encyclopedia of genes and genomes. IMG: Integrated microbial genomes at the Joint Genome Institute. GenBank: National Institute of Health's Genetic sequence database.

The DNA-SIP sequence data revealed that the taxonomic distribution of the SIPf7 library was 90.9% bacteria, 7.1% eukaryota, 1.6% archaea, 0.2% viruses, and 0.2% other (see Figure 3.7). Of the five most abundant glycoside hydrolase gene sequences in the SIPf7 library, two were associated with cellulose degrading GH families (Table 3.5). Endoglucanase C was associated with *Cellvibrio* spp. (Gammaproteobacteria) and

endoglucanase Z was associated with *Dickeya* spp. (Gammeproteobacteria), and both enzymes were associated with GH 5. However, in the bulk soil libraries, none of the five most of the abundant GH enzyme sequences were associated with cellulases (Table 3.5). Similarly, none of the five most abundant GH enzymes sequences from the DNA-SIP MDA libraries were associated with cellulase GH families (Table 3.5). This suggests that MDA introduced a bias into the resulting libraries. Previous research characterizing the active consortia of cellulolytic organisms from Resolute Bay soil determined that the major cellulolytic organisms were found within the Gemmatimonadetes, Gammaproteobacteria, Acidobacteria, and Actinobacteria bacterial classes (Dunford, 2011).

Table 3.5 The five most abundant GH enzyme sequences for each library and their UniProt accession numbers following classification on MG-RAST. * indicates the enzyme sequence was only found in the forward read, and ** indicates the enzyme sequence was only found in the reverse read. Unmarked enzymes were found in both reads.

Sample	Abundant GH enzyme sequences (UniProt accession #)
SIPf7	α-galactosidase (Q9X4Y0), endoglucanase C (P27033), endoglucanase Z (P07103), xyloglucanase (Q3MUH7), α-L-arabinofuranosidase C* (P23031), glucoronoxylanase xynC** (Q45070)
SIPf7MDA1	β-glucanase (P45798), levanase (P05656), cytoplasmic α-amylase (P26612), β-galactosidase (Q8FKG6), β-hexosaminidase (P49008)
SIPf7MDA2	β-glucanase (P45798), levanase (P05656), β-galactosidase (Q8FKG6), β-hexosaminidase (P49008), α-N-acetylgalactosaminidase* (A4Q8F7), endo-1,4-β-xylanase Z** (P10478)
RB1A	α-N-arabinofuranosidase 2 (P94552), β-xylosidase (P23552), β-galactosidase (P26257), β-glucanase* (P45798), endo-1,4 β-xylanase Z* (P10478), levanse** (P05656), α-N-arabinofuranosidase 1** (P94531)
RB1B	α-N-arabinofuranosidase 2 (P94552), α-galactosidase (Q9X4Y0), β-galactosidase (P77989), β-galactosidase (P26257), glucoamylase* (P29761), 4-α-glucanotransferase** (O87172)
RB1AMDA	β-galactosidase (Q8FKG6), α-N-acetylgalactosaminidase (A4Q8F7), β-glucanase (P45798), α-N-arabinofuranosidase 2 (P94552), β-glucosidase** (P27034), levanase* (P05656)
RB1BMDA	β-glucosidase (P27034), α-N-acetylgalactosaminidase (A4Q8F7), levanase (P05656), β-glucanase* (P45798), β-hexosaminidase* (P49008), α-N-arabinofuranosidase 2** (P94552), α-N-acetylgalactosaminidase (A4Q8G1)
CDE	thermostable β-glucosidase B (P14002), endoglucanase 1 (Q02934), endoglucanase Z (P23659), exoglucanase 2 (P50900), endoglucanase A (P22534),

The CAZy database describes the families of enzymes that degrade, modify, or create glycosidic bonds, including glycoside hydrolases. Following the download of all enzymes annotated as glycoside hydrolases, each sequence was assigned to its CAZy family. The abundance of sequences affiliated with glycoside hydrolase enzymes and cellulases for each sample can be seen in Table 3.6.

Table 3.6 The number of enzymes annotated as glycoside hydrolases (GH), and cellulose degrading GH per 1 million sequences following the MG-RAST quality control pipeline (replicate 1/replicate 2). Cellulases were determined as enzymes belong to the GH5-12, 26, 44, 45 and 48 CAZy families. These families were previously referred to as cellulase families A-H, I, J, K, and L.

Sample	GH / million sequences	Cellulases / million sequences
RB1A (forward/reverse)	359/353	54/57
RB1B (forward/reverse)	326/308	53/43
RB1AMDA (forward/reverse)	263/231	38/36
RB1BMDA (forward/reverse)	238/234	36/36
SIPf7 (forward/reverse)	536/511	157/147
SIPf7MDA1 (forward/reverse)	294/293	61/55
SIPf7MDA2 (forward/reverse)	326/243	57/46
CDE (forward/reverse)	2649/2669	1461/1458

These data suggest that performing a DNA-SIP enrichment before the generation of metagenomic libraries increases the abundance of glycoside hydrolase enzymes found within the library. The results also demonstrate that the DNA-SIP enrichment resulted in a substantial increase in the relative abundance of cellulase enzymes in particular (Table 3.6). The data also indicate that MDA introduced a bias in SIP fraction 7 with regard to glycoside hydrolases, because the total number of glycoside hydrolases per million sequences decreased from ~525 in the SIPf7 library to ~290 in the SIP MDA libraries. This MDA-associated decrease in abundance was greater in cellulose degrading glycoside hydrolases; they were three times less abundant following MDA than they were in the SIPf7 library. The bias introduced by MDA on the bulk soil libraries was much less pronounced, which was expected based on the patterns observed with DGGE fingerprinting (see Figure 3.5).

Importantly, these results suggest that DGGE fingerprint similarity of MDA-amplified and non-MDA-amplified templates is a good measure of expected outcomes following sequence-based analysis of DNA. That said, also note that the proportion of glycoside hydrolases and cellulases decreased following MDA of bulk soil samples (Table 3.6), but by a smaller magnitude than in the DNA-SIP samples.

Prior to MDA, bacteria comprised ~91% of both the DNA-SIP and bulk soil Illumina libraries and ~7% of the libraries were comprised of eukaryota (Figure 3.7). However, the proportion of bacteria decreased to ~80% in the DNA-SIP MDA libraries and to ~ 85% in the bulk soil MDA libraries. The proportion of eukaryota rose to ~13% for both the DNA-SIP SIP MDA libraries and bulk soil MDA libraries (Figure 3.7). Previous research has shown that the MDA protocol used in my research (GenomiPhi kit) exhibited bias against high G+C content template DNA (Yilmas *et al*., 2010). A possible reason for the decreased proportion of bacteria following MDA is that most of the high G+C content template DNA was likely bacterial, and therefore was not represented in the post-MDA libraries. This was supported by the observation that the proportion of Actinobacteria, a bacterial class with high G+C content (Ventura *et al.*, 2007), decreased from ~17% in the DNA-SIP library to ~4% in the SIP-MDA libraries (data not shown). The mean G+C contents of pre- and post-MDA libraries further suggested that high G+C content template DNA was biased against (Table 3.4). It also suggested that greater bias was introduced with lower amounts of template DNA, supporting the observed DGGE community fingerprints. The functional category hits distribution was also determined for each Illumina library using the SEED Subsystem annotation on the MG-RAST server (Figure 3.8). Despite a change in GC content, the

distribution of functional categories changed very little between pre- and post-MDA libraries for both DNA-SIP and bulk soil samples (Figure 3.8). The lack of functional change following the application of MDA may be a result of poor annotation coverage of sequences in our libraries (~7-10%) or functional redundancy across diverse soil microbial genomes.

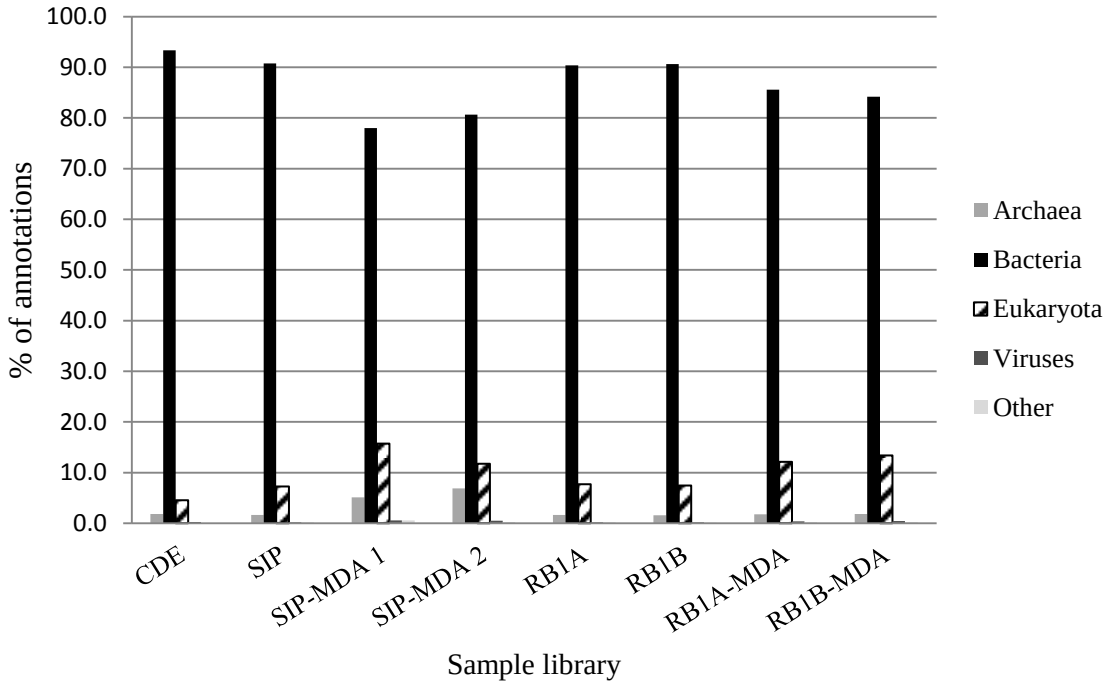
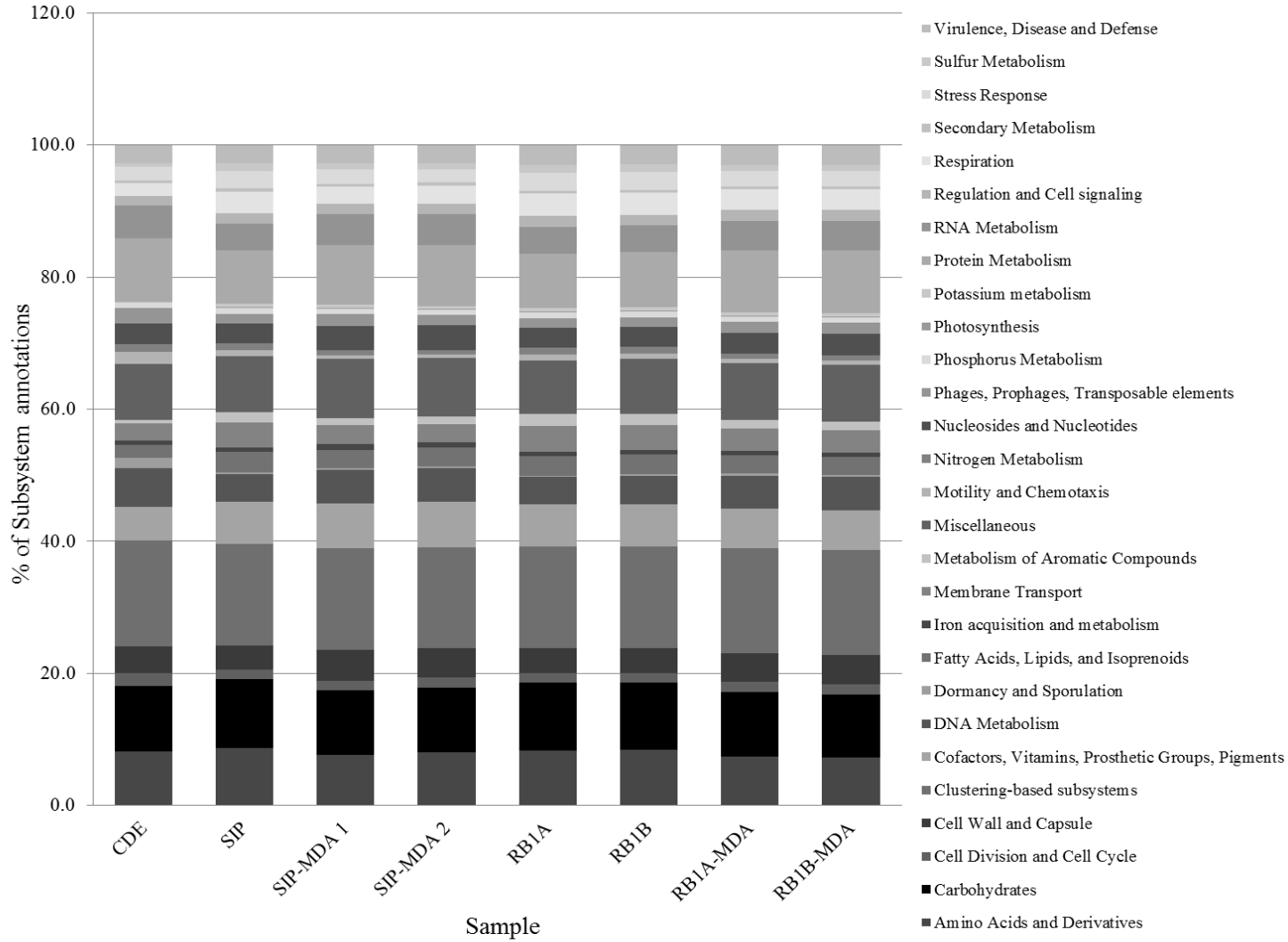


Figure 3.7 The taxonomic distributions of domains for each Illumina library following upload to MG-RAST. Annotations are from all of the databases MG-RAST accesses (GenBank, SwissProt, TrEMBL, SEED, KEGG, IMG, PATRIC, RefSeq). See Figure 3.6 for the complete names of each database.

Figure 3.8 The distribution of functional categories for each Illumina library. Annotation was performed using the SEED Subsystem annotation system.



The diversity of glycosyl hydrolases was measured using the Shannon index, and the proportion of glycoside hydrolase enzymes that made up the ten most abundant glycoside hydrolase families in each sample. The Shannon index is a measure of species diversity based on communication theory: the entropy associated with predicting the next letter in a message is represented by the Shannon function ('H') (Spellerberg & Fedor, 2003). In this research, the Shannon function represents the diversity (entropy) of glycoside hydrolases within each sample. The CDE library possessed very high proportions of glycoside hydrolases (Table 3.6), but with a very low diversity (Tables 3.7, 3.8), reflecting the low diversity observed in the DGGE fingerprints (Figure 3.5). This low diversity in the CDE library was observed using the ten abundant enzyme family counts, with the vast majority of all annotated enzymes falling into a few glycoside hydrolase families (Table 3.4, Appendix A). The Shannon index also demonstrated low glycoside hydrolase diversity for the CDE library, with Shannon values lower than (Table 3.8) those of the other libraries. Both the proportion of enzymes in the ten most abundant families and the Shannon index demonstrated higher diversity in the bulk soil, DNA-SIP, and MDA libraries (Table 3.7, 3.8) than in the CDE library. Though it would seem logical that SIPf7 should have lower diversity than the bulk soil, because diversity was measured within the glycoside hydrolase families only this was not the case. Diversity measures for all MDA samples were similar as well. Shannon index values are shown in Table 3.8, and represent the glycoside hydrolase diversity of each library.

Table 3.7 Proportion of the GH enzymes comprising the ten most abundant GH families for each sample (replicate 1/replicate 2).

Library	GH enzymes (%)
RB1A (forward/reverse)	57/58
RB1B (forward/reverse)	59/61
RB1AMDA (forward/reverse)	59/59
RB1BMDA (forward/reverse)	59/60
SIPf7 (forward/reverse)	59/61
SIPf7MDA1 (forward/reverse)	57/59
SIPf7MDA2 (forward/reverse)	60/61
CDE (forward/reverse)	88/88

Table 3.8 Shannon index values for each library (replicate 1/replicate 2). Values represent a measure of glycoside hydrolase diversity for each library.

Library	Shannon index
RB1A (forward/reverse)	3.48/3.42
RB1B (forward/reverse)	3.45/3.39
RB1AMDA (forward/reverse)	3.43/3.42
RB1BMDA (forward/reverse)	3.42/3.41
SIPf7 (forward/reverse)	3.45/3.41
SIPf7MDA1 (forward/reverse)	3.51/3.48
SIPf7MDA2 (forward/reverse)	3.42/3.40
CDE (forward/reverse)	2.55/2.54

Using the CAZy database, enzyme representation in sequence data was compared across the approximately 130 glycoside hydrolase families (www.cazy.org; January 2012). The cellulose-degrading enrichment had the highest proportion of glycoside hydrolases, and the highest proportion of cellulases (Table 3.6, Table 3.9). Within the CDE library,

approximately 55% of all glycoside hydrolases were cellulase enzymes (Table 3.9, Appendix A). The majority (~65%) of enzymes annotated as cellulases belonged to GH-9 and GH-48 (Table 3.9), which are glycoside hydrolase families consisting primarily of bacterial cellulases (Wilson and Urbanowisc, 2010). GH-9 is the second largest cellulase family, and is comprised largely of endocellulases, which randomly cut internal bonds creating oligosaccharides of various lengths from the polysaccharide cellulose chain (Lynd *et al.*, 2002). GH-48 is, at present, comprised solely of cellulases, and cellulase enzyme components. GH-48 cellulases occur in free enzymes systems, multi-enzyme systems, and in every cellulosome system known to date (Dassa, 2010). Both of these CAZy GH families contain cellulases with current industrial uses. The high proportion of cellulases, and low sample diversity supported what DGGE had demonstrated previously, and can be attributed to a few microorganisms with the ability to efficiently metabolize cellulose..

The DNA-SIP library contained the next highest proportion of cellulases (Table 3.6, Table 3.9), with approximately 30% of all glycoside hydrolases within the sample annotated as cellulases. The two most abundant cellulase families in the DNA-SIP library were GH-5 and GH-9 (Table 3.9), accounting for approximately 58% of all annotated cellulase enzymes. GH-5 is one of the largest of all glycoside hydrolase families, containing a wide variety of cellulases, including endoglucanases, exoglucanases, and β-glucosidases widely distributed across bacteria, archaea, fungi, and plants (Davies, 2011). Similar to GH-9 and GH-48, GH-5 also contains cellulase enzymes currently being used for industrial purposes. Other cellulase families of interest to industry, namely GH-6 and GH-48, represented approximately 10% of enzymes annotated as cellulases in the DNA-SIP library (Table 3.9).

Table 3.9 Number of sequences annotated as cellulases per one million sequences for each library (replicate 1/replicate 2).

CAZy Family	Known Activities	CDE (for/rev)	SIPf7 (for/rev)	SIPMDA1 (for/rev)	SIPMDA2 (for/rev)
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	270/271	63/60	24/22	20/18
GH6	Endoglucanase, cellobiohydrolase	1/1	5/6	3/2	3/2
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	0/0	0/1	0/0	0/0
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	17/18	6/6	3/2	3/2
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	601/596	31/28	9/7	8/7
GH10	Xylanase, β-1,3-endoxylanase	192/195	25/24	17/16	18/15
GH11	Xylanase	10/7	7/5	3/2	2/2
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	0/0	0/0	0/0	0/0
GH26	β-1,3-Xylanase, mannanase	5/3	3/4	2/2	3/2
GH44	Endoglucanase, xyloglucanase	3/5	11/10	2/1	0/1
GH45	Endoglucanase (mainly eukaryotic)	14/11	2/1	0/1	0/0
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	354/353	6/6	1/1	2/0

CAZy Family	Known Activities	RB1A (for/rev)	RB1AMDA (for/rev)	RB1B (for/rev)	RB1BMDA (for/rev)
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	12/17	12/11	18/14	12/12
GH6	Endoglucanase, cellobiohydrolase	3/3	1/2	3/2	2/2
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	1/0	1/0	1/0	0/1
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	2/3	3/2	2/2	2/1
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	11/9	6/6	9/8	5/5
GH10	Xylanase, β-1,3-endoxylanase	18/15	12/12	11/10	12/11
GH11	Xylanase	2/1	1/1	2/2	1/1
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	0/0	0/0	0/0	0/0
GH26	β-1,3-Xylanase, mannanase	1/1	3/2	1/2	1/2
GH44	Endoglucanase, xyloglucanase	1/3	1/1	4/2	1/1
GH45	Endoglucanase (mainly eukaryotic)	0/1	0/2	0/0	1/0
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	4/5	3/1	4/5	2/2

3.3.5 Bray-Curtis dissimilarity unrooted, neighbor-joining trees

Along with measuring the diversity of each library, the Bray-Curtis similarity coefficient was used to compare the compositional dissimilarity between sample libraries. An important aspect of the Bray-Curtis dissimilarity measure is that it takes into account the abundance representation of each entry, whereas several other dissimilarity measures (ie. Jaccard's) do not. This results in glycoside hydrolases families with large numbers of annotated enzymes having a greater effect on the similarity coefficient than families containing small numbers of enzymes. From the similarity coefficients for all glycoside hydrolase families, and cellulase families, unrooted neighbor-joining trees were generated using the APE platform in the program R (Saitou & Nei, 1987, Paradis *et al.*, 2004). These trees, which are shown in Figure 3.9, were constructed to demonstrate the similarity between all libraries. Branch lengths are inversely related to similarity; shorter branch lengths represent greater similarity.

The branch lengths between the CDE library and the closest related library (SIPf7) demonstrated that the CDE library was distinct from all other samples (Figure 3.9). The reason for the large dissimilarity is the lack of diversity in the enrichment library, coupled with large proportions of glycoside hydrolases and cellulases in relation to the other samples. The SIPf7 library was most closely related to the CDE library because it had the next highest proportion of glycoside hydrolases. It also had a higher proportion of glycoside hydrolases than the bulk soil samples, and a higher proportion of cellulases. The relationships between CDE, SIPf7, and bulk soil libraries were expected due to the proportion and types (GH

families) of sequences annotated as glycoside hydrolases in each library (Table 3.6, Table 3.9, Appendix A). As noted above, the application of MDA introduced representational bias into the resulting libraries. The MDA bias was greater for the SIP DNA samples than the bulk soil samples, and the neighbor-joining trees reflect this observation. Note that SIPf7 was more similar to the bulk soil samples than it was to the SIPf7-MDA libraries (Figure 3.9), for both total glycoside hydrolases and cellulases. Representation bias was less in the bulk soil libraries, as their MDA libraries had considerably shorter branch lengths, indicating a higher similarity. Representational bias introduced through the use of MDA has implications for research targeting cellulases with DNA-SIP and functional metagenomic studies. The screening or selection of the metagenomic clone library will be less effective, because of the decrease in the representation of glycoside hydrolases, and more specifically, cellulases. Unless the bias is reduced, it may cause researchers to do far more screening or selection to discover positive clones. To combat this problem, maximum amounts of DNA should be added to the density gradient prior to ultracentrifugation. Following fractionation, this should result in larger quantities of DNA being associated with each fraction, providing more template DNA for the MDA reaction, which should decrease the representation bias.

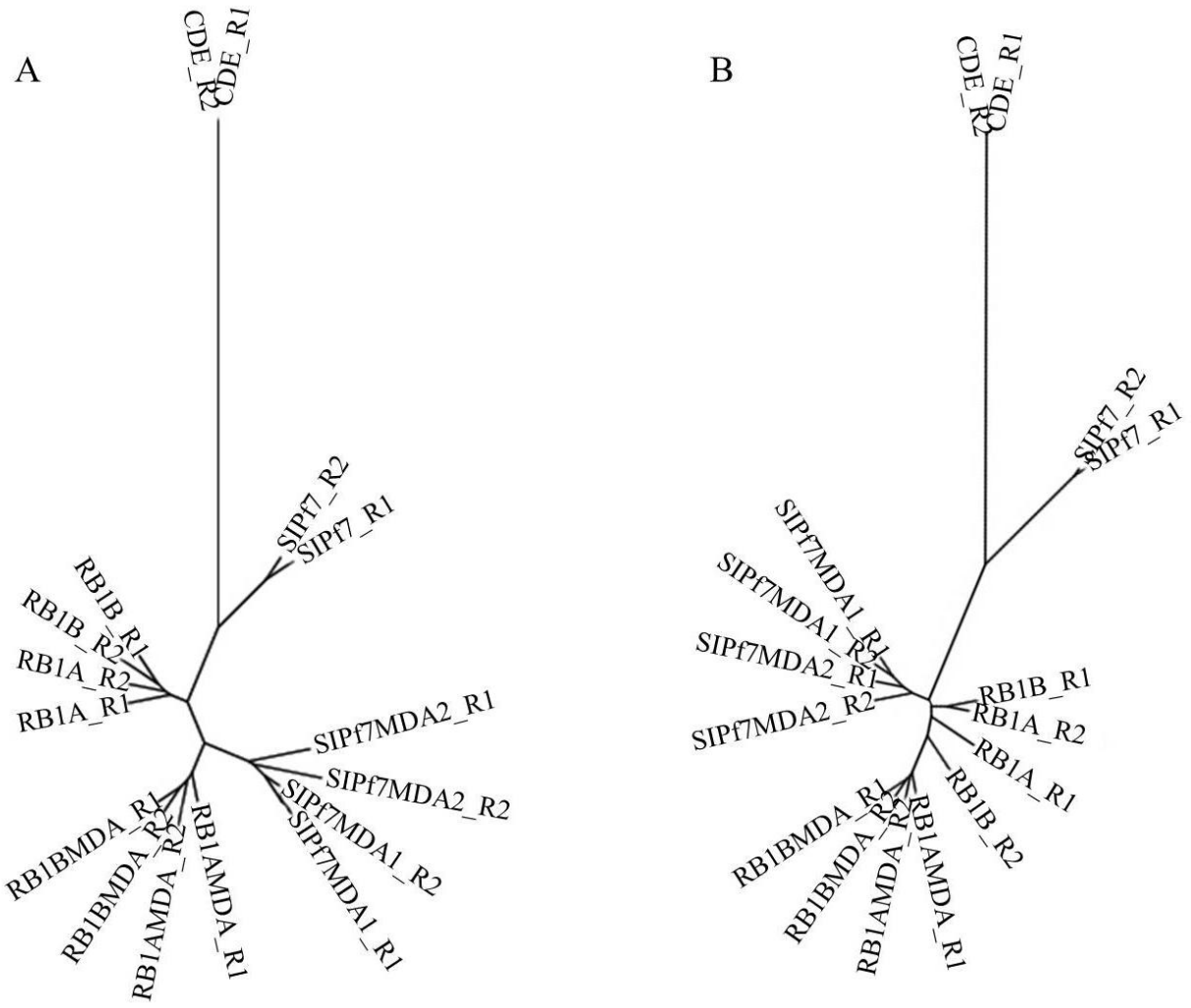


Figure 3.9 Unrooted neighbor-joining trees constructed using Bray-Curtis similarity coefficients in R. "A" represents all glycoside hydrolases and "B" represents glycoside hydrolases belonging to cellulose degrading families (GH5-12, 26, 44, 45, 48). R1 signifies the forward read and R2 signifies the reverse read.

Chapter 4 Conclusions and Future Considerations

4.1 Conclusions

Characterization of tundra glycoside hydrolases and the discovery of novel, cold-adapted enzymes has implications for our planet's biogeochemical cycling and our civilization's industrial processes. This study represents the first attempt to apply DNA-SIP and metagenomics to study glycoside hydrolases in Arctic tundra. This thesis described the optimization of a DNA isolation procedure to retrieve high-quality, high molecular weight DNA, and determined that DNA-SIP pre-enrichment increases the relative abundance of glycoside hydrolases in Arctic tundra soil metagenomic libraries, which is an important prerequisite to the discovery of cold-adapted, industrially relevant glycoside hydrolases.

Two of the most promising culture-independent approaches for linking taxonomy and metabolic activity are DNA-SIP and metagenomics. The combination of these two approaches has enormous potential for increasing understanding of our planet's biogeochemical cycling, and for industrial applications. Prior to this research, the coupling of these two methods had not yet been applied to the discovery of glycoside hydrolases, which are a group of enzyme families that have enormous industrial potential. Cellulases, enzymes capable or breaking down cellulose and its constituents are considered to be some of the most industrially relevant glycoside hydrolases (Bayer *et al.*, 2007, Wilson, 2009). Because cellulose is the most abundant organic compound on the planet (Lynd *et al.*, 2002, O'Sullivan *et al.*, 2007), cellulases are also major contributors to the global carbon cycle. With climate

change becoming a major research focus, furthering our knowledge of glycoside hydrolases and cellulase enzymes is of great importance. Metagenomic studies can expand upon our current knowledge of glycoside hydrolases by uncovering which enzymes are found in which environments, and metagenomic studies can also discover novel enzymes useful for industrial applications.

A problem facing metagenomic studies targeting cellulases is that many microorganisms harboring the desired genes and enzymes are not the most abundant organisms in the environment. The inclusion of a DNA-SIP enrichment before the construction of metagenomic libraries should increase the abundance of target genes or organisms. This research is the first to combine DNA-SIP and metagenomics to study active cellulose-degrading consortia within a tundra soil community, and is the first to apply these two techniques for the discovery of novel cellulases. The inclusion of a DNA-SIP pre-enrichment step to increase the abundance of sequences from organisms containing cellulases will allow for more efficient metagenomic research in the future.

The primary research conducted in this thesis builds upon work done by Eric Dunford, who characterized the active consortia of cellulolytic bacteria in an Arctic tundra soil using DNA-SIP (Dunford, 2011). His research was the first to characterize cellulose degrading soil bacteria in the Arctic using DNA-SIP. My research attempted to determine whether DNA-SIP was a suitable pre-enrichment step for increasing the abundance of cellulase genes in metagenomic libraries of Arctic tundra soil. I hypothesized that a DNA-SIP pre-enrichment step would increase the abundance of cellulases. The hypothesis was confirmed because the Illumina data for the SIPf7 sample resulted in 157 and 147 annotated

cellulases per 1 million sequences, in contrast to 54 and 57 for RB1A and 53 and 43 for RB1B (Table 3.3). Thus, there was a ~3 fold increase in abundance of cellulases following enrichment using DNA-SIP.

This research demonstrated that the application of MDA to DNA-SIP heavy fraction DNA resulted in an approximately 33% decrease the abundance of annotated cellulase genes (~50 per 1 million sequences), and a bias against high G+C content template DNA(see section 3.3.4 and Figure 3.8). Bias against high G+C content templates had been previously observed in the research of Yilmas *et al.* (2010). This representation bias could be detrimental to the application of DNA-SIP as a pre-enrichment for functional metagenomics. However, the addition of larger quantities of DNA to the CsCl SIP gradient and more DNA added to MDA reactions (e.g. >5 ng) should help circumvent this bias for future studies.

Previous research by Schwarz *et al.* (2006), targeting B12-dependent glycerol dehydratases in Wadden Sea sediment, found that the addition of a DNA-SIP enrichment before the construction of metagenomic libraries increased their gene detection frequencies by 2.1-3.8 fold over non DNA-SIP enriched sediment. Although the Schwarz research used ^{13}C-labelled glycerol as a substrate and utilized small-insert functional metagenomic screening, not sequence-based shotgun metagenomics, they demonstrated a similar increase in gene frequencies (~3 fold) to my research. Successful functional metagenomic screening of DNA-SIP enriched soil targeting cellulases has considerable potential to discover truly novel genes.

Preliminary research in this thesis attempted to optimize a DNA extraction and purification technique suitable for the production of large-insert (30-50 kb) functional

metagenomic libraries. I hypothesized that bead-beating would result in DNA that was too highly sheared because of the vigorous, mechanical nature of the extraction and that a gentler extraction method would be needed. Results generated from the three methods tested confirmed this hypothesis. Pulsed field gel electrophoresis demonstrated that bead-beating resulted in fragments that were too small (<20 kb) for the construction of large-insert metagenomic libraries. A more gentle extraction method - the high-salt buffer and SDS based modified Zhou procedure - was necessary and resulted in DNA in a suitable size range (>30 kb). Previous research (Riesenfeld *et al.*, 2004, Turnbaugh *et al.*, 2009, Lanzen *et al.*, 2011) has used bead-beating extractions for sequence-based, small-insert libraries but found that the method was unsuitable for large-insert libraries. That research was supported by the research presented in this thesis. Along with Neveu *et al.* (2011), Chen *et al.* (2008), and Sul *et al.* (2009), who used the Zhou extraction method for large-insert libraries, this research demonstrated that it is an effective method for extraction of high-molecular weight DNA suitable for the construction of cosmid, fosmid, or BAC metagenomic libraries.

4.2 Future Considerations

Microbial communities are extremely complex, making them difficult to study. The methods applied in this thesis represent strategies for analyzing subsets of communities, and for determining microorganisms and their genes important for specific community functions. Limitations of DNA-SIP as a method for identifying unknown organisms include a difficulty in characterizing microorganisms with long generation times, dilution of applied labeled

substrate with other environmental carbon sources, and substrate concentration requirements much greater than those found in natural environments (Neufeld *et al.*, 2007b). Researchers have developed variations of stable-isotope probing to circumvent these limitations. Phospholipid fatty acid stable-isotope probing (PFLA-SIP) can achieve characterization of organisms actively metabolizing substrates through the comparison of cell membrane lipids (Treonis *et al.*, 2004), which does not require long incubation times and is useful for the identification of groups with unique membrane structures. RNA-SIP is similar to DNA-SIP, but uses isotopically labeled RNA molecules. Similar to PFLA-SIP, the benefit of RNA-SIP is increased sensitivity.

An important consideration when using a DNA-SIP pre-enrichment is the determination of which heavy fraction to amplify using MDA. This study used fraction 7 because it provided higher amounts of template DNA for MDA reactions than fractions 5 and 6. However, the use of fractions 5 or 6 may have resulted in a higher proportion of cellulase genes sequences in the resulting metagenomic libraries because fraction 7 shared some bacterial community characteristics with the lighter fractions (Figure 3.3). It would be interesting to see if the use of fraction 5 or 6 would have resulted in an increase in the three fold increase seen in cellulase gene sequences. I would suggest future studies using a DNA-SIP pre-enrichment should attempt MDA amplification on all heavy fractions and, depending on the amount of bias introduced through MDA, use the most unique heavy fraction for metagenomic library construction.

Following stable-isotope probing, ideally this research would have included the screening of cosmid, fosmid or BAC functional metagenomic libraries as well. It would have

been very interesting to observe how the screening efficiencies of each library compared to the proportions of cellulases calculated using Illumina sequencing and functional annotation. However, with functional screening the possibility of having a fraction 7 sample without MDA applied is not possible. There would not be enough DNA to generate a library of adequate size. Observing whether the screening efficiency was better in the SIP-MDA library versus the bulk soil would be very interesting although DGGE and sequence data suggest that greater template concentrations must be used to minimize MDA bias (Table 3.3, Figure 3.4).

This research also provided evidence that DGGE is a useful method for indicating whether representation bias has been introduced by MDA. In Figure 3.5, the SIP-MDA samples, though sharing some bands, contained visually different bacterial fingerprints than SIPf7. This suggested that a considerable amount of representation bias had been introduced following MDA, which was confirmed followed Illumina sequencing and annotation. Future researchers using MDA to amplify SIP DNA should consider using DGGE to provide a preliminary measure of representation bias introduced by MDA. If DGGE indicates bias has been introduced it may be beneficial to perform new DNA-SIP and/or MDA experiments to reduce the bias before proceeding to sequencing or functional cloning and screening. The use of DGGE as an "early warning" method to assess MDA representational bias has the potential to save researchers from wasting time and resources by proceeding with further experiments on samples that have been biased through MDA reactions.

With relation to a DNA-SIP and functional metagenomic research, I believe a study similar to the research by Schwarz *et al.* (2006) could provide very interesting results, and hopefully provide numerous truly novel cellulase genes. Schwarz *et al.* (2006) incubated

marine sediment with ^{13}C-labelled glycerol as a DNA-SIP enrichment prior to using functional metagenomics. Following DNA-SIP fractionation they used the heavy and light DNA to generate metagenomic libraries. Using colony hybridization they found 24 positive clones in the heavy DNA library and 9 positive clones in the light DNA library.

I believe a study employing a ^{13}C-cellulose DNA-SIP enrichment on tundra soil prior to functional metagenomic screening would enhance gene detection capabilities for cellulases. MDA would be necessary to increase the amount of DNA to a level suitable for functional library generation, which could hinder the success rate of the functional screening due to representation bias (Table 3.4, Table 3.7, Figure 3.5, Appendix A). A possible way to circumvent this bias would be to pool the heavy DNA-SIP fractions prior to MDA, thereby increasing the amount of template DNA. Following MDA, cosmid, fosmid, or BAC librarys could be constructed followed by Congo Red overlay to screen for cellulases. I believe this future research would not only provide interesting results, but also discover many industrial relevant cellulase genes.

An important next step will be to conduct DNA-SIP and metagenomics investigations with the inclusion of different types of soils. Canadian MetaMicroBiome Library research currently being undertaken should provide this information from many different soil ecosystem samples from across Canada (Neufeld *et al*., 2011). This project uses multiple ^{13}C-labeled carbohydrate substrates, which should result in the characterization and discovery of numerous novel carbohydrate-active enzymes.

The results presented in this thesis advance the understanding of the effect of DNA-SIP on metagenomic libraries, as well as the effect of MDA. The research presented here represents a proof-of-principle experiment for the targeted enrichment of cellulases using DNA stable-isotope probing and metagenomics from tundra soil. The data collected during this experiment provides insight into the usefulness of DNA-SIP as an enrichment step for metagenomics, and will strengthen the application of this type of research for both environmental and industrial applications.

Appendix A

Glycoside Hydrolase Counts for Illumina Libraries

The following tables show the total number of sequences annotated as glycoside hydrolases for each Illumina library (replicate 1/replicate 2). Note that a combined table of all GH families and sequence representation has been prepared in MS Excel and included on the attached CD.

CAZy Family	Known Activities	SIPf7 (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	193/196
GH2	β-Galactosidase, β-mannosidase, others	150/154
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, others	333/317
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	78/73
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	392/373
GH6	Endoglucanase, cellobiohydrolase	33/40
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	1/4
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	40/39
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	194/178
GH10	Xylanase, β-1,3-endoxylanase	156/150
GH11	Xylanase	44/34
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	2/2
GH13	α-Amylase, catalytic domain, and related enzymes	263/281
GH14	β-amylase	3/3
GH15	glucoamylase, glucodextranase, α-trehalase	20/12
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	54/51
GH17	licheninase, glucan β-1,3-glucosidase	13/10
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	73/71
GH19	chitinase, lysozyme	9/7
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	21/23
GH22	lysozyme type c and i, α-lactalbumin	1/1
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	52/41
GH26	β-1,3-Xylanase, mannanase	18/24
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	57/69
GH28	Polygalacturonase, rhamnogalacturonase, others	23/25
GH29	α-L-fucosidase	24/31
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	59/63
GH31	α-Glucosidase, α-xylosidase, others	71/51
GH32	invertase, endo-inulinase, endo-levanase, others	48/47
GH34	sialidase	1/1
GH35	β-Galactosidase	36/36

Con't	Known Activities	SIPf7 (forward/reverse)
GH38	α-Mannosidase	27/22
GH39	β-Xylosidase, α-L-iduronidase	25/22
GH42	β-Galactosidase	32/24
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	34/35
GH44	Endoglucanase, xyloglucanase	68/60
GH45	Endoglucanase (mainly eukaryotic)	10/6
GH46	chitosanase	8/7
GH47	α-Mannosidase	12/16
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	36/39
GH50	β-agarase	2/2
GH51	Endoglucanase, α-L-arabinofuranosidase	117/134
GH52	β-Xylosidase	3/8
GH53	β-1,4-Endogalactanase	16/12
GH54	α-L-arabinofuranosidase, β-xylosidase	9/4
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	2/0
GH56	hyaluronidase	5/3
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	56/50
GH58	Endo-N-acetylneuraminidase or endo-sialidase	1/0
GH59	galactocerebrosidase	2/2
GH61	endoglucanase (needs others cellulases)	2/0
GH62	α-L-arabinofuranosidase	56/44
GH63	α-glucosidase, α-1,3-glucosidase	1/2
GH64	β-1,3-glucanase	1/0
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	13/14
GH66	dextranase	1/0
GH68	β-fructofuranosidase, others	6/1
GH71	α-1,3-glucanase	1/0
GH73	peptidoglycan hydrolase	1/4
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	115/93
GH77	4-α-Glucanotransferase, amylomaltase	77/51
GH79	β-glucuronidase, heparanase, others	2/0
GH81	endo-β-1,3-glucanase	4/5
GH83	neuraminidase	3/1
GH85	endo-β-N-acetylglucosaminidase	2/1
GH95	α-L-Fucosidase	9/11
GH96	α-agarase	9/8
GH99	endo-α-1,2-mannosidase	2/2
GH109	α-N-Acetylgalactosaminidase	21/12
GH110	α-galactosidase, α-1,3-galactosidase	6/3
GH116	acid β-glucosidase, β-glucosidase, β-xylosidase	5/9

CAZy Family	Known Activities	SIPMDA1 (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	142/187
GH2	β-Galactosidase, β-mannosidase, others	245/262
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	300/305
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	16/21
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	304/280
GH6	Endoglucanase, cellobiohydrolase	32/23
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	3/5
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	35/22
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	115/88
GH10	Xylanase, β-1,3-endoxylanase	206/208
GH11	Xylanase	40/31
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	0/1
GH13	α-Amylase, catalytic domain, and related enzymes	266/282
GH14	β-amylase	13/13
GH15	glucoamylase, glucodextranase, α-trehalase	20/13
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	124/127
GH17	licheninase, glucan β-1,3-glucosidase	17/8
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	130/139
GH19	chitinase, lysozyme	10/9
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	71/70
GH22	lysozyme type c and i, α-lactalbumin	6/6
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	71/88
GH26	β-1,3-Xylanase, mannanase	28/29
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	62/81
GH28	Polygalacturonase, rhamnogalacturonase, others	21/25
GH29	α-L-fucosidase	75/68
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	32/29
GH31	α-Glucosidase, α-xylosidase, others	118/97
GH32	invertase, endo-inulinase, endo-levanase, others	129/130
GH33	sialidase, trans-sialidase, others	35/29
GH34	sialidase	5/3
GH35	β-Galactosidase	45/34
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	3/3
GH37	α,α-trehalase	236/261
GH38	α-Mannosidase	34/30
GH39	β-Xylosidase, α-L-iduronidase	30/19
GH42	β-Galactosidase	15/12
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	83/73
GH44	Endoglucanase, xyloglucanase	19/15
GH45	Endoglucanase (mainly eukaryotic)	5/10

Con't	Known Activities	SIPf7MDA1 (forward/reverse)
GH47	α-Mannosidase	36/38
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	11/15
GH49	dextranase, isopullulanase, others	1/0
GH50	β-agarase	1/1
GH51	Endoglucanase, α-L-arabinofuranosidase	73/81
GH52	β-Xylosidase	1/1
GH53	β-1,4-Endogalactanase	34/30
GH54	α-L-arabinofuranosidase, β-xylosidase	8/5
GH56	hyaluronidase	5/7
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	42/39
GH58	Endo-N-acetylneuraminidase or endo-sialidase	1/0
GH59	galactocerebrosidase	0/3
GH61	endoglucanase (needs others cellulases)	1/4
GH62	α-L-arabinofuranosidase	30/21
GH63	α-glucosidase, α-1,3-glucosidase	2/4
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	27/26
GH66	dextranase	0/1
GH68	β-fructofuranosidase, others	5/3
GH73	peptidoglycan hydrolase	11/77
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	57/65
GH76	α-1,6-mannanase	1/1
GH77	4-α-Glucanotransferase, amylomaltase	45/55
GH79	β-glucuronidase, heparanase, others	4/0
GH81	endo-β-1,3-glucanase	9/9
GH83	neuraminidase	1/2
GH85	endo-β-N-acetylglucosaminidase	1/1
GH89	α-N-Acetylgalactosaminidase	1/1
GH95	α-L-Fucosidase	7/7
GH96	α-agarase	10/19
GH99	endo-α-1,2-mannosidase	1/3
GH102	Peptidoglycan lytic transglycosylase	1/0
GH103	Peptidoglycan lytic transglycosylase	1/2
GH109	α-N-Acetylgalactosaminidase	79/90
GH110	α-galactosidase, α-1,3-galactosidase	12/11
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	6/7

CAZy Family	Known Activities	SIPMDA2 (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	174/112
GH2	β-Galactosidase, β-mannosidase, others	297/255
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	411/332
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	5/4
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	260/234
GH6	Endoglucanase, cellobiohydrolase	41/28
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	3/3
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	34/29
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	108/94
GH10	Xylanase, β-1,3-endoxylanase	239/193
GH11	Xylanase	22/21
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	2/0
GH13	α-Amylase, catalytic domain, and related enzymes	331/229
GH14	β-amylase	13/7
GH15	glucoamylase, glucodextranase, α-trehalase	12/13
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	166/129
GH17	licheninase, glucan β-1,3-glucosidase	10/12
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	95/80
GH19	chitinase, lysozyme	8/3
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	126/75
GH22	lysozyme type c and i, α-lactalbumin	7/4
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	28/28
GH24	lysozyme	0/2
GH25	Lysozyme	0/1
GH26	β-1,3-Xylanase, mannanase	39/22
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	91/76
GH28	Polygalacturonase, rhamnogalacturonase, others	42/19
GH29	α-L-fucosidase	92/61
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	21/21
GH31	α-Glucosidase, α-xylosidase, others	82/51
GH32	invertase, endo-inulinase, endo-levanase, others	143/100
GH33	sialidase, trans-sialidase, others	23/20
GH34	sialidase	10/6
GH35	β-Galactosidase	44/30
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	4/4
GH37	α,α-trehalase	399/281
GH38	α-Mannosidase	90/79
GH39	β-Xylosidase, α-L-iduronidase	24/11
GH42	β-Galactosidase	15/12
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase	64/62

Con't	Known Activities	SIPMDA2 (forward/reverse)
GH44	Endoglucanase, xyloglucanase	4/8
GH45	Endoglucanase (mainly eukaryotic)	5/6
GH46	chitosanase	0/1
GH47	α-Mannosidase	47/32
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	23/3
GH50	β-agarase	1/1
GH51	Endoglucanase, α-L-arabinofuranosidase	85/67
GH52	β-Xylosidase	2/2
GH53	β-1,4-Endogalactanase	33/25
GH54	α-L-arabinofuranosidase, β-xylosidase	6/12
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	4/1
GH56	hyaluronidase	1/0
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	43/22
GH58	Endo-N-acetylneuraminidase or endo-sialidase	4/3
GH59	galactocerebrosidase	2/1
GH62	α-L-arabinofuranosidase	15/15
GH64	β-1,3-glucanase	1/0
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	38/23
GH66	dextranase	0/1
GH68	β-fructofuranosidase, others	1/0
GH71	α-1,3-glucanase	1/0
GH73	peptidoglycan hydrolase	15/13
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	47/57
GH76	α-1,6-mannanase	0/1
GH77	4-α-Glucanotransferase, amylomaltase	61/56
GH79	β-glucuronidase, heparanase, others	2/2
GH81	endo-β-1,3-glucanase	9/7
GH83	neuraminidase	4/3
GH85	endo-β-N-acetylglucosaminidase	7/2
GH89	α-N-Acetylgalactosaminidase	2/0
GH95	α-L-Fucosidase	8/6
GH96	α-agarase	9/7
GH109	α-N-Acetylgalactosaminidase	137/84
GH110	α-galactosidase, α-1,3-galactosidase	44/37
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	0/2

CAZy Family	Known Activities	CDE (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	183/166
GH2	β-Galactosidase, β-mannosidase, others	195/207
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	1272/1264
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	58/98
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	1756/1902
GH6	Endoglucanase, cellobiohydrolase	6/10
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	113/124
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	3903/4188
GH10	Xylanase, β-1,3-endoxylanase	1246/1370
GH11	Xylanase	62/51
GH13	α-Amylase, catalytic domain, and related enzymes	1743/2239
GH14	β-amylase	5/11
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	343/361
GH17	licheninase, glucan β-1,3-glucosidase	0/1
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	91/113
GH19	chitinase, lysozyme	2/1
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	18/23
GH24	lysozyme	29/32
GH25	Lysozyme	25/30
GH26	β-1,3-Xylanase, mannanase	31/22
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	3/2
GH28	Polygalacturonase, rhamnogalacturonase, others	10/18
GH29	α-L-fucosidase	3/8
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	54/50
GH31	α-Glucosidase, α-xylosidase, others	82/77
GH32	invertase, endo-inulinase, endo-levanase, others	218/240
GH33	sialidase, trans-sialidase, others	6/3
GH34	sialidase	14/14
GH35	β-Galactosidase	82/68
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	28/22
GH38	α-Mannosidase	119/113
GH39	β-Xylosidase, α-L-iduronidase	377/392
GH42	β-Galactosidase	115/141
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	10/10
GH44	Endoglucanase, xyloglucanase	19/38
GH45	Endoglucanase (mainly eukaryotic)	89/78
GH47	α-Mannosidase	2/6
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	2299/2481
GH52	β-Xylosidase	0/2

Con't	Known Activities	CDE (forward/reverse)
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	5/7
GH66	dextranase	333/385
GH68	β-fructofuranosidase, others	1/0
GH73	peptidoglycan hydrolase	5/9
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	371/382
GH77	4-α-Glucanotransferase, amylomaltase	82/70
GH95	α-L-Fucosidase	6/7
GH109	α-N-Acetylgalactosaminidase	1/0
GH110	α-galactosidase, α-1,3-galactosidase	3/1

CAZy Family	Known Activities	RB1A (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	101/90
GH2	β-Galactosidase, β-mannosidase, others	103/119
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	244/221
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	29/25
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	63/86
GH6	Endoglucanase, cellobiohydrolase	15/18
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	6/1
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	8/14
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	56/47
GH10	Xylanase, β-1,3-endoxylanase	91/75
GH11	Xylanase	12/6
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	1/0
GH13	α-Amylase, catalytic domain, and related enzymes	167/208
GH14	β-amylase	4/5
GH15	glucoamylase, glucodextranase, α-trehalase	12/17
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	27/16
GH17	licheninase, glucan β-1,3-glucosidase	19/5
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	39/49
GH19	chitinase, lysozyme	4/3
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	29/35
GH22	lysozyme type c and i, α-lactalbumin	2/3
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	51/51
GH24	lysozyme	11/5
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	50/41
GH28	Polygalacturonase, rhamnogalacturonase, others	20/16
GH29	α-L-fucosidase	21/30
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	9/2
GH31	α-Glucosidase, α-xylosidase, others	58/44
GH32	invertase, endo-inulinase, endo-levanase, others	31/39
GH33	sialidase, trans-sialidase, others	13/11
GH34	sialidase	26/1
GH35	β-Galactosidase	5/24
GH37	α,α-trehalase	62/52
GH38	α-Mannosidase	12/10
GH39	β-Xylosidase, α-L-iduronidase	40/36
GH42	β-Galactosidase	17/13
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	20/22
GH44	Endoglucanase, xyloglucanase	7/17
GH45	Endoglucanase (mainly eukaryotic)	0/3
GH47	α-Mannosidase	21/20

Con't	Known Activities	RB1A (forward/reverse)
GH49	dextranase, isopullulanase, others	0/1
GH50	β-agarase	4/2
GH51	Endoglucanase, α-L-arabinofuranosidase	107/110
GH52	β-Xylosidase	1/1
GH53	β-1,4-Endogalactanase	10/8
GH54	α-L-arabinofuranosidase, β-xylosidase	10/11
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	1/0
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	23/31
GH58	Endo-N-acetylneuraminidase or endo-sialidase	2/1
GH61	endoglucanase (needs others cellulases)	2/0
GH62	α-L-arabinofuranosidase	10/2
GH63	α-glucosidase, α-1,3-glucosidase	1/2
GH64	β-1,3-glucanase	0/3
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	2/1
GH68	β-fructofuranosidase, others	1/2
GH73	peptidoglycan hydrolase	3/4
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	32/29
GH76	α-1,6-mannanase	3/0
GH77	4-α-Glucanotransferase, amylomaltase	58/52
GH81	endo-β-1,3-glucanase	4/1
GH83	neuraminidase	1/4
GH89	α-N-Acetylgalactosaminidase	0/1
GH95	α-L-Fucosidase	5/5
GH96	α-agarase	4/6
GH99	endo-α-1,2-mannosidase	2/0
GH103	Peptidoglycan lytic transglycosylase	0/1
GH109	α-N-Acetylgalactosaminidase	21/26
GH110	α-galactosidase, α-1,3-galactosidase	3/3
GH111	keratan sulphate hydrolase	1/0
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	0/2

CAZy Family	Known Activities	RB1AMDA (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	147/129
GH2	β-Galactosidase, β-mannosidase, others	168/159
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	368/292
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	22/22
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	121/110
GH6	Endoglucanase, cellobiohydrolase	14/16
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	5/4
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	27/16
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	57/63
GH10	Xylanase, β-1,3-endoxylanase	117/119
GH11	Xylanase	5/8
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	1/2
GH13	α-Amylase, catalytic domain, and related enzymes	216/231
GH14	β-amylase	7/4
GH15	glucoamylase, glucodextranase, α-trehalase	13/10
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	60/57
GH17	licheninase, glucan β-1,3-glucosidase	21/16
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	64/49
GH19	chitinase, lysozyme	8/12
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	41/50
GH22	lysozyme type c and i, α-lactalbumin	3/3
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	46/38
GH24	lysozyme	6/11
GH25	Lysozyme	5/4
GH26	β-1,3-Xylanase, mannanase	31/16
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	81/94
GH28	Polygalacturonase, rhamnogalacturonase, others	20/30
GH29	α-L-fucosidase	58/34
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	7/3
GH31	α-Glucosidase, α-xylosidase, others	91/82
GH32	invertase, endo-inulinase, endo-levanase, others	86/65
GH33	sialidase, trans-sialidase, others	23/28
GH34	sialidase	3/2
GH35	β-Galactosidase	26/27
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	2/9
GH37	α,α-trehalase	118/107
GH38	α-Mannosidase	11/4
GH39	β-Xylosidase, α-L-iduronidase	36/40
GH42	β-Galactosidase	16/14

Con't	Known Activities	RB1AMDA (forward/reverse)
GH44	Endoglucanase, xyloglucanase	13/8
GH45	Endoglucanase (mainly eukaryotic)	1/1
GH46	chitosanase	1/0
GH47	α-Mannosidase	34/27
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	25/24
GH51	Endoglucanase, α-L-arabinofuranosidase	110/93
GH53	β-1,4-Endogalactanase	15/9
GH54	α-L-arabinofuranosidase, β-xylosidase	4/3
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	2/0
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	21/18
GH58	Endo-N-acetylneuraminidase or endo-sialidase	2/1
GH61	endoglucanase (needs others cellulases)	9/0
GH62	α-L-arabinofuranosidase	2/4
GH63	α-glucosidase, α-1,3-glucosidase	5/3
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	11/6
GH66	dextranase	2/3
GH68	β-fructofuranosidase, others	1/1
GH71	α-1,3-glucanase	0/1
GH73	peptidoglycan hydrolase	5/2
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	31/27
GH77	4-α-Glucanotransferase, amylomaltase	46/42
GH81	endo-β-1,3-glucanase	3/2
GH83	neuraminidase	2/2
GH84	hyaluronidase, N-acetyl β-glucosaminidase	1/0
GH85	endo-β-N-acetylglucosaminidase	0/1
GH89	α-N-Acetylgalactosaminidase	0/1
GH95	α-L-Fucosidase	4/1
GH96	α-agarase	8/10
GH99	endo-α-1,2-mannosidase	1/0
GH109	α-N-Acetylgalactosaminidase	79/77
GH110	α-galactosidase, α-1,3-galactosidase	4/2
GH112	β-mannanase	0/3
GH115	acid β-glucosidase, β-glucosidase, β-xylosidase	2/0
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	1/4

CAZy Family	Known Activities	RB1B (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	121/109
GH2	β-Galactosidase, β-mannosidase, others	128/170
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	314/295
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	44/40
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	130/100
GH6	Endoglucanase, cellobiohydrolase	24/11
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	5/2
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	11/11
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	66/62
GH10	Xylanase, β-1,3-endoxylanase	84/73
GH11	Xylanase	12/15
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	1/0
GH13	α-Amylase, catalytic domain, and related enzymes	262/296
GH14	β-amylase	7/6
GH15	glucoamylase, glucodextranase, α-trehalase	24/28
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	28/30
GH17	licheninase, glucan β-1,3-glucosidase	12/10
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	55/55
GH19	chitinase, lysozyme	7/9
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	42/24
GH22	lysozyme type c and i, α-lactalbumin	4/2
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	80/60
GH24	lysozyme	19/10
GH25	Lysozyme	1/1
GH26	β-1,3-Xylanase, mannanase	9/12
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	52/47
GH28	Polygalacturonase, rhamnogalacturonase, others	29/28
GH29	α-L-fucosidase	30/18
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	11/7
GH31	α-Glucosidase, α-xylosidase, others	79/59
GH32	invertase, endo-inulinase, endo-levanase, others	36/38
GH33	sialidase, trans-sialidase, others	11/17
GH34	sialidase	0/2
GH35	β-Galactosidase	21/20
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	4/5
GH37	α,α-trehalase	56/57
GH38	α-Mannosidase	15/16
GH39	β-Xylosidase, α-L-iduronidase	25/36
GH42	β-Galactosidase	17/14
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	24/20

Con't	Known Activities	RB1B (forward/reverse)
GH45	Endoglucanase (mainly eukaryotic)	2/0
GH46	chitosanase	2/4
GH47	α-Mannosidase	21/26
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	27/34
GH50	β-agarase	0/4
GH51	Endoglucanase, α-L-arabinofuranosidase	125/115
GH52	β-Xylosidase	1/0
GH53	β-1,4-Endogalactanase	10/7
GH54	α-L-arabinofuranosidase, β-xylosidase	15/17
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	1/2
GH56	hyaluronidase	0/2
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	41/29
GH58	Endo-N-acetylneuraminidase or endo-sialidase	1/0
GH61	endoglucanase (needs others cellulases)	8/4
GH62	α-L-arabinofuranosidase	5/4
GH63	α-glucosidase, α-1,3-glucosidase	4/3
GH64	β-1,3-glucanase	2/0
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	4/3
GH66	dextranase	4/4
GH68	β-fructofuranosidase, others	1/2
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	50/32
GH76	α-1,6-mannanase	2/1
GH77	4-α-Glucanotransferase, amylomaltase	81/89
GH79	β-glucuronidase, heparanase, others	0/1
GH81	endo-β-1,3-glucanase	1/1
GH82	I-carrageenase	0/1
GH83	neuraminidase	2/1
GH85	endo-β-N-acetylglucosaminidase	4/0
GH88	delta-4,5 Unsaturated b-glucuronyl hydrolase	0/1
GH89	α-N-Acetylgalactosaminidase	2/0
GH95	α-L-Fucosidase	7/8
GH96	α-agarase	8/12
GH99	endo-α-1,2-mannosidase	1/1
GH109	α-N-Acetylgalactosaminidase	22/22
GH110	α-galactosidase, α-1,3-galactosidase	2/2
GH113	endo-α-1,4-polygalactosaminidase	1/0
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	4/3

CAZy Family	Known Activities	RB1BMDA (forward/reverse)
GH1	β-Glucosidase, β-galactosidase, β-mannosidase, others	131/133
GH2	β-Galactosidase, β-mannosidase, others	125/105
GH3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-laraβinofuranosidase, others	349/342
GH4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	16/15
GH5	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	115/120
GH6	Endoglucanase, cellobiohydrolase	15/17
GH7	β-1,4-endoglucanase,cellobiohydrolase, chitosanase	0/6
GH8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	19/11
GH9	Endoglucanase, cellobiohydrolase, β-glucosidase	50/49
GH10	Xylanase, β-1,3-endoxylanase	116/116
GH11	Xylanase	5/9
GH12	Endoglucanase, β-1,3-1,4-glucanase, others	0/1
GH13	α-Amylase, catalytic domain, and related enzymes	197/237
GH14	β-amylase	6/8
GH15	glucoamylase, glucodextranase, α-trehalase	14/15
GH16	β-1,3-Endoglucanase, β-agarase, xyloglucanase, β-1,3-galactanase, others	41/38
GH17	licheninase, glucan β-1,3-glucosidase	12/17
GH18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	44/56
GH19	chitinase, lysozyme	8/11
GH20	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	41/40
GH22	lysozyme type c and i, α-lactalbumin	7/8
GH23	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	37/48
GH24	lysozyme	16/16
GH25	Lysozyme	8/1
GH26	β-1,3-Xylanase, mannanase	9/25
GH27	α-Galactosidase, α-N-acetylgalactosaminidase, isomalto-dextranase	62/63
GH28	Polygalacturonase, rhamnogalacturonase, others	26/22
GH29	α-L-fucosidase	49/46
GH30	β-1,6-glucanase, β-xylosidase, β-glucosidase, β-fucosidase, otheers	1/3
GH31	α-Glucosidase, α-xylosidase, others	83/81
GH32	invertase, endo-inulinase, endo-levanase, others	67/62
GH33	sialidase, trans-sialidase, others	19/16
GH34	sialidase	2/3
GH35	β-Galactosidase	23/23
GH36	α-Galactosidase, α-N-acetylgalactosaminidase	8/8
GH37	α,α-trehalase	117/113
GH38	α-Mannosidase	20/11
GH39	β-Xylosidase, α-L-iduronidase	33/25
GH42	β-Galactosidase	9/7
GH43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	30/19

Con't	Known Activities	RB1B (forward/reverse)
GH44	Endoglucanase, xyloglucanase	13/14
GH45	Endoglucanase (mainly eukaryotic)	6/4
GH46	chitosanase	2/2
GH47	α-Mannosidase	36/32
GH48	endoglucanase, chitinase, cellobiohydrolase, endo-cellulases	19/21
GH51	Endoglucanase, α-L-arabinofuranosidase	94/103
GH52	β-Xylosidase	0/1
GH53	β-1,4-Endogalactanase	9/6
GH54	α-L-arabinofuranosidase, β-xylosidase	4/14
GH55	exo-β-1,3-glucanase, endo-β-1,3-glucanase	1/1
GH57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	23/27
GH58	Endo-N-acetylneuraminidase or endo-sialidase	3/1
GH61	endoglucanase (needs others cellulases)	0/2
GH62	α-L-arabinofuranosidase	1/2
GH63	α-glucosidase, α-1,3-glucosidase	3/2
GH64	β-1,3-glucanase	1/0
GH65	Trehalase, maltose phosphorylase, trehalose phosphorylase	7/3
GH66	dextranase	1/5
GH68	β-fructofuranosidase, others	0/1
GH73	peptidoglycan hydrolase	5/1
GH74	Endoglucanase, cellobiohydrolase, xyloglucanase	29/40
GH75	chitosanase	1/0
GH77	4-α-Glucanotransferase, amylomaltase	53/50
GH81	endo-β-1,3-glucanase	3/2
GH83	neuraminidase	2/4
GH85	endo-β-N-acetylglucosaminidase	2/2
GH89	α-N-Acetylgalactosaminidase	0/1
GH95	α-L-Fucosidase	5/1
GH96	α-agarase	16/10
GH99	endo-α-1,2-mannosidase	4/2
GH109	α-N-Acetylgalactosaminidase	69/70
GH110	α-galactosidase, α-1,3-galactosidase	6/3
GH116	β-Glucosidase, β-galactosidase, β-mannosidase, others	2/2

References

Abulencia, C. B., Wyborski, D. L., Garcia, J. A., Podar, M., Chen, W., Chang, S. H. et al. (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol* **72:**3291-3301.

Amann, R. I., Ludwig, W., Schleifer, K. H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59:**143-169.

Andrews, S., Taylor, E., Pell, G., Vincent, F., Ducros, V., Davies, G. et al. (2004) The use of forced protein evolution to investigate and improve stability of family 10 xylanases - The production of Ca^{2+}-independent stable xylanases RID A-9042-2011. *J Biol Chem* **279:**54369-54379.

Bayer, E. A., Setter, E., & Lamed, R. (1985) Organization and distribution of the cellulosome in *Clostridium thermocellum*. *J Bacteriol* **163:**552-559.

Bayer, E. A., Lamed, R., & Himmel, M. E. (2007) The potential of cellulases and cellulosomes for cellulosic waste management. *Curr Opin Biotechnol* **18:**237-245.

Béjà, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P. et al. (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2:**516-529.

Bergen, A., Qi, Y., Haque, K., Welch, R., Chanock, S. (2005) Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. *BMC Biotechnol* **5:** 24.

Berlemont, R., Delsaute, M., Pipers, D., D'Amico, S., Feller, G., Galleni, M., Power, P. (2009) Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *ISME J* **3:**1070-1081.

Bhat, M. K. (2000) Cellulases and related enzymes in biotechnology. *Biotechnol Adv* **18:**355-383.

Binga, E. K., Lasken, R. S., Neufeld, J. D. (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* **2:**233-241.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365-370.

Brady, S. F., Simmons, L., Kim, J. H., Schmidt, E. W. (2009) Metagenomic approaches to natural products from free-living and symbiotic organisms. *Nat Prod Reports* **26:** 1488-1503.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* **37:** D233-D238.

Chapin, F. S., Sturm, M., Serreze, M. C., McFadden, J. P., Key, J. R., Lloyd, A. H. et al. (2005) Role of land-surface changes in Arctic summer warming. *Science* **310:** 657-660.

Chen, Y., Wu, L. Q., Boden, R., Hillebrand, A., Kumaresan, D., Moussard, H. et al. (2009) Life without light: microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave. *ISME J* **3:** 1093-1104.

Chen, Y., Dumont, M. G., Neufeld, J. D., Bodrossy, L., Stralis-Pavese, N., McNamara, N. P. et al. (2008) Revealing the uncultivated majority: combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated *Methylocystis* in acidic peatlands. *Environ Microbiol* **10:** 2609-2622.

Coker, J., Sheridan, P., Loveland-Curtze, J., Gutshall, K., Auman, A., Brenchley, J. (2003) Biochemical characterization of a beta-galactosidase with a low temperature optimum obtained from an Antarctic *Arthrobacter* isolate. *J Bacteriol* **185:** 5473-5482.

Daniel, R. (2005) The metagenomics of soil. *Nat Rev Microbiol* **3:** 470-478.

Davidson, E., Janssens, I. (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440:** 165-173.

Davies, G. (2011) "Glycoside Hydrolase family 5" in *CAZypedia*, available at http://www.cazypedia.org, accessed 5 December 2011

DeSantis, G., Zhu, Z., Greenberg, W. A., Wong, K., Chaplin, J., Hanson, S. R. et al. (2002) An enzyme library approach to biocatalysis: Development of nitrilases for enantioselective production of carboxylic acid derivatives. *J Am Chem Soc* **124:** 9024-9025.

Detter, J., Jett, J., Lucas, S., Dalin, E., Arellano, A., Wang, M. et al. (2002) Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* **80:** 691-698.

Doi, R., & Kosugi, A. (2004) Cellulosomes: Plant-cell-wall-degrading enzyme complexes. *Nat Rev Microbiol* **2:** 541-551.

Duan, C., Feng, J. (2010) Mining metagenomes for novel cellulase genes. *Biotechnol Lett* **32:** 1765-1775.

Dunford, E. (2010) Characterization of active cellulolytic consortia from Arctic tundra. University of Waterloo. M.Sc. thesis.

Engel, K., Pinnell, L. J., Cheng J., Charles T. C., Neufeld J. D. (2011) Nonlinear electrophoresis for purification of soil DNA for metagenomics. *J Microbiol Meth* **88**: 35-40

Entcheva, P., Liebl, W., Johann, A., Hartsch, T., Streit, W. R. (2001) Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *Appl Environ Microbiol* **67:** 89-99.

Epstein, H. E., Walker, M. D., Chapin, F. S., Starfield, A. M. (2000) A transient nutrient-based model of Arctic plant community response to climatic warming. *Ecol Appl* **10:** 824-841.

Falkowski, P., Scholes, R. J., Boyle, E., Canadell, J., Canfield, D., Elser, J. et al. (2000) The global carbon cycle: A test of our knowledge of earth as a system. *Science* **290:** 291-296.

Georlette, D., Bentahir, M., Claverie, P., Collins, T., D'amico, S., Delille, D., Feller, G., Gratia, E., Hoyoux, A., Lonhienne, T. (2002) Cold-adapted enzymes. *Phys Chem Bas Biotechnol* **7**: 177-196.

Gerday, C., Aittaleb, M., Bentahir, M., Chessa, J., Claverie, P., Collins, T. et al. (2000) Cold-adapted enzymes: from fundamentals to biotechnology. *Trends Biotechnol* **18:** 103-107.

Gupta, R., Beg, Q. K., Lorenz, P. (2002) Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl Microbiol Biotechnol* **59:** 15-32.

Han, B., Chen, Y., Abell, G., Jiang, H., Bodrossy, L., Zhao, J. G. et al. (2009) Diversity and activity of methanotrophs in alkaline soil from a Chinese coal mine. *FEMS Microbiol* **70:** 196-207.

Handelsman, J. (2004) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol and Mol Biol Rev* **68:** 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem and Biol* **5:** R245-R249.

Hanson, P. J., Edwards, N. T., Garten, C. T., Andrews, J. A. (2000) Separating root and soil microbial contributions to soil respiration: A review of methods and observations. *Biogeochem* **48:** 115-146.

Houghton, R. A., Davidson, E. A., Woodwell, G. M. (1998) Missing sinks, feedbacks, and understanding the role of terrestrial ecosystems in the global carbon balance. *Global Biogeochem Cycles* **12:** 25-34.

Kim, S. J., Lee, C. M., Han, B. R., Kim, M. Y., Yeo, Y. S., Yoon, S. H. et al. (2008) Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS Microbiol Lett* **282:** 44-51.

Knietsch, A., Bowien, S., Whited, G., Gottschalk, G., Daniel, R. (2003a) Identification and Characterization of coenzyme B^{12}-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl Environ Microbiol* **69:** 3048-3060.

Knietsch, A., Waschkowitz, T., Bowien, S., Henne, A., Daniel, R. (2003b) Construction and screening of metagenomic libraries derived from enrichment cultures: Generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. *Appl Environ Microbiol* **69:** 1408-1416.

Krause, L., Diaz, N. N., Edwards, R. A., Gartemann, K., Kroemeke, H., Neuweger, H. et al. (2008) Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. *J Biotechnol* **136:** 91-101.

Lanzen, A., Jorgensen, S. L., Bengtsson, M. M., Jonassen, I., Ovreas, L., Urich, T. (2011) Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiol Ecol* **77:** 577-589.

Levin, D. B., Islam, R., Cicek, N., Sparling, R. (2006) Hydrogen production by *Clostridium thermocellum* 27405 from cellulosic biomass substrates. *Int J Hydrogen* **31:** 1496-1503.

Lorenz, P., & Eck, J. (2005) Metagenomics and industrial applications. *Nat Rev Microbiol* **3:** 510-516.

Lynd, L. R., Weimer, P. J., van Zyl, W. H., Pretorius, I. S. (2002) Microbial cellulose utilization: Fundamentals and biotechnology. *Microbiol Mol Biol Rev* **66:** 739-739.

Mack, M. C., Schuur, E. A. G., Bret-Harte, M. S., Shaver, G. R., Chapin, F. S. (2004) Ecosystem carbon storage in Arctic tundra reduced by long-term nutrient fertilization. *Nature* **431:** 440-443.

Melillo, J. M., Steudler, P. A., Aber, J. D., Newkirk, K., Lux, H., Bowles, F. P. et al. (2002) Soil warming and carbon-cycle feedbacks to the climate system. *Science* **298:** 2173-2176.

Muyzer, G., Dewaal, E., Uitterlinden, A. (1993) Profiling of complex microbial-populations by denaturing gradient gel-electrophoresis: Analysis of polymerase chain reaction-amplified genes-coding for 16s ribosomal-RNA. *Appl Environ Microbiol* **59:** 695-700.

Neufeld, J. D., Chen, Y., Dumont, M. G., Murrell, J. C. (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ Microbiol* **10:** 1526-1535.

Neufeld, J. D., Engel K., Cheng J., Moreno-Hagelsieb G., Rose D. R., Charles T. C. (2011) Open resource metagenomics: a model for sharing metagenomic libraries. *Stand Gen Sci* **5:** 203-210

Neufeld, J. D., Schafer, H., Cox, M. J., Boden, R., McDonald, I. R., Murrell, J. C. (2007a) Stable-isotope probing implicates *Methylophaga* spp and novel Gammaproteobacteria in marine methanol and methylamine metabolism. *ISME J* **1:** 480-491.

Neufeld, J. D., Vohra, J., Dumont, M. G., Lueders, T., Manefield, M., Friedrich, M. W., Murrell, J. C. (2007b) DNA stable-isotope probing. *Nat Prot* **2:** 860-866.

Neveu, J., Regeard, C., & DuBow, M. S. (2011) Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. *Appl Microbiol Biotechnol* **91:** 635-644.

O'Sullivan, C., Burrell, P. C., Clarke, W. P., Blackall, L. L. (2007) A survey of the relative abundance of specific groups of cellulose degrading bacteria in anaerobic environments using fluorescence in situ hybridization. *J Appl Microbiol* **103:** 1332-1343.

Overbeek, R., Bartels, D., Vonstein, V., Meyer, F. (2007) Annotation of bacterial and archaeal genomes: Improving accuracy and consistency. *Chem Rev* **107:** 3431-3447.

Pang, H., Zhang, P., Duan, C., Mo, X., Tang, J., Feng, J. (2009) Identification of cellulase cenes from the cetagenomes of compost soils and functional characterization of one novel endoglucanase. *Curr Microbiol* **58:** 404-408.

Paradis, E., Claude, J., Strimmer, K. (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289-290.

Pel, J., Broemeling, D., Mai, L., Poon, H. L., Tropini, G., Warren, R. L. et al. (2009) Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proc Natl Acad Sci USA* **106:** 14796-14801.

Pinnell, L. J., Charles T. C., Neufeld, J. D. (2010) Stable-isotope probing and metagenomics. *Stable Isotopes in Microbial Molecular Ecology* Whiteley A., Murrell JC. (Eds.) ASM Press.

Post, W. M., Emanuel, W. R., Zinke, P. J., Stangenberger, A. G. (1982) Soil carbon pools and world life zones. *Nature* **298:** 156-159.

Radajewski, S., McDonald, I. R., Murrell, J. C. (2003) Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Curr Opin Biotechnol* **14:** 296-302.

Rahmstorf, S., Ganopolski, A. (1999) Long-term global warming scenarios computed with an efficient coupled climate model. *Clim Change* **43:** 353-367.

Rappe, M. S., Giovannoni, S. J. (2003) The uncultured microbial majority. *Ann Rev Microbiol* **57:** 369-394.

Riesenfeld, C., Goodman, R., Handelsman, J. (2004) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* **6:** 981-989.

Saitou, N., Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-425.

Schellenberger, S., Kolb, S., Drake, H. L. (2010) Metabolic responses of novel cellulolytic and saccharolytic agricultural soil Bacteria to oxygen. *Environ Microbiol* **12:** 845-861.

Schlesinger, W. H., Andrews, J. A. (2000) Soil respiration and the global carbon cycle. *Biogeochem* **48:** 7-20.

Schmeisser, C., Steele, H., Streit, W. R. (2007) Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* **75:** 955-962.

Schuur, E. A. G., Vogel, J. G., Crummer, K. G., Lee, H., Sickman, J. O., Osterkamp, T. E. (2009) The effect of permafrost thaw on old carbon release and net carbon exchange from tundra. *Nature* **459:** 556-559.

Schuur, E. A. G., Bockheim, J., Canadell, J. G., Euskirchen, E., Field, C. B., Goryachkin, S. V. et al. (2008) Vulnerability of permafrost carbon to climate change: Implications for the global carbon cycle. *Bioscience* **58:** 701-714.

Schwarz, S., Waschkowitz, T., Daniel, R. (2006) Enhancement of gene detection frequencies by combining DNA-based stable-isotope probing with the construction of metagenomic DNA libraries. *World J Microbiol Biotechnol* **22:** 363-367.

Shiratori, H., Reno, H., Ayame, S., Kataoka, N., Miya, A., Hosono, K. et al. (2006) Isolation and characterization of a new *Clostridium* sp that performs effective cellulosic waste digestion in a thermophilic methanogenic bioreactor. *Appl Environ Microbiol* **72:** 3702-3709.

Spellerberg, I., Fedor, P. (2003) A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Global Ecol Biogeogr* **12:** 177-179.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., DeLong, E. F. (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178:** 591-599.

Sul, W. J., Park, J., Quensen, J. F., Rodrigues, J. L. M., Seliger, L., Tsoi, T. V. et al. (2009) DNA-stable isotope probing integrated with metagenomics for retrieval of biphenyl dioxygenase genes from polychlorinated biphenyl-contaminated river sediment. *Appl and Environ Microbiol* **75:** 5501-5506.

Treonis, A., Ostle, N., Stott, A., Primrose, R., Grayston, S., Ineson, P. (2004) Identification of groups of metabolically-active rhizosphere microorganisms by stable isotope probing of PLFAs. *Soil Biol Biochem* **36:** 533-537.

Treusch, A. H., Leininger, S., Kletzin, A., Schuster, S. C., Klenk, H. P., Schleper, C. (2005) Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ Microbiol* **7:** 1985-1995.

Tringe, S. G., Rubin, E. M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Gen* **6:** 805-814.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E. et al. (2009) A core gut microbiome in obese and lean twins. *Nature* **457:** 480-U7.

Ueno, Y., Haruta, S., Ishii, M., Igarashi, Y. (2001) Microbial community in anaerobic hydrogen-producing microflora enriched from sludge compost. *Appl Microbiol Biotechnol* **57:** 555-562.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Voget, S., Leggewie, C., Uesbeck, A., Raasch, C., Jaeger, K. E., Streit, W. R. (2003) Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol* **69:** 6235-6242.

Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., van Sinderen, D. (2007) Genomics of *Actinobacteria*: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* **71:** 495-548.

Whitman, W. B., Coleman, D. C., Wiebe, W. J. (1998a) Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95:** 6578-6583.

Wilson, D. B. (2009) Cellulases and biofuels. *Curr Opin Biotechnol* **20:** 295-299.

Wilson, D. B., Urbanowisc B. (2010) "Glycoside hydrolase family 9" in *CAZypedia*, available at http://www.cazypedia.org, accessed 5 December 2011

Yilmas, S., Allgaier, M., Hugenholtz, P. (2010) Multiple displacement amplification compromises quantitative analysis of metagenomics. *Nat Meth* **7**: 943-944

Yokoi, H., Maki, R., Hirose, J., Hayashi, S. (2002) Microbial production of hydrogen from starch-manufacturing wastes. *Biomass Bioenergy* **22:** 389-395.

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W., Church, G. M. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24:** 680-686.

Zhou, J., Bruns, M., & Tiedje, J. (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62:** 316-322.

Zimov, S., Schuur, E., Chapin, F. (2006) Permafrost and the global carbon budget. *Science* **312:** 1612-1613.