Targeting novel soil glycosyl hydrolases by combining stable isotope probing and metagenomics

by

Yris Milusqui Verastegui Peña

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, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. 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

Soil represents the largest global reservoir of microbial diversity for the discovery of novel genes and enzymes. Both stable-isotope probing (SIP) and metagenomics have been used to access uncultured microbial diversity, but few studies have combined these two methods for accessing the biotechnological potential of soil genetic diversity and fewer yet have employed functional metagenomics for recovering novel genes and enzymes for bioenergy or bioproduct applications. In this research, I demonstrate the power of combining functional metagenomics and SIP using multiple plant-derived carbon substrates and diverse soils for characterizing active soil bacterial communities and recovering glycosyl hydrolases based on gene expression. Three disparate Canadian soils (tundra, temperate rainforest and agricultural) were incubated with five native carbon (^{12}C) or stable-isotope labelled (^{13}C) carbohydrates (glucose, cellobiose, xylose, arabinose and cellulose). Sampling at defined time intervals (one, three and six weeks) was followed by DNA extraction and cesium chloride density gradient ultracentrifugation. Denaturing gradient gel electrophoresis (DGGE) of all gradient fractions confirmed the recovery of labeled nucleic acids. Sequencing of original soil samples and labeled DNA fractions demonstrated unique heavy DNA patterns associated with all soils and substrates. Indicator species analysis revealed many uncultured and unclassified bacterial taxa in the heavy DNA for all soils and substrates. Among characterized taxa, *Salinibacterium* (*Actinobacteria*), *Devosia* (*Alphaproteobacteria*), *Telmatospirillum* (*Alphaproteobacteria*), *Phenylobacterium* (*Alphaproteobacteria*) and *Asticcacaulis* (*Alphaproteobacteria*) were the bacterial "indicator species" for the heavy substrates and soils tested. Both *Actinomycetales* and *Caulobacterales* (genus *Phenylobacterium*) were associated with metabolism of cellulose.

Members of the *Alphaproteobacteria* were associated with the metabolism of arabinose and members of the order *Rhizobiales* were strongly associated with the metabolism of xylose.

Annotated metagenomic data suggested diverse glycosyl hydrolase gene representation within the pooled heavy DNA. By screening only 2876 inserts derived from the ^{13}C-cellulose heavy DNA, stable-isotope probing and functional screens enabled the recovery of six clones with activity against carboxymethylcellulose and methylumbelliferone-based substrates.

Acknowledgements

I want to express my sincere gratitude to:

My supervisor Dr. Josh D. Neufeld for giving me the opportunity to do this research and for all his support.

Dr. Trevor Charles and Dr. David Rose for their advice and support as a committee members.

My family, for their unconditional love, care and support regardless the distance.

Special thanks to:

Soil sampling facilitators: Paul Grogan from Queen's University (Arctic tundra soil) Richard Winder from the Pacific Forestry Centre (British Columbia; temperate rainforest) and Kari Dunfield from University of Guelph (agricultural soil).

Novozymes for their collaboration in screening clones with 4-methylumbelliferone-based (4-MU) fluorogenic substrates

Dr. Brendan McConkey for all his advice, support, tolerance and double thanks for accepting to be a proofreader of this thesis.

Dr. Tanya Romantsov for her teaching, support care and many good moments. Dr. Jijiun Cheng and Katja Engel (MSc.), thanks for all the advice provided.

Brianne McCallum for all your support at all times, for many laughs and our writing time in Dana Porter's10th floor where all the wonderful sunsets can be seen. Chris Wilson, thank you for your edits. The Neufeld Lab, including all the people who pass through, thank you all.

Mike Hall, thank you for taking time to attend my requests. The Natural Sciences and Engineering Research Council (NSERC), Canada who provided financial support for this study and to everyone who made this work possible.

"I know that with You and your unfailing love, all things work together for my good"

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1AT	Arctic tundra 1
7TR	Temperate rainforest
11AW	Agricultural Soil–Wheat
AXIOME	automation, extension and integration of microbial ecology
BLAST	basic local alignment search tool
CAZy	carbohydrate-active enzymes database
C	carbon atom
CBM	carbohydrate binding module
CM^2BL	Canadian metamicrobiome library
CMC	carboxymethylcellulose
CO_2	carbon dioxide
CsCl	cesium chloride
DGGE	denaturing gradient gel electrophoresis
DMSO	dimethyl sulfoxide
DNA-SIP	deoxyribonucleic acid stable-isotope probing
dNTP	deoxyribonucleoside triphosphate
EtBr	ethidium bromide
EDTA	ethylenediaminetetraacetic acid
FID	flame ionization detector
GH	glycosyl hydrolase
GT	glycosyl transferases
H_2	hydrogen gas
IS	indicator species
IV	indicator value
Kb	kilo base
LB	Luria broth
Mb	mega base
MDA	multiple displacement amplification
MG-RAST	metagenomics rapid annotation using subsystem technology server

mmol	millimol(s)
mM	millimolar
MRPP	multiple response permutation procedure
MU	methylumbelliferone units
ng	nanograms
N_2	dinitrogen gas
NCBI	National Center for Biotechnology Information
O_2	oxygen gas
OTU	operational taxonomic unit
PANDAseq	paired-end assembler for DNA sequences
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulsed field gel electrophoresis
pH	power of hydrogen
QIIME	quantitative insights into microbial ecology
RDP	ribosomal database project
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid
RNA-SIP	ribonucleic acid stable-isotope probing
SIP	stable-isotope probing
Swissprot	manually annotated universal protein resource database
Tc	tetracycline
TCD	thermal conductivity detector
T-RFLP	terminal fragment length polymorphism
UniprotKB TrEmbl	computer-annotated protein sequence database
UniFrac	Unit fraction metric
UniProt	universal protein resource
UV	ultraviolet
V3	variable region 3
V	volts
V/cm	volts per centimeter
w/v	weight per volume
WFPS	water-filled pore space

1.0. Introduction and literature review

1.1. Soil overview

Soil is a naturally occurring material composed of organic and inorganic matter. It develops gradually by complex pedogenic processes (Agriculture and Agri-Food Canada, 1998; Osman, 2013). It is in constant change, transforming energy and recycling materials through physical, chemical and biological processes. Among these processes, the degradation of organic matter plays an important role because it releases nutrients, which support plant, animal and microbial diversity (Oades, 1984).

Soils components are organic matter and different types of mineral particles (sand, silt and clay), which vary in size. The combined proportions of these particles determine soil texture, which is a stable characteristic of different soils. Texture regulates the soil environment - for example, in fine texture soils, microbial populations and activities are higher (Sessitsch *et al.*, 2001; Osman, 2013) as texture regulates the pore size and the capacity of these soils to retain water (Thomsen *et al.*, 1999). Soil physicochemical properties are mainly governed by precipitation and temperature, which directly affect the distribution of plants, animals and microorganisms (Campanella & Mitchell, 1968; Lin *et al.*, 2003; Gelsomino & Cacco, 2006; Seyfried & Grant, 2007). The combination of the climatic conditions and specific soil-forming processes contribute to the development of specific soil types (Osman, 2013). On the other hand, soil structure is the arrangement of soil particles in aggregates and its formation is a result of different processes that include mechanical, pedogenic and microbial processes (Dexter, 1988; Astrow, 1996). Soil structure can be negatively or positively impacted by different natural phenomena such as drought, rain, freezing and thawing as well as by soil management practices such as tillage, crop rotation, fertilization and irrigation (Dexter, 1988;

Balesdent *et al.*, 2000).

In soil pores, air, water and gas exchange with the environment occurs mainly through diffusion. Gases vary in volume and proportions depending of the type of components present in each soil (Conen & Smith, 2000; Smith *et al.*, 2003). Also, modification of the porosity and pore size distribution can indirectly affect the rate of gas exchange (Del Grosso *et al.*, 2000; Horn & Smucker, 2005).

When properly managed, soils can be a significant carbon sink. One of the practices that have been used for carbon sequestration is conservation tillage, which involves different methods of tillage to avoid soil erosion. Other methods include usage of semiarid lands, water management (e.g. irrigation systems and drainage), regrowth of native vegetation and biochar application (Lal & Kimble, 1977; Post *et al.*, 2012). Soil organic matter is not only a substrate that provides food for microorganisms, but also influences the physical, chemical and biological properties of soils such as enhancement of aggregation, aeration, water movement, water retention, reduction of evaporation, soil conditioning, structure, porosity, water holding capacity, drainage, pH, stability and nutrient storage (Tisdall & Oades, 1982; Gregorich *et al.*, 1994; Li *et al.*, 2007; Osman, 2013).

There are three types of organic matter present in soils. Firstly, soils contain fresh organic matter from biological materials such as decomposing plant and animal tissues. These are in the process of incorporation into soil and can be identified by their origin and structure; they represent 1-10% of the total organic matter in soil. Secondly, partially decomposed organic matter comprises about 10-40% of the total organic matter usually present in soils. Together these two types of organic matter are known as “the active soil organic matter” because microorganisms from soils actively degrade these materials. Thirdly, decomposed or

stable organic matter (also known as humus), is represented by stable, natural organic complexes coming from the products of biological decomposition and resynthesis. This matter can coat soil particles as a gum in the aggregates making them inaccessible to decomposing microorganisms. Humus comprises about 40-60% of the total organic matter in soils (Anderson, 1979; Osman, 2013).

1.2. Soil classification

Soil classification systems organize information in order to be useful and it is a reflection of the existing knowledge about soils, as the knowledge increases and new concepts are developed the classification systems tend to be modified (Agriculture and Agri-Food Canada, 1998). Classification of soils has been prepared based on measurable and observable properties such as morphological (horizon differentiation, soils depth), physical (color, texture, structure, compaction) and chemical and mineralogical (pH, organic matter, clay, iron, aluminum oxides) (Agriculture and Agri-Food Canada, 1998).

Soil texture is a physical property that has been used to classify soils using three defined textural fractions defined by the USDA: clay (< 0.002 mm), silt (0.002 – 0.05 mm) and sand (> 0.05 mm). Using the particle size distribution and the textural triangle developed by the USDA, soils are classified in twelve textural classes. This method is the most widely used to qualitatively classify soils (Twarakavi *et al.*, 2010). Soil classification in Canada is based on soil properties that reflect processes of soil development and environmental factors and it is based on knowledge acquired from the extensive number of soils present in Canada. This classification has been influenced by concepts elaborated by other countries where the U.S. system influenced the most (Agriculture and Agri-Food Canada, 1998). The Canadian system

for soil classification is designed to cover existing soils in Canada in order to organize and make more feasible sharing and understanding the information of the soils and the environmental factors that affect them.

1.3. Soil overview from biomes and ecosystems used in this study

1.3.1. Arctic tundra soils

Arctic tundra soils are dominated by the presence of permafrost (permanently frozen ground). The soil horizons remain separate and an organic layer can be part of the surface of the soil. This type of soil is classified as cryosolic order by the Canadian System of Soil Classification (Agriculture and Agri-Food Canada, 1998). Tundra is characterized by lower temperatures, lower precipitation and a short summer where only the surface layer thaws (active layer). In this layer the annual temperature ranges between -30°C and +15°C and is where many biogeochemical processes occur (Boike *et al.*, 2012). The permafrost layer underneath allows the upper portion of the soils to remain saturated with water, therefore Arctic tundra soils are cool, anaerobic and the organic matter accumulated make these soils the largest reservoirs of carbon in the Earth (Billings, 1987). The estimated carbon content in permafrost is 1672 Pg and when it thaws the organic matter trapped becomes available for microbial metabolism (Mackelprang *et al.*, 2011).

1.3.2. Temperate rainforest soils

Forest soils have well differentiated layers or horizons with the surface layer known as the O horizon. The O horizon is divided in three sub-horizons each one of them contains large amounts and different types of organic matter where the undecomposed organic matter is at the

top and the decomposed organic matter is at the bottom. These three sub-horizons compose the forest floor and are characterized by intense microbial activity, which combined with physico-chemical characteristics of these soils make them highly fertile (Gaudinski *et al.*, 2000; Davidson *et al.*, 2006; Lal & Lorenz, 2012). Temperate rainforest biome is characterized by two defined seasons, including a hot summer (30°C) and a cold winter (-30°C) with annual precipitation ranging from 50-200 cm. Because these soils are highly porous, rates of infiltration and percolation are rapid (Lorenz & Lal, 2012; Osman, 2013). According to the Canadian System of Soil Classification, luvisolic, podzolic and brunisolic orders are common soils in the Canadian temperate forest. These orders are characterized based on the presence of different components (texture, parental material, organic matter and color) in the B horizon (Agriculture and Agri-Food Canada, 1998). In this biome, trees have a large influence over the soil properties like pH and exchangeable cations. For example, it has been reported that *Pinus sylvestris* and *Picea abies* acidify soils due to the production of organic acids (Nilsson *et al.*, 1982; Priha & Smolander, 1999; Augusto *et al.*, 2002; Lindroos *et al.*, 2011). Also temperate forests are reservoirs of carbon where high amounts of carbon are stored in the soils and in the vegetation (Lal, 2005; Pan *et al.*, 2011).

1.3.3. Agricultural soils

Agricultural soils have originated from forest soils and were selected because of their desirable characteristics for agricultural use. These soils were managed using some practices such as fertilization, irrigation, cultivation, mechanization and land conservation in order to improve the characteristics and transform them into deep productive soils suitable to sustain annual crops of commercial value (Boyer & Groffman, 1996; Lorenz & Lal, 2012). These soils

do not show defined profiles because they were highly disturbed through different agricultural practices (Boyer & Groffman, 1996; Osman, 2013). These disturbances cause shifts in microbial communities (Altieri, 1999; Kaisermann *et al.*, 2013). The incorporation of the organic matter is low, due to reduced organic residues, deposition and annual crop removal. Therefore the microbial activity is reduced. Also due to their smaller pores, the water percolation and infiltration is poor (Boyer & Groffman, 1996; Schjonning *et al.*, 2002). Most agricultural soils have pH ranges between 6.5-7.5 because neutral pH values benefit the majority of the crops and the nitrogen-fixing bacteria grow better in neutral soils (Osman, 2013).

The input of carbon in agricultural soils is from plant photosynthesis in addition to manure and organic residues that increase carbon content (Paustian *et al.*, 1997; Lorenz & Lal, 2012). Decomposition of soil organic matter by bacteria and fungi helps to recycle the carbon in agricultural soils (Viaud *et al.*, 2000; Johnson *et al.*, 2003). Fungi decompose fresh organic matter and bacteria consume both fresh and old organic matter. Fungi are more efficient in carbon assimilation than bacteria, and fungal communities are dispersed near the soil surface while bacterial communities are dominant in deeper layers. Bacteria decompose more labile substrates such as low molecular weight soluble sugars, amino acids and polyphenols that are solubilized and leached to the deepest layers of the soil (Cleveland *et al.*, 2004; Lorenz & Lal, 2012).

1.4. Microbial diversity in soil

Biodiversity is a measure of the number of species that exist in a geographic region (Nannipieri *et al.*, 2003). Total diversity includes ecological, metabolic and genetic diversity;

therefore, the study of microbial diversity in soils requires the understanding of the ecological (physical and chemical properties of soils), metabolic (carbon and energy sources present in soils) and genetic aspects (distribution of genetic information between organisms present in soils) (Finlay *et al.*, 1997; Brussaard *et al.*, 1997; Nannipieri *et al.*, 2003; Fierer *et al.*, 2007). Bacterial community structure and diversity can be used to test the quality of soils (Sharma *et al.*, 2010). Therefore, it is important to comprehend the relationship between microbial diversity and soil function (Torsvik & Øvreås, 2002; Sharma *et al.*, 2010; Monard *et al.*, 2011); understanding the relations between genetic diversity in soils and bacterial community structure and between community structure and function is required (Neufeld *et al.*, 2007a; Liebner *et al.*, 2008; Urich *et al.*, 2008).

Soils host diverse microhabitats with extensive physicochemical gradients and environmental conditions where microorganisms live in consortia, actively interacting with other members of the soil biota (Ladd *et al.*, 1993; Stotzky, 1997; Brussaard *et al.*, 1997). Microbial growth is larger on the surfaces of soil particles, usually within plant rhizosphere. Even in small soil aggregates there can be found many different microenvironments, where high microbial diversity can be present (Stotzky, 1997; Nannipieri *et al.*, 2003; Gardner *et al.*, 2011). Among chemical characteristics of soil, pH seems to be the most important factor that influences soil bacterial richness and diversity with higher diversity found in neutral pH and lower diversity associated with acidic soils (Fierer & Jackson, 2006). A study in Arctic soils showed that the composition of the bacterial communities reflects pH-dependent trends apparent in other biomes and ecosystems (Chu *et al.*, 2010) with *Acidobacteria*, *Alphaproteobacteria*, *Actinobacteria*, *Betaproteobacteria* and *Bacteroidetes* the dominant phyla in these soils. On the other hand, in agricultural soils, soil type (sand loam and clay) and

plant species seem to be the most important factors defining bacterial community composition (Girvan *et al.*, 2003; Marschner *et al.*, 2004).

1.5. Microbial Processes in soil

Soil represents a complex and essential component of terrestrial ecosystems, critical for myriad important processes that sustain life. For example, soil represents a physical support for living creatures, a reservoir for organic and mineral substances and a medium in which a great variety of organisms live (Gobat *et al.*, 2004; Bardgett, 2005). Soil microorganisms catalyze Earth's biogeochemical reactions, including the degradation of organic matter and the recycling of nutrients (Van Veen & Kuikman, 1990; Bardgett, 2005; Borch *et al.*, 2010). Microbial processes in soils such as organic matter decomposition, ammonification, denitrification, nitrification, nitrogen fixation, phosphorous and sulfur transformations are important because they govern soil productivity and fertility (Quastel, 1965; Gil-Sotres *et al.*, 2005; Osman, 2013). All of these are processes are regulated by enzymes with rates that depend on factors like soil temperature and oxygen (Davidson & Janssens, 2006; Gardner *et al.*, 2011; Schimel & Schaeffer, 2012; Osman, 2013). For example, organic matter turnover is slow at low temperatures and in anaerobic soils (Bridgham *et al.*, 1998; Koch *et al.*, 2007). Other important microbial processes include synthesis of antibiotics and degradation of contaminants (Krueger *et al.*, 1991; Raaijmakers *et al.*, 2002; Handelsman, 2007; Prusov, 2013).

There is still information that needs to be revealed about microbial processes in soils. A better understanding of how a microbial community is structured, which microorganisms are part of that community and how members interact with each other to maintain the community

functioning, is necessary (Handelsman, 2007). Attesting to the heterogeneity, interactivity and connectivity of the soil niche, traditional culture-based techniques grossly underestimate microbial diversity. Readily cultured microorganisms typically represent a very small proportion of soil microbial communities, typically less than 1% (Amann *et al.*, 1995). The "uncultured majority" harbor an enormous reservoir of uncharacterized organisms, genes and enzymatic processes (Torsvik and Øvreås, 2002). An outstanding methodological challenge remains: how best to access the biotechnological potential contained within the DNA of uncultured microorganisms?

1.6. Tools to assess the biotechnological potential in soil microbial communities

1.6.1. Metagenomics

Metagenomics is currently defined as any functional or sequence-based cultivation-independent analysis of the microorganisms from a determined habitat or environment (Handelsman, 2005; Sleator *et al.*, 2008; Simon & Daniel, 2009). This broad methodological approach uses PCR-dependent or independent techniques to analyze natural microbial communities (Zhou *et al.*, 1997; Xu, 2006) and enables studies of organisms that are not readily cultured (Handelsman, 2005; Guazzaroni *et al.*, 2009). Metagenomics can complement or replace culture-based analysis and overcome some of their limitations. It is also a powerful tool for the evaluation of the phylogenetic diversity of the unexplored and uncharacterized microbial diversity present in soils, sediments or water (Neufeld & Mohn, 2006).

Metagenomics is a rapidly expanding area of research, providing new information about microbial life and providing access to novel biomolecules, enzymes and drugs of industrial importance. This methodological approach enables the analysis of the genetic diversity and

metabolic potential of microbial communities, as well as interactions between microorganisms and the environment, and the role that microbes potentially play in each community (Simon & Daniel, 2009). Even though applied metagenomics has unveiled the importance and potential of microbial diversity in different fields of study, still more studies in microbial metabolism are necessary.

Metagenomics captures environmental genomes from extracted community DNA, circumventing the need for cultivation and enabling the exploration of microbial genetic diversity and biotechnological potential (Simon & Daniel, 2009). Metagenomic analyses have exposed new microbial pathways and reactions, yielding novel enzymes and products of economic importance. Given that metagenomic studies demonstrate that the majority of total genetic diversity space remains unexplored, "it will be far more efficient and productive to seek new enzymes from metagenome libraries than to tweak the activities of existing ones" (Ferrer *et al.*, 2009). Indeed, there are several recent examples of glycosyl hydrolases (e.g., cellulases) recovered by functional screening metagenomic libraries from terrestrial environments (e.g., Kim *et al.*, 2008; Wang *et al.*, 2009; Liu *et al.*, 2011; Nacke *et al.*, 2012). These studies reflect an important limitation of bulk DNA metagenomic library construction: in the absence of a suitable growth selection for phenotypes, many clones (e.g, tens of thousands) must be screened laboriously prior to recovering targets of interest. In addition, recovered clones are theoretically the most abundant target genes in the microbial community of interest. Targeted metagenomic approaches thus offer the potential to filter for sequences of particular microbial taxa of relevance to an activity of environmental or industrial relevance.

1.6.2. Stable-isotope probing (SIP)

Stable-isotope probing (SIP) is a culture-independent method for identifying microorganisms that assimilate a particular growth substrate (Radajewski *et al.*, 2000; Radajewski *et al.*, 2002; Dumont & Murrell, 2005; Neufeld *et al.*, 2007b). For the analysis of nucleic acids of active organisms, a stable-isotope-labeled substrate (e.g., ^{13}C-labelled or ^{15}N-labelled) incorporates within the DNA (DNA-SIP) or RNA (RNA-SIP) of active organisms and isopycnic ultracentrifugation can physically recover labeled nucleic acids from those of unlabeled community members. Molecular analysis of the labelled DNA or RNA can reveal phylogenetic and functional information about the microorganisms responsible for the metabolism of a particular substrate or link the identity of microorganisms in the environment to particular functions (Neufeld *et al.*, 2007a; Chen & Murrell, 2010).

1.6.3. Combining Stable Isotope Probing (SIP) and Metagenomics

Combining SIP with metagenomics can be useful in genome and metabolic network reconstruction of active and uncultivated microorganisms, providing insights into the functions of less-abundant community members and exploring complex environmental processes such as biodegradation (as reviewed in Wackett, 2004; Pinnell *et al.*, 2010; Chen & Murrell, 2010). Until now, a small number of studies that combine DNA–SIP techniques with metagenomics have been published (Pinnell *et al.*, 2010; Chen & Murrell, 2010). Initial studies showed clearly how combining these approaches can have advantages over conventional metagenomics, facilitating the establishment of a direct link between identity and function. None of the earlier studies combining SIP and metagenomics have been used to recover glycosyl hydrolases, although this could have been done in SIP studies related to carbohydrate metabolism. For example Kovatcheva-Datchary *et al.* (2009) used SIP and terminal restriction

fragment length polymorphism (T-RFLP) to study bacteria from the human colon using an *in vitro* model in order to find those involved in starch degradation. Analysis of ^{13}C-labelled 16S rRNA genes indicated that *Ruminococcus bromii, Pevotella* spp*. and Eubacterium rectale* were involved in starch metabolism. Another group used SIP and denaturing gradient gel electrophoresis (DGGE) to identify that *Dyella, Mesorhizobium* sp., *Sphingomonas* sp. and an uncultured deltaproteobacterium affiliated with *Myxobacteria* were linked to cellulose degradation (el Zahar Haichar *et al.*, 2007). Recently, a study used SIP incubations to determine the impact of oxygen on the metabolic responses of bacteria involved in cellulose and cellobiose degradation in an agricultural soil. The results suggested that cellulolytic bacteria are different from saccharolytic bacteria and oxygen availability defined the different taxonomic groups involved (Schellenberger *et al.*, 2010). In each case, using SIP the authors were able to link microorganisms with function. On the other hand, studies using metagenomic approaches alone have reported novel glycosyl hydrolases. These discoveries were identified by screening metagenomic libraries for enzyme activities, or sequence analysis from different environmental samples (Sharma *et al.*, 2008; Li *et al.*, 2009; Simon & Daniel, 2009). Therefore, metagenomics has identified novel genes that encode enzymes that can be useful in industry, and SIP elucidates the functional role of microorganisms in diverse communities. Combining both methods will enable the retrieval of new genes and enzymes from uncultivated and low relative abundance microorganisms actively involved in different metabolic processes (Pinnell *et al.*, 2010). In particular, this research proposal describes the application of DNA-SIP and metagenomics for the recovery of a particular group of industrially relevant enzymes: glycosyl hydrolases.

Although several studies have combined DNA-SIP and metagenomic sequencing to

identify high proportions of genes from active microorganisms, such as those using glycerol (Schwarz *et al.*, 2006), C_1 compounds (Dumont *et al.*, 2006; Chen *et al.*, 2008; Kalyuzhnaya *et al.*, 2008; Neufeld *et al.*, 2008) and biphenyl (Sul *et al.*, 2009; Lee *et al.*, 2011). All of these studies were focused on the analysis of single substrates or individual samples. In addition, only one of these studies used functional metagenomic screens, expressing labeled DNA within a surrogate host for identifying enzyme activity (Schwarz *et al.*, 2006). The ability to identify genes based on function, instead of sequence homology, is arguably the most powerful application of metagenomics for the recovery of novel genes (Neufeld *et al.*, 2011) and a natural bedfellow of the SIP approach for targeting active-yet-uncultured microorganisms (Pinnell *et al.*, 2010).

In this study, I expanded on previous efforts to combine SIP and metagenomics, focusing on the identification of soil microorganisms active in degrading plant-derived organic carbon, and the recovery of glycosyl hydrolases through activity-based functional metagenomic screens. The hypothesis was that combining metagenomics and stable-isotope probing (SIP) would identify novel microorganisms and enzymes, and the recovery of enzymes from these combined methods will be higher than previous efforts using conventional metagenomics.

1.7. Glycosyl Hydrolases

The degradation of plant organic matter by the combined action of glycosyl hydrolase (GH) enzymes is an important soil function. The GH group of enzymes is distributed across a wide variety of organisms. They catalyze the hydrolysis of glycosidic bonds from complex carbohydrates (e.g., cellulose, hemicellulose) to release simple sugars (e.g., pentoses and hexoses) and as a result, GHs include important enzymes for biotechnological applications.

Because glycosidic bonds are considered among the most stable linkages that occur naturally, GHs are credited as some of the most proficient catalysts (Tkacz & Lange, 2004).

Until now, 133 families of glycosyl hydrolases have been defined and they have been grouped into 14 "clans" (Cantarel *et al.*, 2009; Lombard *et al.*, 2013). A protein clan is defined by amino acid sequences sharing common ancestry by virtue of shared tertiary structure, catalytic residues and mode of action (Henrissat & Bairoch, 1996). Glycosyl hydrolases can be classified according to substrate specificity and type of glycosidic bonds involved in hydrolysis (O- or S-glycosides). They can also be classified for their mechanism of action as either retaining or inverting the anomeric configuration (McCarter & Withers, 1994; Davies & Henrissat, 1995; Withers, 1995) the mode of action as exo- or endo-acting or based on amino acid sequence similarity (McCarter & Withers, 1994; Davies & Henrissat, 1995; Henrissat & Davies, 1997; Tkacz & Lange, 2004). Note that some enzymes that hydrolyze the same substrate may be from different families and enzymes with different substrate specificities can belong to the same family (Davies & Henrissat, 1995).

Protein 3-D structures are highly conserved (Davies & Henrissat, 1995) and GHs contain a variety of folds that are not necessary related to enzyme substrate specificity. The type of fold that each GH contains is part of the characterization in the CAZy database (Lombard *et al.*, 2013). The major structural folds described until now include the $(\beta/\alpha)_8$ barrel, which is also known as a TIM barrel because this was first discovered in the triose phosphate isomerase. CAZy clans that have this type of fold are G,H and K and families 14, 29, 31, 67 and 84. The β-jelly roll is also known as a swiss roll, and clans B and C use this type of fold. The β-propeller folds are found in glycosidases from clans F and J. The other major structural

fold involves the $(\alpha/\alpha)_n$ barrel and this type of fold is found in clans L and M, also in GH families 9, 88 and 105 (Hancock & Withers, 2007; Lombard *et al.*, 2013).

GHs can also be divided according to the topology of the active site, which is directly related to the way that the enzyme functions with a determined substrate. Until now, all known GHs can be divided into three classes: a) Pocket or crater classes cleave a carbohydrate from the non reduced end and are found in monosachharidases (β-galactosidase and β-glucosidase) and exopolysaccharidases (β –amylase and glucoamylase); b) Cleft grooves are open structures present in endo-acting polysaccharidases such as endocellulases, chitinases, endoglucanases, where the open structure allow the enzyme to bind randomly in the internal linkages of polysaccharides; c) Tunnel structures involve a protein cleft that is partially covered with loops, giving a tunnel topology, attacking mainly the ends of the carbohydrates. The tunnel topology structure is found mainly in cellobiohydrolases and exocellulases (Davies & Henrissat, 1995; Hancock & Withers, 2007).

1.7.1. Glycosyl hydrolases in industry

Glycosyl hydrolases can have a variety of uses related to the degradation of biomass, which can be used in ethanol production (Schäfer *et al.*, 2007; Maki *et al.*, 2009). Combinations of glycosyl hydrolases are used in the production of fruit and vegetable juice for extraction, clarification and stabilization (Tkacz & Lange, 2004). Xylanases have been used in baking industry to improve bread characteristics and quality by strengthening the gluten and extending shelf life. Also, it is used to make xylan sugar derivatives (xylose, xylocellobiose) (Subramaniyan & Prema, 2002; Tkacz & Lange, 2004; Butt *et al.*, 2008). For pulp and paper pre-bleaching, xylanases help to minimize chlorine inputs in subsequent processing steps,

which improves paper properties (Subramaniyan & Prema, 2002; Tkacz & Lange, 2004; Schäfer *et al.*, 2007). In animal food industries, 1,3-1,4-β glucanases and xylanases are used to hydrolyze non-starch polysaccharides, which are not able to be metabolized by young monogastric animal like chickens and piglets (Tkacz & Lange, 2004). Glycosyl hydrolases are also added to detergents for the washing of cotton fabrics, whiteness and assisting in the maintenance of colours (Schäfer *et al.*, 2007).

1.7.1.1. Cellulases in industry

One of the groups that has a variety of uses in industry are the cellulases, which are used in food processing, textiles, and laundry detergents. However, cellulases have future potential for biofuel production through the hydrolysis of cellulose (Kennedy *et al.*, 2011; Horn *et al.*, 2012). Cellulose is a polymer of ß-1,4 linked glucose. The repeating unit of cellulose polymer is the cellobiose molecule. Cellulose is a stable crystalline structure (Warren, 1996). Cellulases, together with hemicellulases and polysaccharidases, are grouped as O-glycoside hydrolases (Lynd *et al.*, 2002).

Enzymatic hydrolysis of cellulose is not completely understood and only a small proportion of microorganisms are specialized in its degradation (Wilson, 2011). There are diverse mechanisms that cellulolytic microorganisms use to degrade cellulose. Most of the microorganisms that use the free cellulose mechanism are aerobic (Bayer *et al.*, 2004; Wilson, 2010). These cellulose degraders secrete their enzymes (cellulases) that attach to the substrate through a carbohydrate-binding module (CBM), which most of these cellulases possess. Although many anaerobic microorganisms also contain cellulases, the majority of them are cellulolytic enzyme systems known as cellulosome (Leschine, 1995; Wilson, 2011), they are

present in protuberances that are present in the bacterial cell wall when they grow on cellulosic materials (Lynd *et al.*, 2002) and only few of these enzymes have CBM (Lynd *et al.*, 2002; Gowen & Fong, 2010; Wilson, 2011). Cellulosome formation occurs under carbon-limited conditions and the composition of the cellulosome depends of the carbon sources used. For example, cellulases can be produced when soluble carbohydrates are present, such as cellobiose, but the cellulosome may need cellulose to trigger the expression of these enzymes (Lynd *et al.*, 2002).

Cellulases are diverse. They catalyze the hydrolysis of the ß-1,4 linkage between two glucose monomers and have eight different types of protein folding. Cellulases, like other enzymes that have to degrade insoluble substrates, contain a substrate-binding domain. There are 67 recognized families of CBMs (Lombard *et al.*, 2013). The CBMs connect the catalytic domain and the substrate through a linker peptide. The function of the CBM is not completely understood but it is clear that the CBM binds the enzyme to the substrate and enables access to the catalytic site (Lynd *et al.*, 2002; Gowen & Fong, 2010; Wilson, 2011). The CBMs can have different affinity for cellulose, where some can have affinity for both crystalline and amorphous cellulose, and others display specificity for only one of these forms (Warren, 1996).

Most cellulases are endocellulases, with an open active site, which cleaves any accessible ß-1,4 linkages along the chain. Exocellulases have their active site inside of a tunnel, making that possible only the cleavage of one of the ends of the cellulose chain. This could explain why exocellulases have low activity in carboxymethylcellulose (CMC; Wilson, 2011). Among exocellulases, there are two distinctive groups: one group hydrolyses the reducing end and the other group cleaves the non-reducing end of the cellulose molecule (Wilson, 2011). Cellulose is also broken down by glucanases, which are mainly found in bacteria (Wilson,

2011). Endoglucanases act on random internal bonds in cellulose similar to endocellulases. Some exocellulases can have low endoglucanase activity (Warren, 1996).

Cellulose hydrolysis has been studied in the aerobic genera, *Thermobifida* and *Cellulomonas*, among major soil degraders of cellulose (Lynd *et al.*, 2002), and the most studied environments for cellulases were the rumen and compost. Therefore, with the potential of metagenomics, soils become a great source for mining novel cellulases that can be useful in industry (Wilson, 2011). Previous metagenomic studies have examined cellulose degraders (Li *et al.*, 2009) but the mechanisms for this process and microbial communities responsible in soils are still poorly understood (Wilson, 2011). Also, the vast majority of these industrial hydrolases are derived from cultured microorganisms, which may represent less than 1% of the existing diversity (Davies & Henrissat, 1995; Henrissat & Bairoch, 1996; Bauer *et al.*, 1998).

Bacteria and fungi are known as major cellulose and plant polysaccharide degraders (Bagnara *et al.*, 1985; Leschine, 1995; Busk & Lange, 2013). Among the major cellulolytic bacteria, *Cytophaga hutchinsoni* is aerobic and uses cellulose as its only carbon source (Walker & Warren, 1937). In addition, the genus *Cellvibrio,* described by Winogradsky in 1929, is grouped into the aerobic cellulolytic bacteria (Mergaert *et al.*, 2003). Anaerobic cellulolytic bacteria include members of the genera *Clostridium* and *Acetivibrium* (Eichorst & Kuske, 2012). An extensive number of cellulolytic bacteria have been isolated from different ecosystems such as invertebrates gut (termite, snail, caterpillar, bookworm, beetle larvae) (Gupta *et al.*, 2012; Huan *et al.*, 2012), forest and agricultural soils (Hatami *et al.*, 2008), grassland soils (Eichorst & Kuske, 2012) and rumen (Teather & Wood, 1982). Studies in the intestine show that the genus *Fibrobacter* plays an important role in the digestion of plant fibers (Lin & Stahl, 1995; Jun *et al.*, 2007).

Studies in fungal communities exhibited that fungal cellulolytic activities vary according with the ecosystem and may be influenced by elevated CO_2 (Weber *et al.*, 2011). Using Congo Red and dinitrosalicyclic acid reagent method, cellulolytic fungi have been characterized and showed that *Trichoderma reesei*, *Trichoderma harzianum* and *Neurospora crassa* were important cellulose degraders (Sazci *et al.*, 1986). Fungi have been isolated from different ecosystems, the majority of them were isolated from dead and decaying wood, other sources were paper pulp, sugarcane bagasse and cow manure (Naik *et al.*, 2012).

Microbial diversity in the natural environment has only recently been studied with the advent of molecular (DNA- and RNA-based) techniques and this newly described diversity is recognized as an enormous reservoir of metabolic potential, which represents a near limitless source of new enzymes that could help satisfy the specific demands of industry. Unifying research efforts and different molecular and culture-based technologies can make this a possibility within reach (Schäfer *et al.*, 2007). Recent research suggests a broad diversity of bacteria contribute to plant polymer degradation (Bernard *et al.*, 2007; el Zahar Haichar *et al.*, 2007; Bernard *et al.*, 2009; Schellenberger *et al.*, 2010), suggesting that cultivation-independent methods, like metagenomics, are most strategic for the recovery of genes and enzymes from these heterotrophic bacteria.

1.8. Research hypothesis and objectives

The proposed research will test two hypotheses:

1. Combining metagenomics and stable-isotope probing (SIP) will lead to the identification of novel enzymes and microorganisms.

2. The recovery of enzymes from this combined approach will be higher than by conventional

metagenomics.

In order to test these hypotheses, this research had two key objectives:

a. Determination of the active and uncultivated microorganisms involved in the metabolism of plant-derived carbohydrates from a selected subset of Canadian soil samples.
b. Isolation and initial characterization of novel glycosyl hydrolases with potential for industrial applications.

2.0. Materials and Methods

2.1. Soil samples.

Three soil-sampling sites represented by the Canadian MetaMicrobiome Library (http://www.cm2bl.org/) were used for this study: Arctic Tundra 1 (1AT), Temperate Rainforest (7TR) and Agricultural Soil – Wheat (11AW). Triplicate surface soils from the top 10 cm below the litter layer were combined to prepare a single composite for each site. Composite soil samples were sieved (2 mm) and samples sent to the Agriculture and Food Laboratory, Laboratory services from the University of Guelph (Guelph, ON) for analysis of total organic and inorganic carbon, pH, total nitrogen content, particle size distribution, soil moisture content and bulk density.

2.2. Stable isotope probing.

Prior to stable-isotope probing (SIP), composite soil samples were pre-incubated under experimental temperature conditions for two weeks to minimize carbon available to compete with labeled substrates. The conditions for this pre-incubation were dark storage at 15°C for 1AT and dark storage at room temperature (~24°C) for 7TR and 11AW.

The SIP incubations were conducted with both stable-isotope (^{13}C) and native (^{12}C) substrates, in addition to no-substrate controls, for each of the three soils. Ten grams of soil were added to 120-mL serum vials and substrates were added as follows. Finely shredded cellulose was purified in our lab from *Gluconacetobacter xylinus* incubations with ^{13}C or ^{12}C glucose, as previously described (Pinnell *et al.*, 2013), and 200 mg (6.6 mmol C) was mixed into serum vials in a single dose. Glucose, cellobiose, arabinose, and xylose incubations involved the addition of three weekly substrate pulses of 1.5 mmol of C, which are

approximately 5-500 times higher than normally detected in soils (Medeiros *et al.*, 2006; Hill *et al.*, 2008), but were chosen to ensure detectable labeling, similar to a previous experimental approach (Schellenberger *et al.*, 2010). Substrates were D-Glucose (Bio Basic Inc.; Markham, ON), (U-$^{13}C_6$)-D-Glucose (Cambridge Isotope Laboratories Inc.; Cambridge, ON), D-(+)-Cellobiose (Sigma-Aldrich; Oakville, Ontario), (UL-$^{13}C_{12}$)-Cellobiose (Omicron Biochemicals Inc.; South Bend, IN), D-(-)-Arabinose (Sigma-Aldrich), D-(UL-$^{13}C_5$)-Arabinose (Omicron Biochemicals Inc.), D-(+)-Xylose (Sigma-Aldrich), and D-(UL-$^{13}C_5$)-Xylose (Omicron Biochemicals Inc.). Serum bottles were sealed with butyl septa and crimp seals.

Incubation temperatures and conditions were the same as for the pre-incubation. Samples were collected at weeks one and three for glucose, cellobiose, arabinose, and xylose and weeks three and six for cellulose (Table 1). Serum vials were aerated once per week for one hour in a fume hood. The weight of incubation flasks was assessed weekly and water-filled pore space (WFPS) was maintained between 50-60% by adding distilled water and/or substrate for each incubation according to the following formula (Franzluebbers, 1999):

$$WFPS = w\ [\rho_b\ \rho_s / \rho_s - \rho_b\ ],$$

where w is the gravimetric water content (%), ρ_b is the soil bulk density (g/cm^3), and ρ_s is the soil particle density (2.65 g/cm^3).

Table 1. Experimental design and amount of substrate used in the experiments

Soils	Substrate	Sampling time points
	Glucose (3 pulses of 1.5 mmol of carbon)	1 and 3 weeks
Tundra (1AT)	Cellobiose (3 pulses of 1.5 mmol of carbon)	1 and 3 weeks
Temperate rainforest (7TR)	Arabinose (3 pulses of 1.5 mmol of carbon)	1 and 3 weeks
Agricultural soil - wheat (11AW)	Xylose (3 pulses of 1.5 mmol of carbon)	1 and 3 weeks
	Cellulose (200 mg equivalent to 6.6 mmol of carbon)	3 and 6 weeks

2.3. Gas chromatography

All experimental serum vial headspaces were monitored for CO_2 accumulation with a Shimadzu GC-2014 equipped with a thermal conductivity detector (TCD), methanizer and a flame ionization detector (FID). In addition, no added carbon control soil incubations and separate serum bottles amended with ^{12}C-glucose were monitored for headspace gases as surrogates for experimental bottles because an N_2-free headspace was required for measuring O_2 with the gas chromatograph. The headspace of these separate flasks were flushed with helium and supplemented with oxygen (20%). Headspace CO_2 and O_2 were measured every three days by direct injection of 0.5 mL of headspace gas through a packed Poropak Q column with a helium flow of 20 ml/min. The GC temperatures were maintained for the oven (80°C), TCD (280°C), methanizer (380°C) and FID (250°C).

2.4. DNA extraction and isopycnic centrifugation

Two grams of soil were sampled from each vial at weeks three and six for cellulose and at weeks one and three for glucose, cellobiose, arabinose and xylose. These soil samples were used for DNA extraction, isopycnic centrifugation and DGGE analysis. All initial DNA extractions were done with the PowerSoil DNA isolation kit (MO BIO Laboratories; Carlsbad, CA) according to the manufacturer's instructions. Extracted DNA was quantified with spectroscopy (NanoDrop 2000 UV-Vis Spectrophotometer; Thermo Scientific; Montreal, QC) and electrophoresis on a 1% agarose gel. Cesium chloride (CsCl) gradients for all soils, substrates and time points were processed by ultracentrifugation and fractions collection as described previously (Neufeld *et al.*, 2007b; Dunford & Neufeld, 2010).

2.5. Denaturing gradient gel electrophoresis (DGGE)

Bacterial fingerprints of all SIP gradient fractions confirmed substrate-specific labeling of active soil Bacteria. The V3 regions of bacterial 16S rRNA genes were targeted for DGGE using primers 341f-GC (5' - CGCCCGCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGG CCTACGGGAGGCAGCAG - 3') and 518r (5' - ATTACCGCGGCTGCTGG - 3'; Muyzer *et al.*, 1993). Each PCR contained 19.75 µl of UV-treated water, 2.5 µL ThermoPol Reaction Buffer (10X; New England BioLabs; Whitby, ON), 0.05 µl of dNTPs (100 µM), 0.05 µl of forward primer 341f-GC (100 µM), 0.05 µl of reverse primer 518r (100 µM), 1.5 µl of bovine serum albumin (BSA; 10 µg/mL), 0.25 µl of *Taq* DNA polymerase (New England BioLabs; 5U/µL) and 1 µl of DNA template purified from each gradient fraction. PCR amplification was performed on a DNA Engine (Bio-Rad; Mississauga, ON). Reaction conditions were initial denaturation at 95°C for 5 min, 30 cycles of denaturation at 95°C for 1 min, annealing at 55°C for 1 min, extension at 72°C for 1 min, followed by a final extension at 72°C for 7 min. All PCR products were first inspected on a 1% agarose gels prior to DGGE.

Each lane of a 10% polyacrylamide gel with a denaturing gradient of 30-70% was loaded with 5 µl of each PCR product. Gels were run at 60°C for 14 h at 85 V in the DGGEK-2001-110 (C.B.S. Scientific Inc.; San Diego, CA) as previously described (Green *et al.*, 2010). A custom DGGE ladder was run at both ends of the gel for normalizing patterns. Gels were stained for 45 min with SYBR Green I Nucleic Acid Gel Stain (Thermo Fisher) and rinsed in water. Gel images were prepared with a Pharos Plus Molecular Imager System (Bio-Rad).

2.6. Next-generation sequencing

High-throughput sequencing of the 16S rRNA gene V3 region and paired-end read

assembly were conducted as described previously (Bartram *et al.*, 2011; Masella *et al.*, 2012), targeting both the "heavy" and "light" fractions for each gradient for all soils, stable-isotope substrates and time points. Based on DGGE data the fractions that contained all or the majority of bands with ^{13}C-labelled DNA were chosen to represent the heavy DNA and the fraction that had all or the majority of bands related with ^{12}C-unlabelled DNA represented the light DNA. We sequenced fractions six (heavy) and ten (light) for 1AT and fractions five (heavy) and ten (light) for 7TR and 11AW (60 samples in total). Three 25-µL PCR amplifications per sample were conducted, each containing 5 µL of the 5X Phusion HF Buffer (Finnzyme, Finland), 0.125 µL of the V3 F-modified primer (100 µM), 1.25 µL of an indexed reverse primer (10 µM; V3-1R to V3-60R), 0.2 µL of dNTPs (100 µM), 0.25 µL of the Phusion High-Fidelity DNA Polymerase (2U/µL; Finnzyme) and 1 µL of DNA template (~1-10 ng). The PCR conditions were as follows: initial denaturation at 98°C for 2 min, 20 cycles of denaturation at 98°C for 10 sec, annealing at 50°C for 30 sec and extension at 72°C for 15 sec. A final extension was at 72°C for 7 min. Triplicate reactions were pooled and PCR products from individual indexed composites were combined in equal ng ratios. The PCR products were visualized by electrophoresis on a 2% agarose gel. The correct band (330 bp) was excised and purified using Wizard SV Gel and PCR Clean-Up System (Promega; Madison, WI). Libraries were subjected to 108-base paired-end sequencing on the Genome Analyzer IIx (Illumina Inc.; San Diego, CA) at the Plant Biotechnology Institute (Saskatoon, SK).

Shotgun metagenomic sequencing was performed on DNA from three pooled fractions of the ^{13}C-labeled DNA from each treatment. Pooling of heavy DNA resulted in three composite samples for sequencing: 1) Low pH (fractions five, six and seven of A1T, fractions four, five and six of 7TR) for week 3 incubations with glucose, cellobiose, arabinose and

xylose, 2) Agricultural soil (fractions four, five and six for 11AW) for week 3 incubations with glucose, cellobiose, arabinose and xylose and 3) Cellulose (fractions five, six and seven for A1T, fractions four, five and six for 7TR and 11AW) for week 6 incubations with cellulose. Metagenomic DNA samples were prepared for shotgun sequencing on the Illumina sequencing platform using the Nextera DNA Sample Preparation Kit (Illumina Inc.). Beginning with 25-50 ng of pooled heavy DNA, samples were fragmented and purified using the DNA Clean & Concentrator kit (Zymo Research Corporation; Irvine, CA). Purified fragments used as template for limited-cycle PCR amplification (5 cycles); indexed sequencing adapters (Epicentre; Madison, WI) were used for the PCR. Each amplified sample was purified and subjected to size selection (400-800 bp) using a Pippin Prep instrument (Sage Science, Beverly, MA). Afterwards, each library was quantified using the KAPA Library Quantification Kit (KAPA Biosystems; Woburn, MA). Equimolar samples were pooled, concentrated and quantified. Final concentrations were adjusted to 10 nM. Libraries were sequenced using the HiSeq2000 Sequencing System (Illumina) by the Institute for Genomic Biology Core Facility (University of Illinois). Sequencing was performed using a TruSeq SBS Kit (version 3) and data analyzed using the Cassava 1.8 pipeline. Error rates were estimated at below 0.3%. Each sample yielded 42-90 million 100-base paired-end reads and these reads were deposited in MG-RAST with identification numbers: Low pH forward 4482593.3, Low pH reverse 4483544.3, Cellulose forward 4482599.3, Cellulose reverse 4483820.3, Agricultural forward 4482600.3 and Agricultural reverse 4483819.3.

2.7. Statistical analysis

Principal coordinates analysis (PCoA) with weighted UniFrac distances, multi-response

permutation procedures (MRPP) and indicator species (IS) analyses of 16S rRNA gene sequences generated by assembled paired-end Illumina reads were performed using automated exploration of microbial diversity (AXIOME) automation of PANDAseq (Masella et al., 2012), the QIIME pipeline (Caporaso *et al.*, 2010) and custom AXIOME analyses (Lynch *et al.*, 2013).

2.8. MG-RAST analysis and CAZy Annotation

Paired-end shotgun sequences from the pooled heavy DNA samples described above were analyzed for the presence of GHs using the MG-RAST pipeline (Meyer *et al.*, 2008). Reads were annotated by comparison to sequences in the UniProt database (Apweiler *et al.*, 2004), with no maximum e-value cut off, a 54% minimum percentage identity cutoff and a 30 bp minimum alignment length cutoff. Using custom Perl scripts, annotated by Swiss-Prot and Trembl databases (UniProt release 2012-4), hits were paired with matching GH family CAZy identifiers by comparing an extracted database of UniProt accession numbers to CAZy identifiers. Matches to UniProt sequences were annotated by GH family.

2.9. Cellulose-enriched metagenomic library construction

High molecular weight DNA (40-50 kb) was extracted from all three soil samples that were amended with ^{13}C-labeled cellulose (week 6 time point), using a gentle enzymatic lysis (Zhou *et al.*, 1996). Crude DNA was purified from humic acids using the Aurora (Boreal Genomics; Vancouver, BC) by using one wash cycle (70 V/cm, 10°C, 90 min) and two concentration cycles (70 V/cm, 10°C, 60 min). DNA was visualized using a 1% agarose gel and quantified with the NanoDrop 2000. Samples were subjected to cesium chloride density

gradient ultracentrifugation and fraction collection as described above with minor modifications to the DNA precipitation step to avoid potential problems recovering high molecular weight DNA from cesium chloride by PEG precipitation. Specifically, gradient fractions were diluted with one volume of water and two volumes of ethanol and incubated overnight at -20°C. DNA was precipitated by centrifugation for 30 min at 13,000 x g and the supernatant decanted. The DNA was dissolved in 300 µl of water and then precipitated by adding 1/10 volume of 3M sodium acetate and two volumes of ethanol. We used DGGE, as described above, to profile all fractions, confirming that the fingerprints generated from an alternative lysis protocol were the same as those observed previously. Subsequently, pooled samples and fractions for large-insert cosmid cloning were mixed in the same equal ng ratio used to prepare template for sequence-based metagenomics.

2.9.1. Cosmid library construction

To increase the amount of DNA for ^{13}C-cellulose enriched metagenomic library construction, triplicate multiple displacement amplification (MDA) reactions were conducted on the pooled cellulose sample using the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare; Mississauga, ON), according to the manufacturer's instructions. Each reaction included ~7 ng of DNA template in order to minimize potential amplification bias (Binga, Lasken, & Neufeld, 2008), yielding ~3-4 µg of DNA. Positive control reactions with kit-supplied DNA and no-DNA negative controls were run in parallel. All MDA reaction products were quantified on a 1% agarose gel and triplicates pooled.

In order to purify MDA-amplified DNA from residual Φ29 DNA polymerase, 100 µl of DNA solution was mixed with 613 µl of TE, 73 µl of 10X gel loading buffer and 6.8 µl of 20%

SDS. After heating at 65°C for 10 min, the sample was left on ice for 5 min and then centrifuged at 15,900 × g for 5 min to pellet protein. The DNA-containing supernatant was loaded onto a 1% pulsed-field agarose gel (with TAE buffer) in order to size-select metagenomic DNA. The pulsed field gel electrophoresis (PFGE) was run on a CHEF Mapper (Bio-Rad) at 14°C, 5.5 V/cm, an angle 120° and initial 1.0 – final 6.0 sec switch time for 20 h. The outer sides of gel containing size marker and a small portion of DNA sample were cut off and post-stained with SYBR green. A gel slice corresponding to 30-75 kb DNA was excised without UV exposure, followed by DNA electroelution and concentration. Following end-repair using End-It DNA End-Repair kit (Epicentre), DNA was ligated to *Eco*72I-digested and SAP-treated vector pJC8. The ligation reaction mixture was packaged *in vitro* (Gigapack III XL Packaging Extracts, Stratagene) and then transduced into *Escherichia coli* HB101. Following overnight selection on LB-Tc agar plates, the resulting recombinant cosmid clones were pooled and saved in 7% DMSO in 1 ml aliquots at -75°C. Prior to pooling, a random selection *E. coli* clones from the plates were collected for analysis of cosmid DNA restriction patterns and arrayed into 96-well plates for functional screening. The average sizes of cloned metagenomic DNA and coverage of bacterial genomes were calculated based on sizes of *Eco*RI-*Hin*dIII fragments and the number of recombinant library clones.

2.10. Functional screening

The 2876 randomly selected clones were used for functional screening using 0.2% carboxymethylcellulose (CMC) as a substrate and Congo Red dye application to detect substrate degradation (Teather & Wood, 1982). Minor protocol modifications included omitting plate overlays, using one week incubation times at 37°C and washing colonies from

plates prior to Congo Red staining (0.1% aqueous solution).

In addition to a CMC screen, these same library clones were grown in LB-Tc (10 μg/ml) overnight at 37°C. The following day, a 96-well deep well plate with Terrific Broth (TB) and 10 μg/ml tetracycline was inoculated and incubated at 37°C for 24 to 48 hours. Cell pellets were collected and frozen following culturing. For lysis, cell pellets were thawed and chemically lysed using BugBuster Protein Extraction Reagent (Novagen). Clones were characterized using a panel of 4-methylumbelliferone-based (4-MU) fluorogenic substrates. All clones were tested in a first round of screening on all substrates. Clones that demonstrated activity on one or more substrates were cultured again and rescreened on appropriate substrates to eliminate false-positive reactions. Reactions were carried out in opaque 384-well microplates. Library lysates were incubated with 0.1 mM of 4-MU-substrates for 1 h at 50°C in a 40 μL sodium citrate-buffered (50 mM, pH 5) reaction. Reactions were quenched by the addition of 40 μL of 0.2 M glycine (pH 10). Fluorescence was detected at 445 nm following excitation at 370 nm. Substrates screened included α-L-arabino-furanoside/pyranoside, β-D-cellobio-pyranoside, β-D-glucopyranoside, β-D-xylopyranoside and N-acetyl-β-D-galactosaminide.

The average sizes of positive clones were calculated based on sizes of *Eco*RI-*Hin*dIII-*Bam*HI fragments.

2.11. Analysis of end sequences

Cosmid end-sequences were generated by Sanger sequencing of extracted positive hit cosmids, with M13 forward (5'- CACGACGTTGTAAAACGAC - 3') and M13 reverse (5' - GGATAACAATTTCACACAGG - 3') primers flanking the site of metagenomic DNA

insertion. For each clone, two end-sequences were obtained by The Centre for Applied Genomics (TCAG) and are referred to as “reverse” and “forward” reads. We used blastx searches of translated nucleotide sequences against an NCBI protein database and tblastx searches translated nucleotide sequences against a translated nucleotide database. End sequences were deposited in Genbank with accession numbers KG771718-KG771732. Posterior BLAST analysis was done searching for sequences similarities in the three libraries: Low pH, Agricultural and cellulose (forward and reverse). Sequences with >95% similarity and >30 bp were recorded as positive matches.

3.0. Results and discussion

3.1. Characterization of active soil Bacteria

We used DNA-SIP as a targeted approach for identifying active soil microorganisms involved in the metabolism of five plant-derived carbohydrates (glucose, cellobiose, xylose, arabinose and cellulose) in three disparate soil samples. Soil samples were selected from the CM^2BL soil collection based on maximizing physicochemical diversity, encompassing a wide range of pH, geographical distance (e.g., latitude) and land usages. Physicochemical analyses of these three soil samples revealed that soils differed in texture, carbon content, bulk density and pH (Table 2). In particular, the pH was lower for the Arctic tundra and temperate rainforest soil samples, which suggested that microbial composition and diversity of these two samples would be fundamentally different than in the agricultural soil (Stotzky, 1997; Fierer & Jackson, 2006). We maintained the water-filled pore space (WFPS) between 50% and 60% to maximize microbial activity, avoiding decreased aerobic microbial activity at WFPS values >60% (Linn & Doran, 1984; Franzluebbers, 1999).

Table 2. Location and physicochemical characteristics of the soil samples selected for DNA stable-isotope probing incubations. Characterization was done according with the Canadian system of soil classification (NRC-CNR, 1998).

Samples	Description	Location	Bulk density (g/cm^3)	Total carbon (% dry)	Inorganic carbon (% dry)	Organic carbon (% dry)	pH	Moisture (% dry)	Nitrogen (% dry)	Soil type
Arctic tundra (1AT)	Daring Lake, North West Territories	64°52'N 111°35'W	0.21	46.9	0	46.9	3.9	417.7	1.42	Organic
Temperate rainforest (7TR)	Pacific coastal rainforest, Vancouver Island	48°36'N 124°13'W	0.62	10.8	0	10.8	4.9	69.8	0.35	Coarse sandy loam
Agricultural soil - wheat (11AW)	Elora Research Station, Ontario	43°38'N 80°24'W	1.1	1.85	0.12	1.7	7.4	17.9	0.19	Silt loam

Labeled and unlabeled substrates were added to the soil samples in multiple doses over the period of one and three weeks (glucose, cellobiose, xylose and arabinose) or three and six weeks (cellulose; Table 1). The cellulose produced as substrate for SIP incubations by *Gluconacetobacter xylinus* was predominantly amorphous cellulose (Koizumi *et al.*, 2008), which is more readily degraded than crystalline cellulose (Hall *et al.*, 2010), likely because CBMs are not required for the degradation of amorphous cellulose (Wilson, 2011). Although substrate concentrations were likely several times higher than typical bulk soil concentrations (Medeiros *et al.*, 2006; Hill *et al.*, 2008) higher sugar substrate concentrations would be expected in the root rhizosphere and in areas of active plant matter decomposition (as reviewed in Hill *et al.*, 2008), suggesting that our incubation conditions would not be unrealistic for some naturally occurring soils. In addition, these concentrations were chosen because it is critical that labeled isotope be more abundant than exogenous soil carbon sources for the success of DNA-SIP, enabling the separation and recovery of labeled DNA for subsequent molecular analyses (Neufeld *et al.*, 2007a; Neufeld *et al.*, 2007b). A previous proof of principle used similar substrate concentrations and incubation times with glucose and cellulose, demonstrating minimal-yet-detectable labeling of DNA in an Arctic tundra soil sample (Pinnell *et al.*, 2013).

3.2. Carbohydrate metabolism

Metabolism of labeled substrates in DNA-SIP incubations was confirmed by higher headspace CO_2 production in all substrate-amended flasks, compared to uninoculated controls, for each of the three soils (Figure 1). In all cases, cellulose-amended flasks demonstrated reduced CO_2 production compared to the other substrates, further justifying an extended

incubation time for this comparably recalcitrant substrate. The average amount of CO_2 released after six days was 13% of headspace gases (cellulose not included), which, after subtracting the average CO_2 produced in uninoculated flasks, was approximately equivalent to 1.4 mmol of carbon. This represents 93% of the total carbon added per week, respectively (1.5 mmol of carbon).

In addition to monitoring CO_2 production in all flasks, separate soil incubations were prepared with a defined helium/oxygen headspace and glucose amendment, in order to monitor maximal rates of O_2 consumption. The addition of glucose stimulated O_2 consumption but the headspace remained oxic for each of the weekly incubation periods over the first three weeks (Fig. 2), indicating that weekly aeration of experimental flasks was sufficient to deplete CO_2 and replenish O_2. Maintaining oxic conditions was important to ensure that the DNA-SIP incubation recovered DNA from microorganisms involved in degradation of complex carbohydrates like cellulose under oxic conditions, in addition to capturing DNA from microorganisms involved in anaerobic metabolism (Leschine, 1995). Indeed, recent oxic incubations demonstrated activity of anaerobic *Clostridia* (Schellenberger *et al.*, 2010; Pinnell *et al.*, 2013; Ronan *et al.*, 2013), presumably because anoxic microenvironments exist even within oxic experimental microcosms.

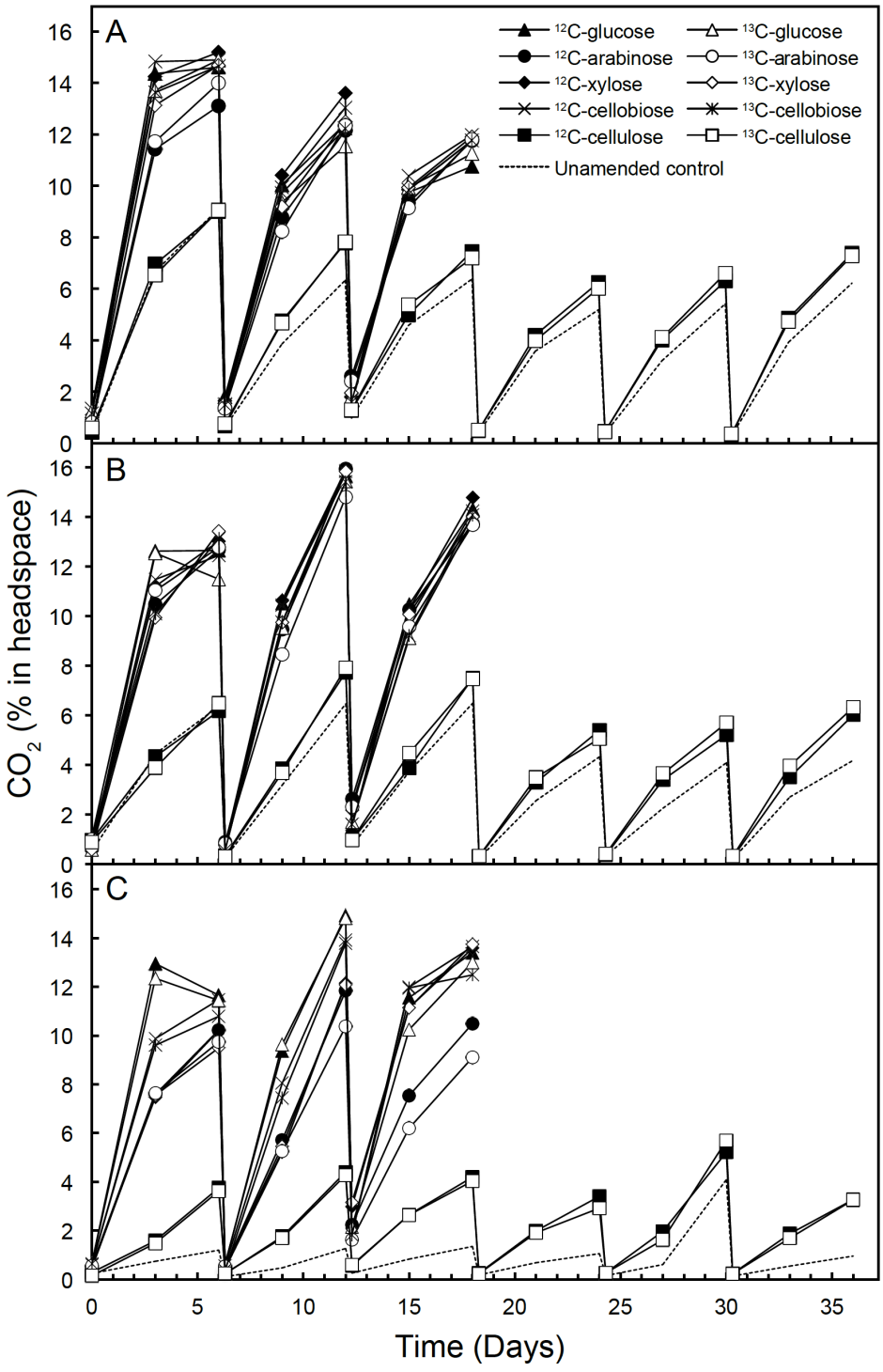

Figure. 1. Carbon dioxide production for Arctic tundra (1AT; A), temperate rainforest (7TR; B) and agricultural (11AW; C) soils. Soil samples were amended with labeled (^{13}C) or unlabeled (^{12}C) substrates. The "control" represents a soil sample incubated without substrate.

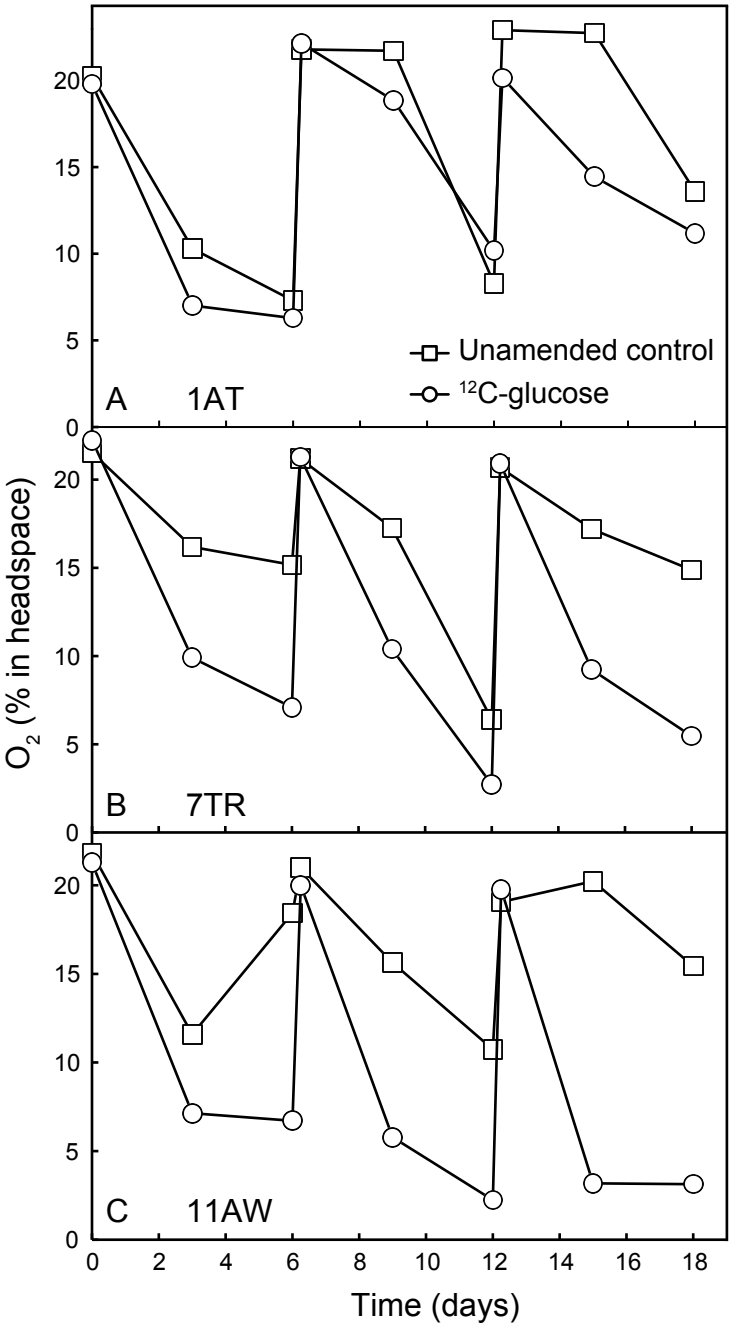

Figure 2. Oxygen concentrations in soil incubations, with and without added glucose. Headspace was flushed with helium and amended with oxygen at weekly intervals for Arctic tundra (1AT; A), temperate rainforest (7TR; B) and agricultural (11AW; C) soils.

3.3. Confirmation of isotope labeling

For the two time points of all incubations (Table 1), DNA was retrieved for the analysis of bacterial community composition by DGGE, targeting the 16S rRNA gene V3 region (Green *et al.*, 2010). Parallel ^{12}C and ^{13}C incubations were used to ensure that isotopic enrichment of nucleic acids occurred, by the demonstration of distinct fingerprint profiles in heavy fractions for ^{13}C-incubated samples, but not for the corresponding ^{12}C controls (Neufeld *et al.*, 2007b). All DNA extracts from microcosm soils were subjected to density gradient ultracentrifugation and recovered in twelve fractions, which were visualized in agarose gels. Agarose gels for the one-week time points exhibited labeled DNA in ^{13}C-incubated heavy fractions (e.g., fractions 1-7), compared to ^{12}C-control fractions, for glucose, cellobiose, arabinose and xylose SIP incubations in temperate rainforest and agricultural soils. Labeled DNA for Arctic tundra soil was not clearly visualized at this time point of incubation (Figs. 3, 5, 7, 9). The three-week time points demonstrated that all soils possessed more DNA than the one-week time point in ^{13}C-incubated heavy fractions compared to ^{12}C-control fractions, for glucose, cellobiose, arabinose and xylose SIP incubations (Figs. 4, 6, 8, 10). For cellulose, only temperate rainforest and agricultural soil incubations resulted in visible DNA in agarose gels corresponding to ^{13}C-sample heavy fractions for both three and six-week time points (Figs. 11,12) At earlier time points the labeled DNA associated with heavy fractions for ^{13}C-incubated samples was lower compared with the later time points.

Bacterial DGGE fingerprints from early time points fractions (Figs. 3, 5, 7, 9, 11) compared with corresponding late time point fractions, demonstrated unique patterns associated with the heavy fractions (e.g., fractions 1-7) for all ^{13}C-incubated SIP microcosms (Figs. 4, 6, 8, 10, 12). Although some cross-gradient fingerprint variations were associated

with ^{12}C control DNA, these differences were likely GC-content shifts, being pronounced only in the lightest fractions (e.g., fractions 10-12) and distinct from shifts associated with fractionated ^{13}C DNA. Substrate- and soil-specific heavy fraction patterns were consistent for early and late time-point samples (Figs. 3-12), which indicated that detected active bacteria were consistent over time rather than changing due to food web dynamics (Neufeld *et al.*, 2007a).

Heavy DNA fingerprints were used to identify fractions containing ^{13}C-labelled DNA for subsequent 16S rRNA gene sequencing, bulk DNA sequencing and functional metagenomics. Based on DGGE patterns, we identified fractions 5 and/or 6 as being representative of heavy DNA and fraction 10 as representing light DNA for all soils, substrates and incubation times (Figs. 4, 6, 8, 10).

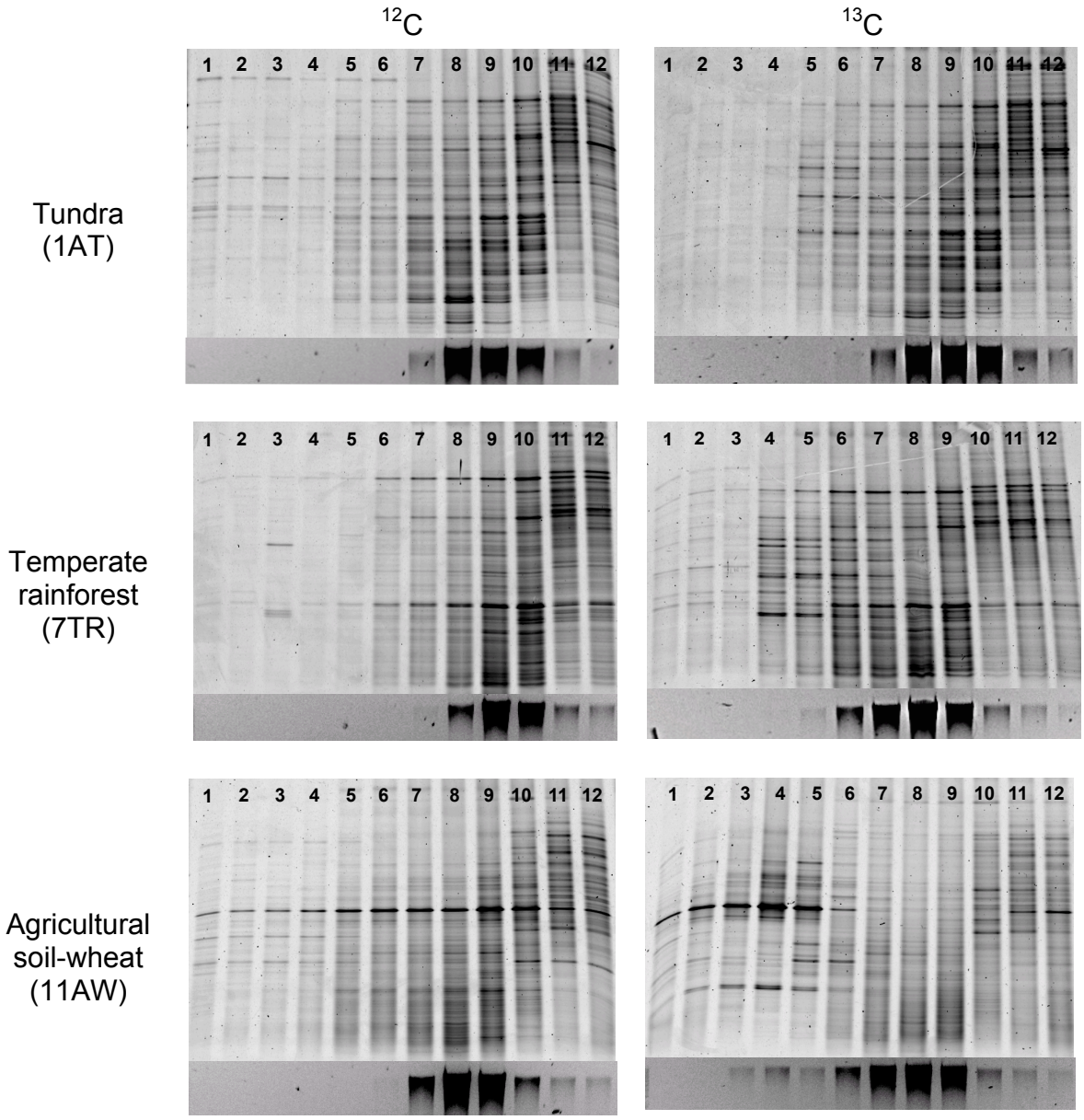

Figure 3. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **glucose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**1 week of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

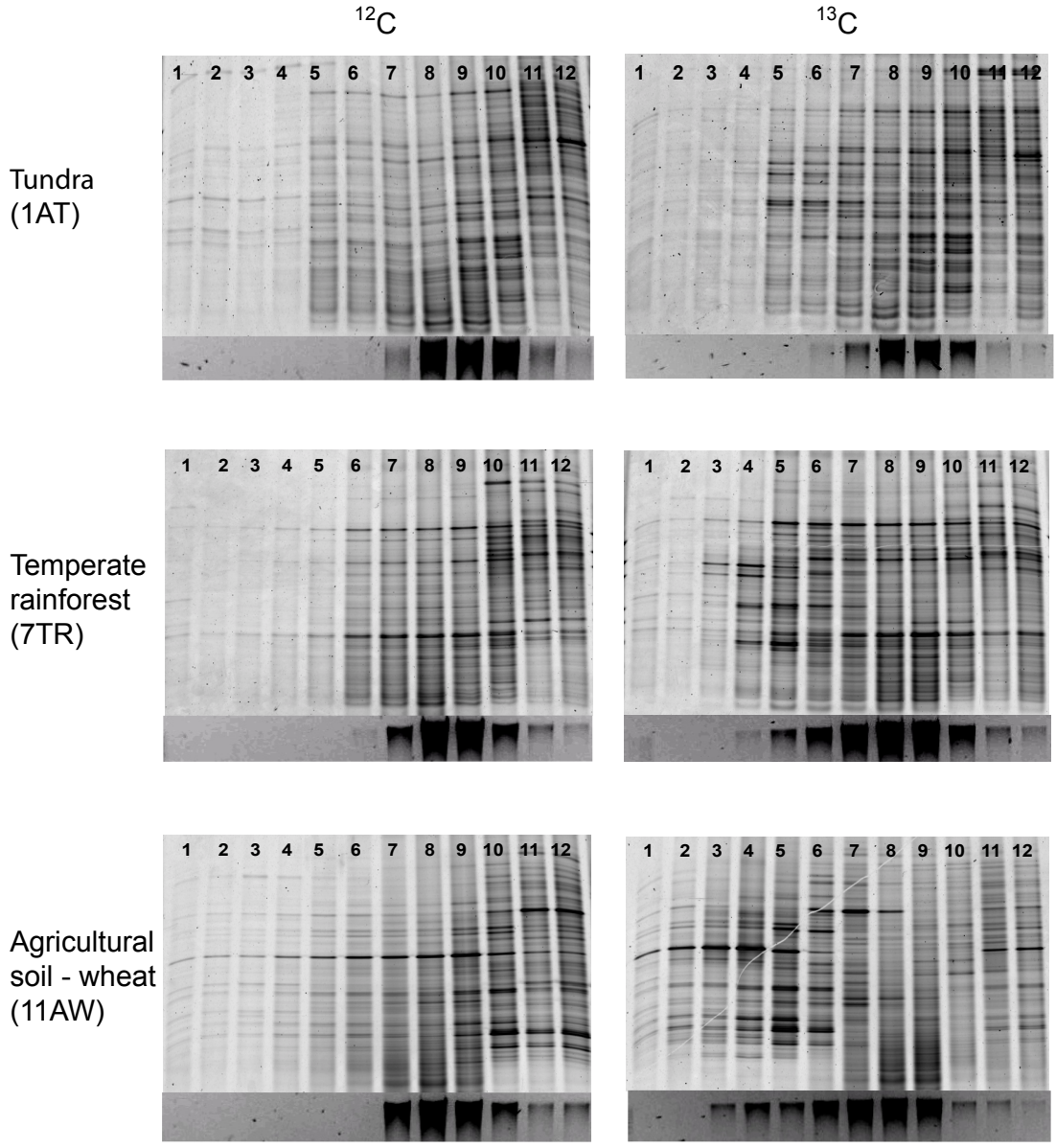

Figure 4. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **glucose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**3 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

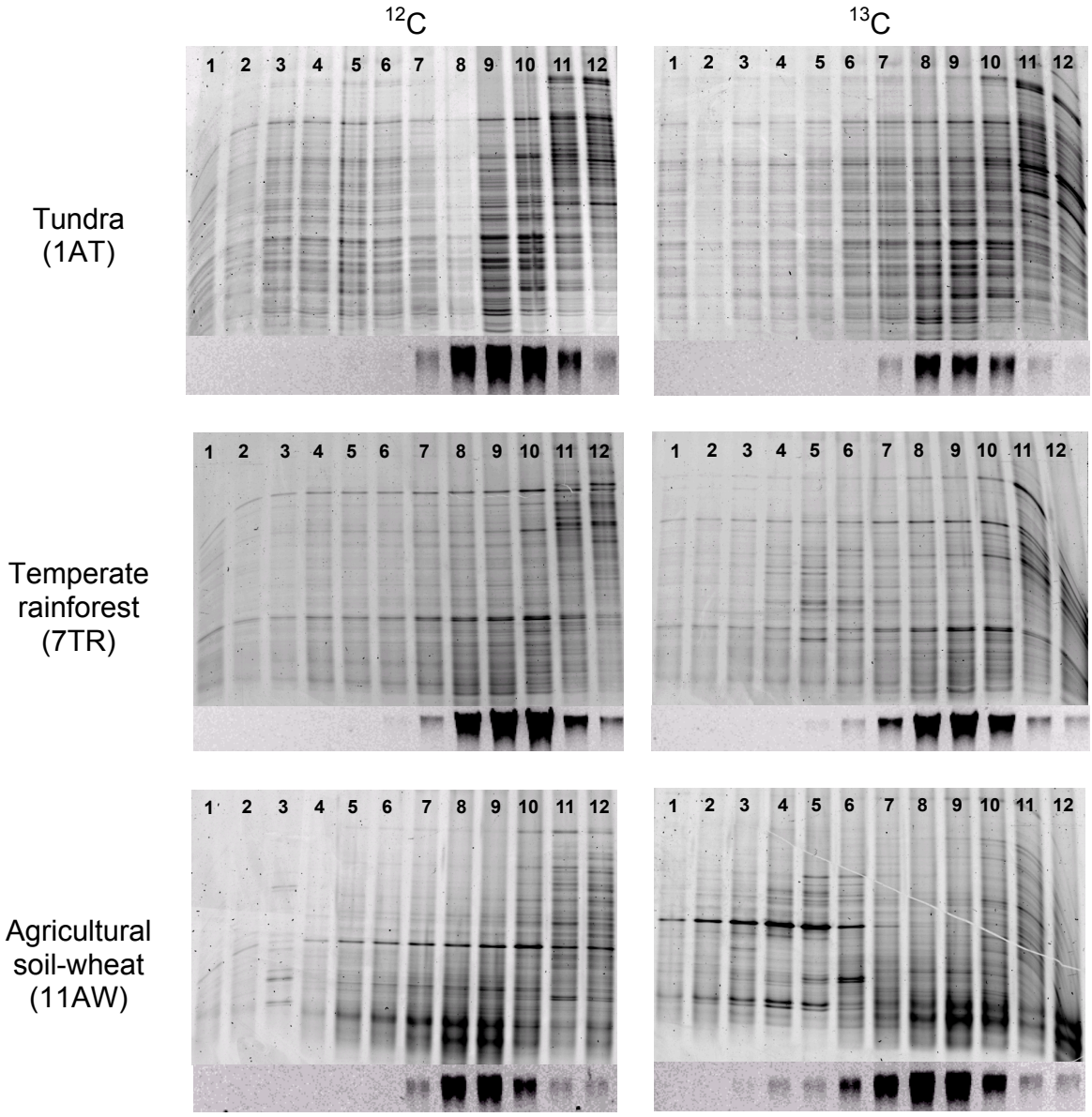

Figure 5. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **cellobiose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**1 week of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

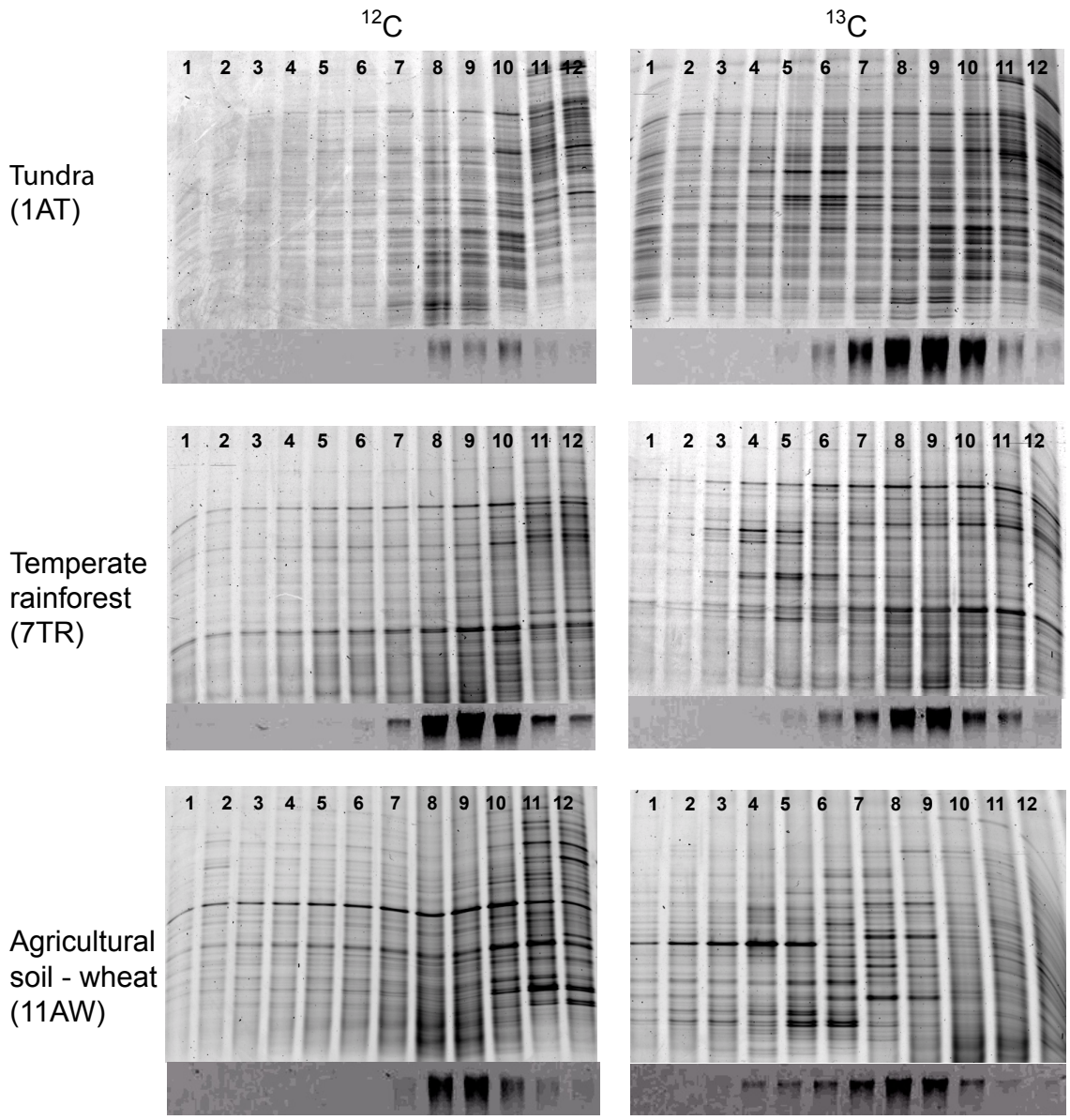

Figure 6. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **cellobiose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**3 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

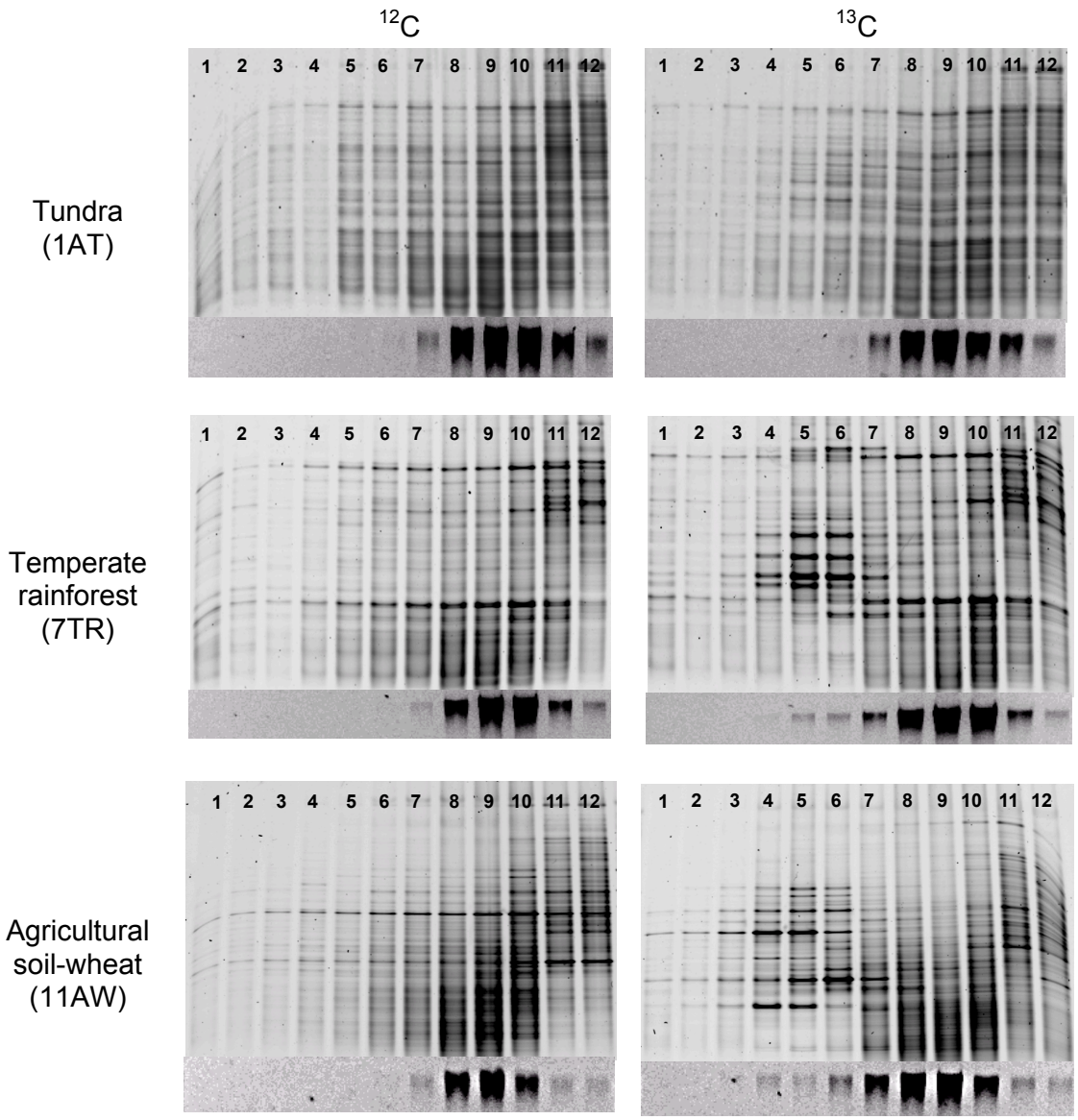

Figure 7. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **arabinose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**1 week of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

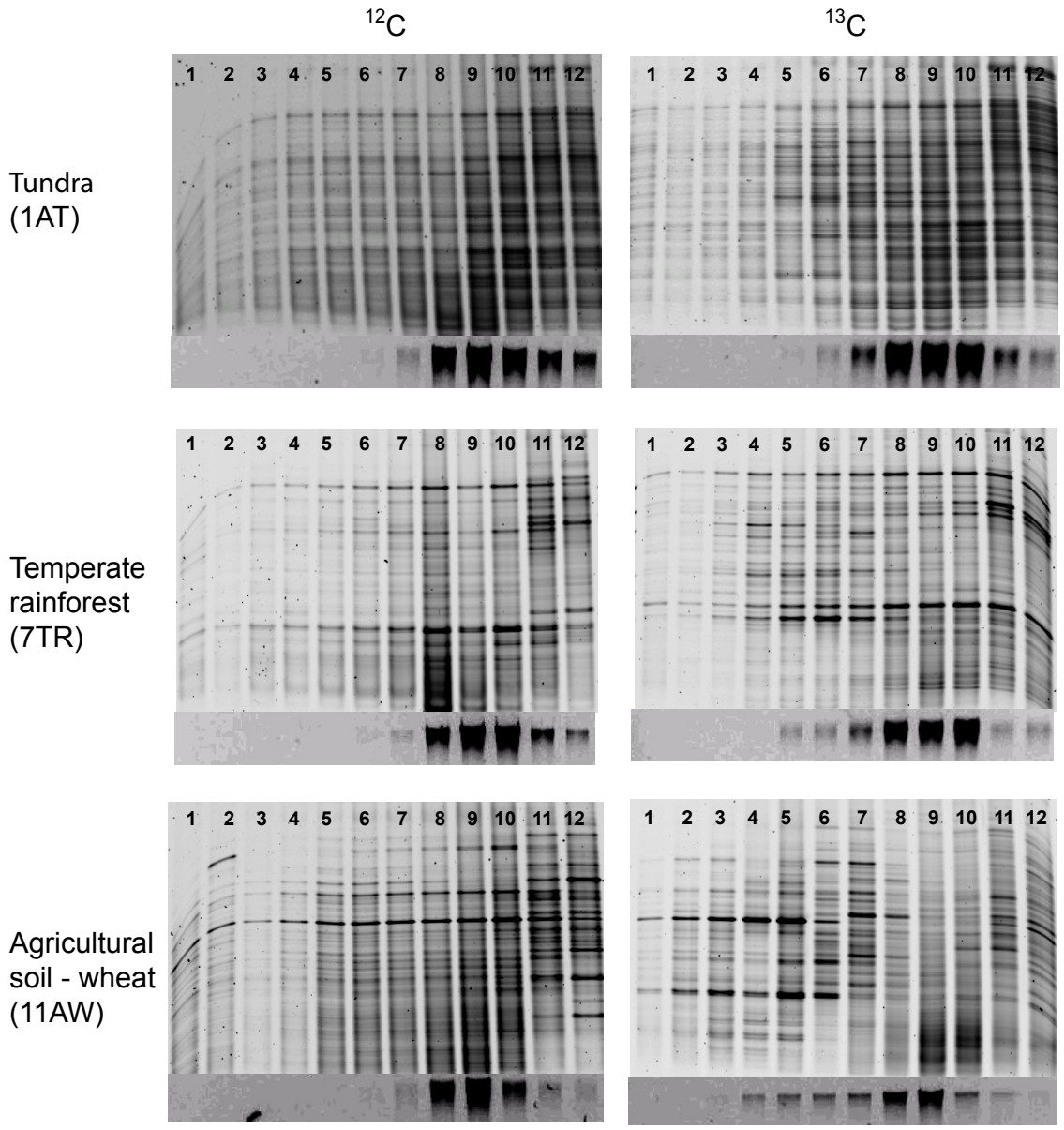

Figure 8. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **arabinose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**3 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

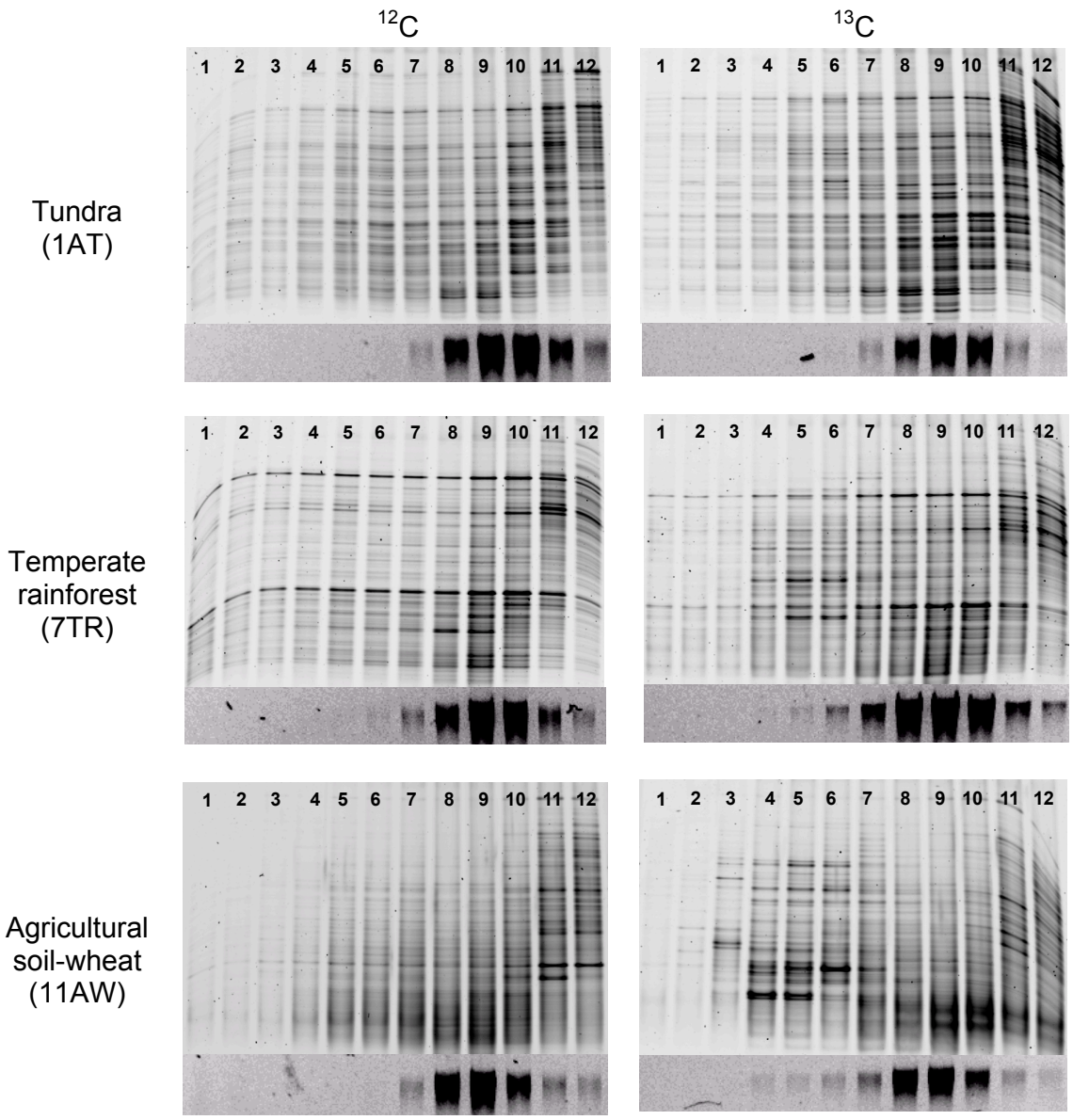

Figure 9. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **xylose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**1 week of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

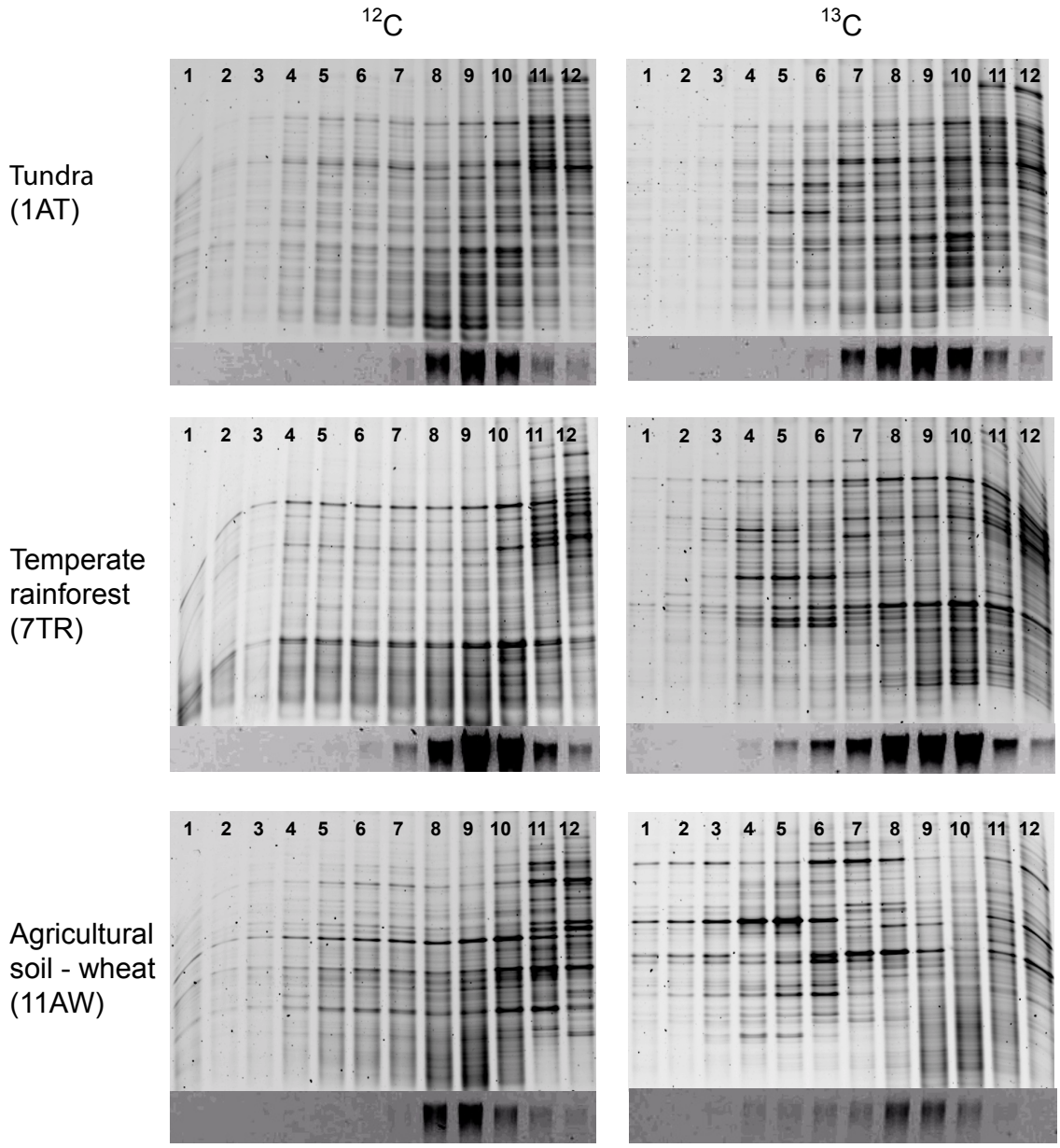

Figure 10. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **xylose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**3 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

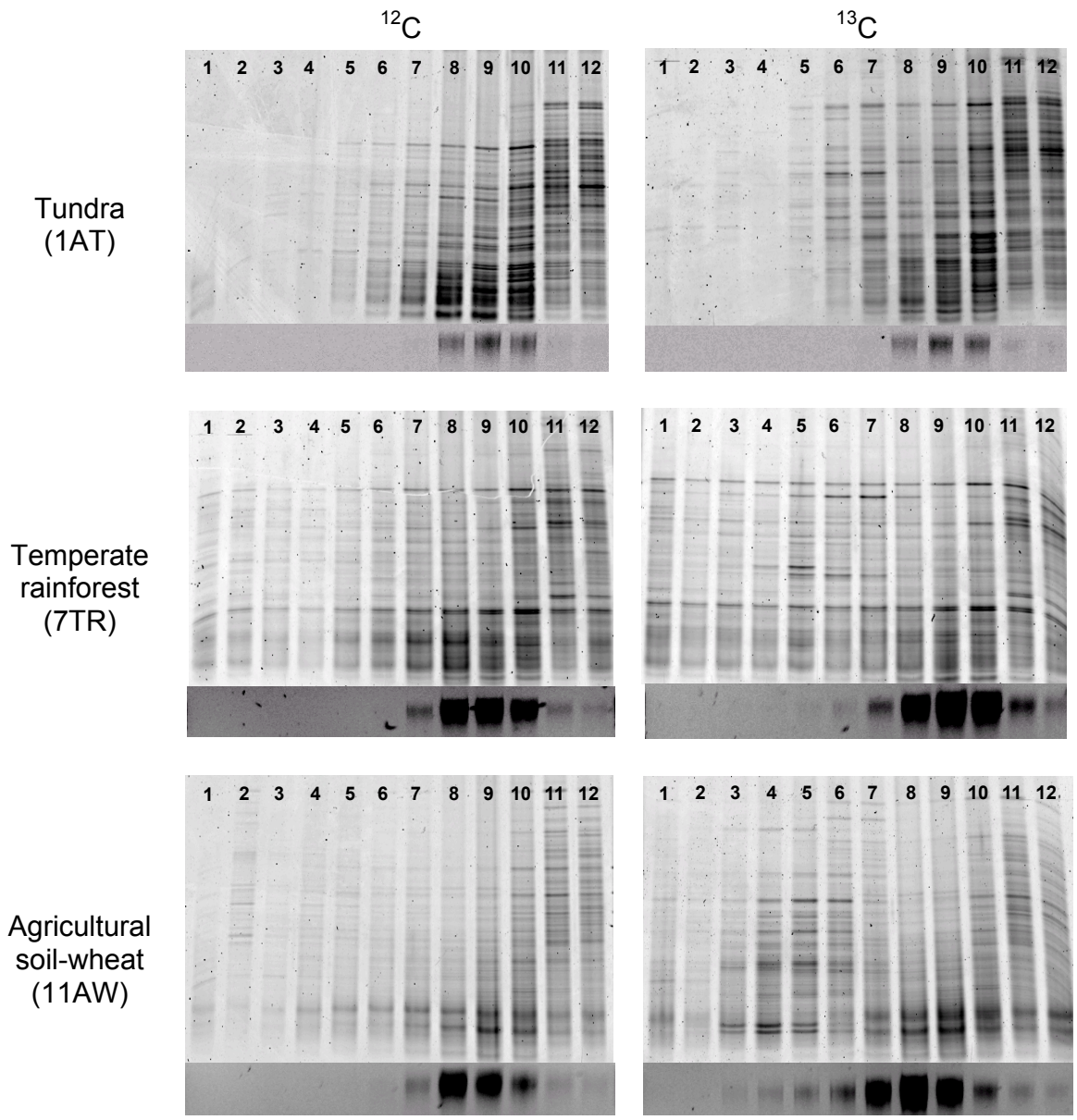

Figure 11. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **cellulose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**3 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

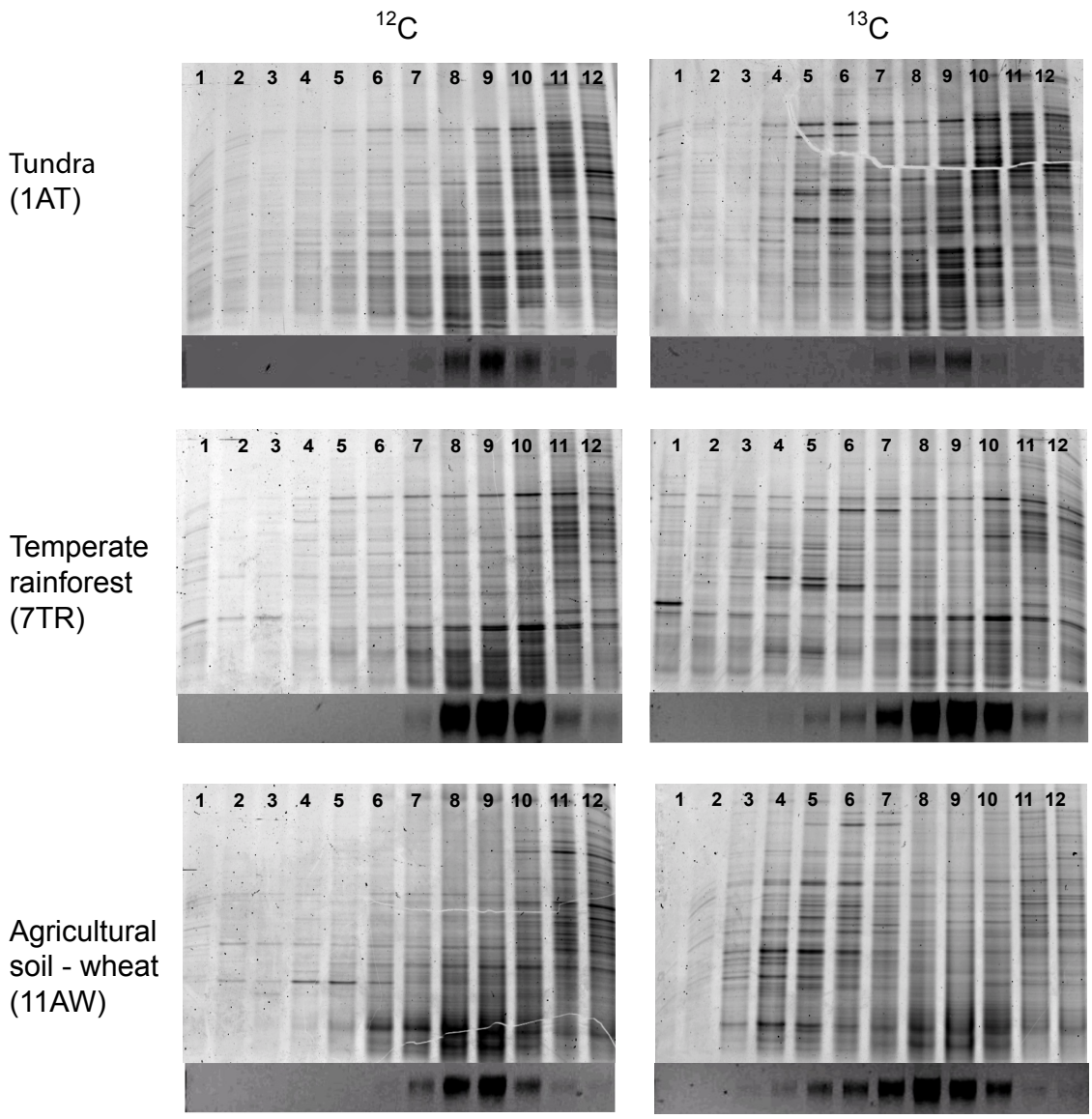

Figure 12. Bacterial denaturing gradient gel electrophoresis fingerprints of density gradient fractions recovered after **cellulose** DNA-SIP incubation of arctic tundra, temperate rainforest and agricultural soils (**6 weeks of incubation**). Both ^{12}C (unlabeled) incubations and ^{13}C (labeled) incubations are shown. The gel strips shown beneath each DGGE are 1% agarose gels stained with ethidium bromide to demonstrate isopycnic separation of DNA within the cesium chloride density gradient.

3.4. Taxonomic characterization of heavy DNA

Fractions were selected from soils, substrates and incubation times, for 16S rRNA gene profiling targeting the bacterial V3 region. Based on DGGE data, we selected fractions 6 (heavy) and 10 (light) for Arctic tundra and fractions 5 (heavy) and 10 (light) for temperate rainforest and the agricultural soil. In addition, we sequenced 16S rRNA genes from DNA extracted from the initial soil samples used to establish SIP incubations, to determine whether light fractions resembled the original soil community as expected. Following paired-end read assembly, we analyzed 630,000 sequences (10,000 reads per sample) using an AXIOME management of QIIME (Caporaso *et al.*, 2010; Lynch *et al.*, 2013) and custom analyses. Beta diversity assessed by weighted UniFrac distances (Lopuzone & Knight, 2005) visualized within principal coordinate analysis (PCoA) plots showed that all samples from within each of the three soil treatments were clustered distinctly according to soil type (Fig. 13A), which was significant based on mixed-response permutation procedures analysis (MRPP; A=0.18, T=-20.4, $p<0.001$). Both the Arctic tundra and temperate rainforest soil profiles clustered more closely to one another, which is likely a result of both soils sharing a low pH (Table 2), a major determinant of soil bacterial diversity and taxonomic composition (Lauber *et al.*, 2009; Bartram *et al.*, 2013). In addition, all heavy and light fraction profiles for the three soils were clustered distinctly (Fig. 13A), which was also highly significant (A=0.40, T=-28.3, $p<0.001$). Native soil phylogenetic profiles clustered with their respective light fractions, as was expected, indicating that the background bacterial community was consistent following SIP incubation. Although the two soil collection time points (Table 1) for some ^{13}C-labelled substrates clustered together (Fig. 13B), the differences between heavy and light fractions were much greater than those observed between the five substrates used in this study.

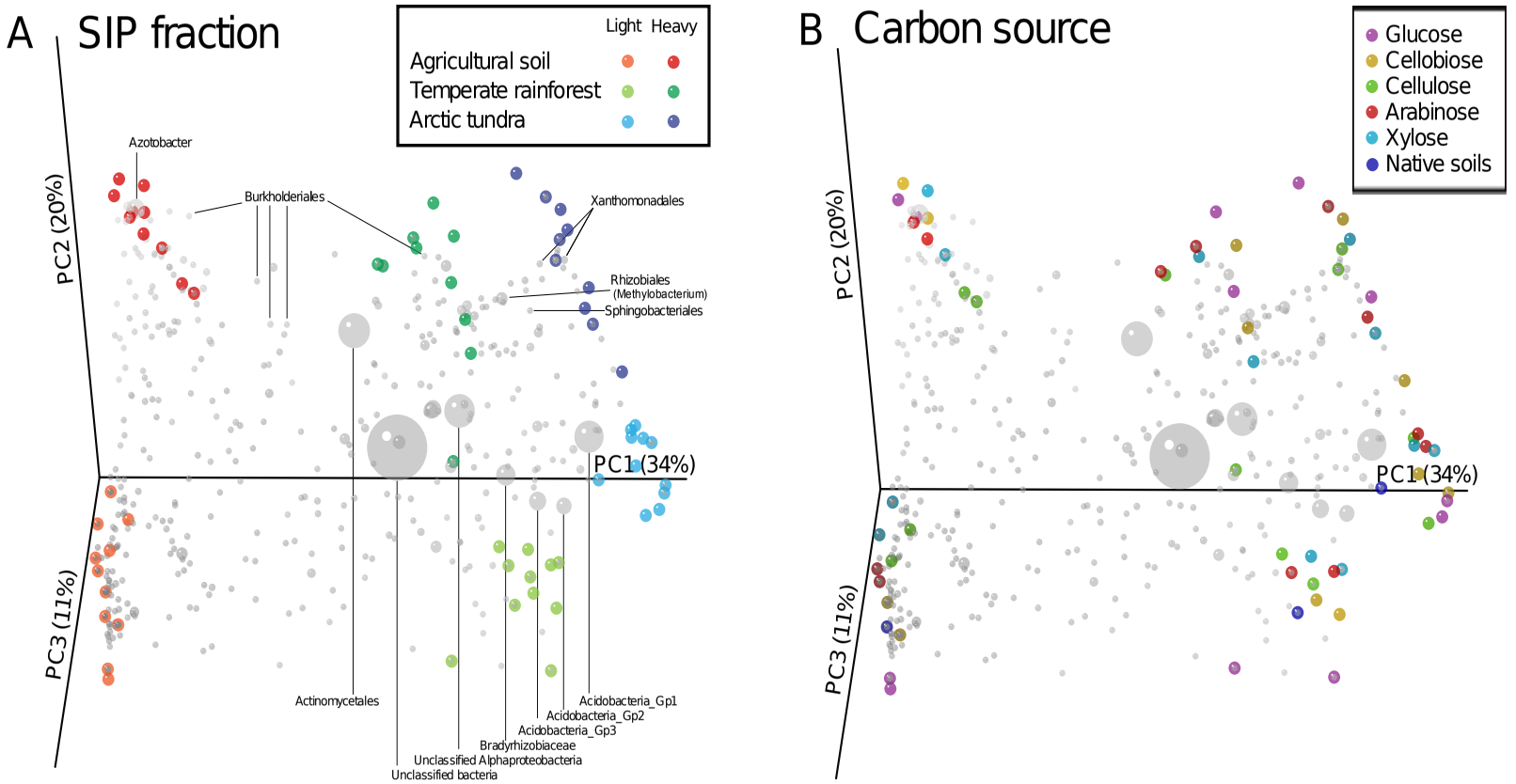

Figure 13. Principal Coordinates Analysis (PCoA) biplots of weighted UniFrac distances of 16S rRNA gene sequences generated by assembled paired-end Illumina reads. (A) Samples separated by soil type and fraction and major taxonomic groups associated with each soil are shown. (B) Samples separated by carbon source. Native soils associated with their respective light fractions. OTUs associated are shown in grey.

Many taxonomic groups were affiliated with heavy DNA, light DNA and with each of the soil types (Fig. 13A). We used indicator species analysis (Dufrene & Legendre, 1997) with an indicator value (IV) threshold of 0.7 and 250 minimum sequence sum to assess the strongest operational taxonomic units most significantly associated with: a) all heavy DNA samples (versus all light DNA samples), b) all heavy DNA samples within each soil type (versus all light DNA for that same soil type), c) each individual substrate across all heavy DNA samples from all soil types (versus the heavy DNA for the other substrates from all soil types) and d) each substrate from within each soil type's heavy DNA (versus the other substrates for that same soil type).

When we compared OTUs associated with all heavy DNA samples versus all light DNA samples from all soils, the indicator species analysis revealed an overall strong presence of the genera *Salinibacterium* (*Actinobacteria*), *Devosia* (*Alphaproteobacteria*), *Telmatospirillum* (*Alphaproteobacteria*), *Phenylobacterium* (*Alphaproteobacteria*) and *Asticcacaulis* (*Alphaproteobacteria*) genera. Also, unclassified members of the uncultured *Alphaproteobacteria* Ellin329 were present in heavy fractions from all soils (Appendix A1, Fig. 14).

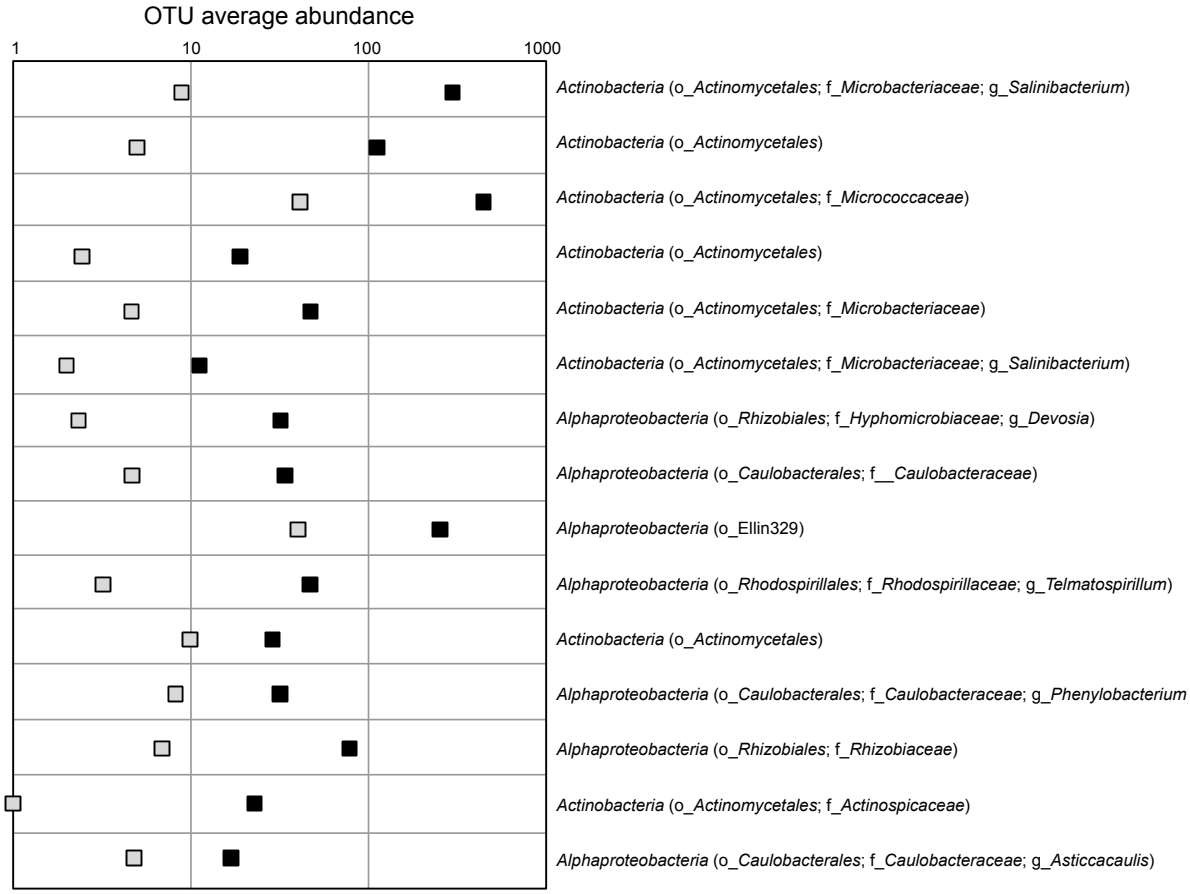

Figure 14. Indicator species associated with all heavy DNA samples from all soils. Average number of OTUs associated with light DNA samples (grey squares) are shown, in addition to the average abundance of indicator species across all heavy DNA samples (black squares).

The indicator species analysis from all heavy DNA samples versus all light DNA within each soil type showed that the predominant genera identified in heavy fractions from Tundra soil (1AT) were *Salinibacterium* (*Actinobacteria*), *Rhodanobacter* (*Gammaproteobacteria*), *Conexibacter* (*Actinobacteria*), *Telmatospirillum* (*Alphaproteobacteria*), *Asticcacaulis* (*Alphaproteobacteria*) and *Burkholderia* (*Betaproteobacteria*), unclassified members of the uncultured *Alphaproteobacteria* Ellin329 and unclassified members of the orders *Sphingomonadales, Acidobacteriales*, *Xanthomonadales*, *Solirubrobacterales*, *Rhodospirillales*, *Caulobacterales*, *Burkholderiales* and *Actinomycetales* were also predominant in this soil (Appendix A2). The temperate rainforest soil (7TR) heavy DNA was dominated by OTUs classified to the genera *Paucibacter* (*Betaproteobacteria*), *Burkholderia* (*Betaproteobacteria*), *Spirochaeta* (*Spirochaetes*), *Salinibacterium* (*Actinobacteria*), *Telmatospirillum* (*Alphaproteobacteria*), *Labrys* (*Alphaproteobacteria*), *Mesorhizobium* (*Alphaproteobacteria*) and *Phenylobacterium* (*Alphaproteobacteria*). Also, members of the uncultured *Alphaproteobacteria* Ellin329, *Betaproteobacteria* Ellin6067, unclassified members of the orders *Rhodospirillales*, *Caulobacterales*, *Burkholderiales*, *Actinomycetales, Rhizobiales* and uncharacterized genera from other phyla such as the *Verrucomicrobia* were found (Appendix A2). The agricultural-soil wheat (11AW) heavy DNA OTUs were represented by the genera *Pseudomonas* (*Gammaproteobacteria*), *Devosia* (*Alphaproteobacteria*), *Pseudoxanthomonas* (*Gammaproteobacteria*), *Salinibacterium* (*Actinobacteria*), *Ramlibacter* (*Betaproteobacteria*), *Ochrobactrum* (*Alphaproteobacteria*), *Paenibacillus* (*Firmicutes*) and *Aeromicrobium* (*Actinobacteria*) and further unclassified members of the orders *Pseudomonadales*, *Rhizobiales*, *Caulobacterales*, *Actinomycetales* and *Burkholderiales* (Appendix A2).

To investigate microorganisms related to the metabolism of the substrates, comparisons of each individual substrate across all heavy DNA samples from all soil types versus the heavy DNA for the other substrates from all soil types were made. The results showed that the orders associated with the metabolism of cellulose were dominated by *Actinomycetales* and *Caulobacterales* (genus *Phenylobacterium*; Appendix A3). Members of the *Alphaproteobacteria* were associated with the metabolism of arabinose, while members of the order *Rhizobiales* were strongly associated with the metabolism of xylose. There were no indicator species associated with the metabolism of glucose and cellobiose (Appendix A3). Comparing heavy DNA from each substrate from within each soil versus the other substrates for the same soil type, the predominant indicator species for the agricultural soil at the taxonomic level of genus associated with the metabolism of glucose was *Paenibacillus* (*Firmicutes*), *Mezorhizobium* (*Alphaproteobacteria*) and *Devosia* (*Alphaproteobacteria*; Appendix A4). The predominant indicators for cellulose in this soil were *Cellvibrio (Gammaproteobacteria*), unclassified members of the order *Sphingomonadales* and *Actinomycetales* (Appendix A4). In the temperate rainforest soil, the predominant order associated with the metabolism of cellulose was the *Myxococcales* (*Deltaproteobacteria*; Appendix A4). An OTU affiliated with *Caulobacterales* was associated with the metabolism of glucose in Arctic tundra. Nevskia (*Gammaproteobacteria*) and two OTUs affiliated with the *Acidobacteria* were associated with tundra cellulose assimilation (Appendix A4). No other OTUs were significant indicators for the remaining substrates (i.e., cellobiose, arabinose, xylose) for the three soils.

Although our DNA-SIP incubation revealed many poorly classified indicator taxa, many of the indicator species associated with the heavy DNA were expected based on previous

studies. For example, *Salinibacterium* was isolated from seawater samples, frozen soils from glaciers (Han *et al.*, 2003; Zhang *et al.*, 2008) and permafrost Antarctic soil (Shin *et al.*, 2012). Different species related to this genus utilize sucrose, glucose, cellobiose, D-mannose, melibiose, maltose, galactose, arabinose and fructose as sole carbon sources (Han *et al.*, 2003; Zhang *et al.*, 2008). Also, members of the genus *Devosia* were isolated from greenhouse soil and beach sediments. They were positive for the hydrolysis of aesculin, β-galactosidase, β-glucosidase and N-acetyl-β-glucosaminidase. However, they also tested negative for the hydrolysis of CMC (Yoo *et al.*, 2006; Lee, 2007). *Phenylobacterium* and *Burkholderia* were found among the most abundant taxa in *Picea abis* forest soils when communities derived from RNA were analyzed (Baldrian *et al.*, 2012). *Asticcacaulis* was identified in tundra wetland soils from samples taken from a depth of 3-6 cm. The species belonging to this genus were identified as aerobic chemoorganoheterotrophs able to use glucose, sucrose, xylose, maltose, galactose arabinose, lactose, fructose, rhamnose and threalose among other carbon sources (Vasilyeva *et al.*, 2006). The genus *Spirochaeta* was isolated from diverse environments, mainly from extremophilic aquatic environments. Some species from this genus are free-living saccharolytic and obligate or facultative anaerobes (Hoover *et al.*, 2003; Angelov *et al.*, 2011). *Spirochaeta americana* was reported as a consumer of D-glucose, fructose, maltose, sucrose starch and D-mannitol (Hoover *et al.*, 2003) and *Spirochaeta thermophila* was described as a cellulolytic organism; the study of its genome revealed a high proportion of genes encoding for more than 30 GHs (Angelov *et al.*, 2011). Species from the genus *Labrys* were found in different rhizosphere habitats, degrading various monosaccharides and disaccharides as sole carbon and energy source (Islam *et al.*, 2007). Additionally, Schellenberg and coworkers reported that in an agricultural soil (clay loam soil, pH 6.6), cellulose was metabolized by

Bacteroidetes, *Chloroflexi* and *Planctomycetes*; cellobiose and glucose were degraded predominantly by *Actinobacteria* (Schellenberger *et al.*, 2010). The results also suggested that cellulolytic bacteria were different from saccharolytic bacteria and that oxygen availability defined the different taxonomic groups involved. Other study showed that under anoxic conditions, cellulose was metabolized by *Actinobacteria*, *Bacteroidetes* and *Firmicutes*; cellobiose and glucose were degraded by *Firmicutes* and members of the *Burkholderiales*, *Caulobacteriales*, *Rhizobiales*, *Sphingobacteriales*, *Xanthomonadales* and group 1 *Acidobacteria* were associated with three different soils amended with cellulose (Eichorst & Kuske, 2012). A recent survey of active *Bacteria* in an Arctic tundra sample found *Clostridium* and *Sporolactobacillus* involved in ^{13}C-glucose assimilation and *Betaproteobacteria*, *Bacteroidetes* and *Gammaproteobacteria* involved in the assimilation of carbon derived from ^{13}C-cellulose (Pinnell *et al.*, 2013). Others have used SIP and labeled cellulose carbon to identify *Dyella, Mesorhizobium*, *Sphingomonas* and uncultured *Deltaproteobacteria* (affiliated with *Myxobacteria*) linked to cellulose degradation (el Zahar Haichar *et al.*, 2007).

3.5. MG-RAST analysis and Functional Annotation

We used next-generation sequence analysis of bulk DNA to survey the prevalence of annotated glycosyl hydrolases (GHs) within three pooled samples targeted for subsequent functional metagenomic screens. Using the UniFrac-based PCoA plot (Fig. 13), we pooled heavy DNA samples representing all substrates (except cellulose) associated with low pH (i.e., temperate rainforest, Arctic tundra), heavy DNA for all substrates (except cellulose) from the agricultural soil and the cellulose-enriched DNA from the three soils. Posterior analysis of paired-end reads was performed by MG-RAST using annotations derived from the Swiss-

Prot/Uniprot database. Only 19.4% (Low pH library) 19.6% (Cellulose library) and 22% (Agricultural library) of sequences were annotated by Swiss-Prot in MG-RAST using a threshold of e-value cutoff 0.01 and only a small percent of these sequences were annotated as GHs (Table 3), which is an important consideration for subsequent analysis of annotation data based on a minority of sequences. Nonetheless, using a custom perl script to convert Swiss-Prot annotations to CAZy GH identifiers, we detected differential abundances of 81 unique GH families for the pooled cellulose library and 80 GH families for each of the low pH and agricultural soil composite libraries. The distribution of annotated GHs varied between samples and the most abundant families in the three pooled samples were GH1, 2, 3, 5, 9, 13, 23, 28 and 35 (Appendix B). In addition, the three next-generation sequence datasets were very similar in their distributions (i.e., $r > 0.99$) for the three libraries and all had representation among GH families commonly associated with known cellulases (GH1, 3, 5, 6, 7, 8, 9, 12, 45, 48, 61; Wang *et al*., 2011; Lombard et al. 2013), hemicellulases (GH 8, 10, 11, 12, 26, 28, 53, 74; Wang *et al*.,2011; Lombard et al. 2013) and debranching enzymes (GH51, 54, 62, 67, 78, 74; Lombard et al. 2013). From the GH families mentioned above, the GH families involved in the metabolism of cellulose that were most abundant in our data were GH families 3, 5 and 9 (Fig. 15; Appendix B). Approximately 48% of the total GH sequences found in MG-RAST were annotated by Swiss-Prot and listed by CAZy database (Appendix B). Given that most GH family annotations were not represented by known CAZy identifiers and that only ~20% of our paired-end reads annotated in Swiss-Prot, the abundance and distribution of functional GH families in our pooled DNA is almost certainly underrepresented. As a result, we used functional screens of large-insert metagenomic libraries for the recovery of glycosyl hydrolases to circumvent these limitations of sequence-based analysis.

Table 3. Number of sequences with identified function categories by MG-RAST pipeline compared with the number of sequences annotated by Swiss-Prot and the number of glycosyl hydrolases annotated by MG-RAST pipeline.

Library	Sequences with identified functional categories	Sequences annotated by Swiss-Prot database	Sequences annotated as GH
Low pH Forward	17,127,682	3,337,989	3,133
Low pH Reverse	14,729,220	2,876,782	2,796
Cellulose Forward	9,002,909	1,768,661	2,849
Cellulose Reverse	7,999,214	1,560,613	2,533
Agricultural Forward	12,969,405	2,978,794	2,956
Agricultural Reverse	4,435,701	1,059,949	1,960

Functional categories were assigned by MG-RAST pipeline using annotations given by one or more protein databases. Sequences annotated as GH were assigned using custom Perl scripts annotated by Swiss-Prot and TrEmbl databases and paired with matching GH family from CAZy identifiers.

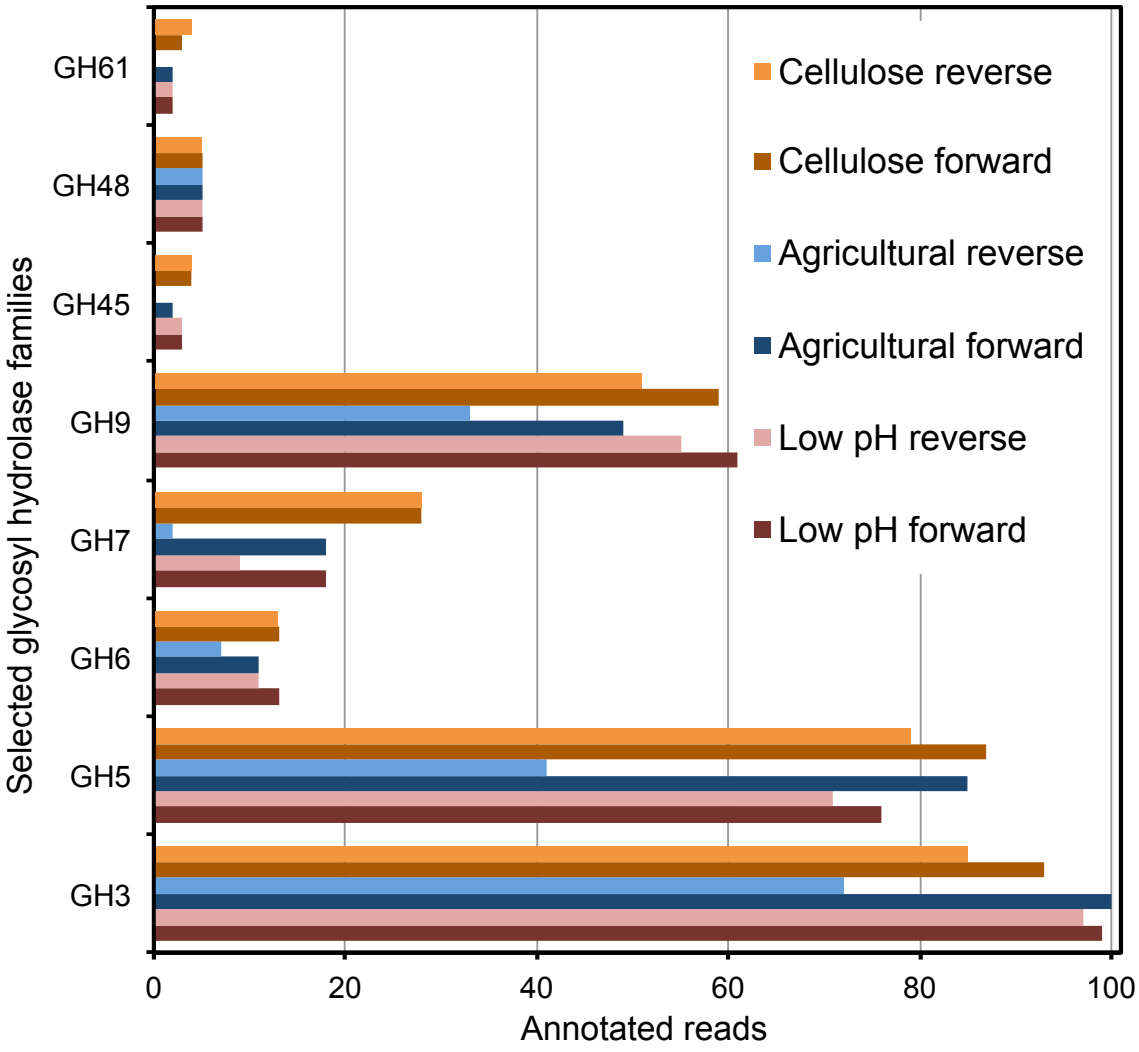

Fig. 15 Glycosyl hydrolase families related with cellulases present in the pooled heavy DNA. Functional annotation of the metagenomic data reveals diverse glycosyl hydrolase (GH) gene representation within the pooled heavy DNA. Reads were annotated by comparison to sequences in the UniprotKB/Swiss-Prot database and using custom Perl scripts annotated by Swiss-Prot and TrEmbl databases hits were paired with matching GH family CAZy identifiers.

3.6. Enriched metagenomic library

Pooled high molecular weight DNA from the ^{13}C-cellulose enriched SIP incubations for the three soils were captured in cosmid libraries and screened for GHs involved in the degradation of cellulose and other plant-derived polymers based on activity. Multiple displacement amplification (MDA) increased the amount of nucleic acids derived from pooled cellulose DNA-SIP incubations prior to the isolation of 20-40 kb DNA fragments via pulsed field gel electrophoresis (PFGE). Cosmid pJC8 (Neufeld *et al.*, 2011) was derived from the low copy and broad-host-range cosmid pRK7813 (Jones & Gutterson, 1987) and accommodates inserts of ~33 kb. To construct pJC8, recombination sequences (*attL1* and *attL2*) flanking a gentamicin resistance marker were cloned into *Hin*dIII/*Eco*RI sites of pRK7813, increasing transfer efficiency of DNA from *E. coli* to many different bacterial hosts via conjugation (Neufeld *et al.*, 2011). The cellulose-SIP metagenomic library generated ~83,000 clones with an average insert size of ~31 kb based on restriction digestion of a subset of clones (Fig. 16). These results were similar to results from a library of ~10,500 clones generated from MDA-amplified SIP-enriched seawater DNA, which had an average insert size of 27 kb, ranging from 17 to 40 kb (Neufeld *et al.*, 2008).

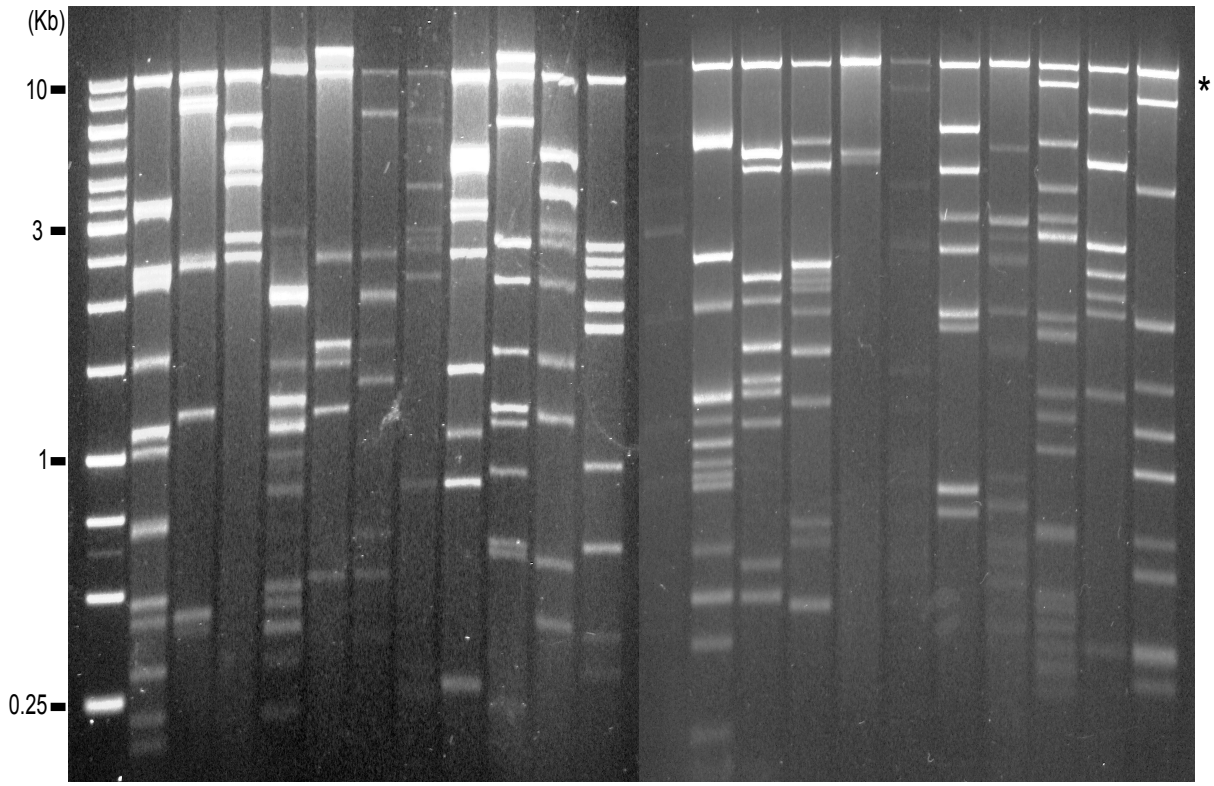

Figure 16. Size determination of inserts within the ^{13}C-cellulose enriched metagenomic library in kilobases (Kb). The ^{13}C-DNA from three soil samples were combined, MDA amplified and cloned into a cosmid to construct a metagenomic library. Randomly selected recombinant cosmids were analyzed by *Eco*RI/*Hin*dIII digestion. The upper shared band (denoted by "*") represents the pJC8 cosmid backbone common to all clones.

3.7. Functional Screening

We used a combined functional screening approach for screening 2,876 randomly selected clones (i.e., ~120 bacterial genomes screened assuming 2 Mb average genome size) from the cellulose-enriched metagenomic library. Qualitative agar plate growth of clones using carboxymethylcellulose (CMC) as a substrate and Congo Red staining (Teather & Wood, 1982) helped identify clones expressing both endoglucanase and glucosidase activities (EC 3.2.1; Enebro *et al.*, 2009). The results of this experiment showed two positive clones (2380 and 2044) capable of clearing CMC (Appendix C). The same 2,876 clones were screened on five different methylumbelliferone-based substrates (Table 4) and six clones were identified with high activity on one (i.e., C122, C2194) or more than one (i.e, C424, C762, C1024, C1088) of these fluorogenic substrates. Substrate activity profiles of C424 and C1088 were very similar, as were those of C122 and C2194, suggesting that genomic DNA captured in these pairs of clones may have derived from the same active organism overlapping in part of the sequence and/or they were duplicate inserts. Another possibility is that the enzymes encoded in these two clones have similar activities, with a possible common conserved region. The restriction digestion patterns showed that these positive clones have an insert size between 8 -21 kb (Table 4, Fig. 17) and also that C424 and 1088 had restriction sites in common. Therefore, it is likely that these clones contained overlapping inserts, but were not identical; the C122 and C2194 restriction patterns were distinct (Fig.17). Complete sequencing of these sequences would help clarify the relation between these two inserts.

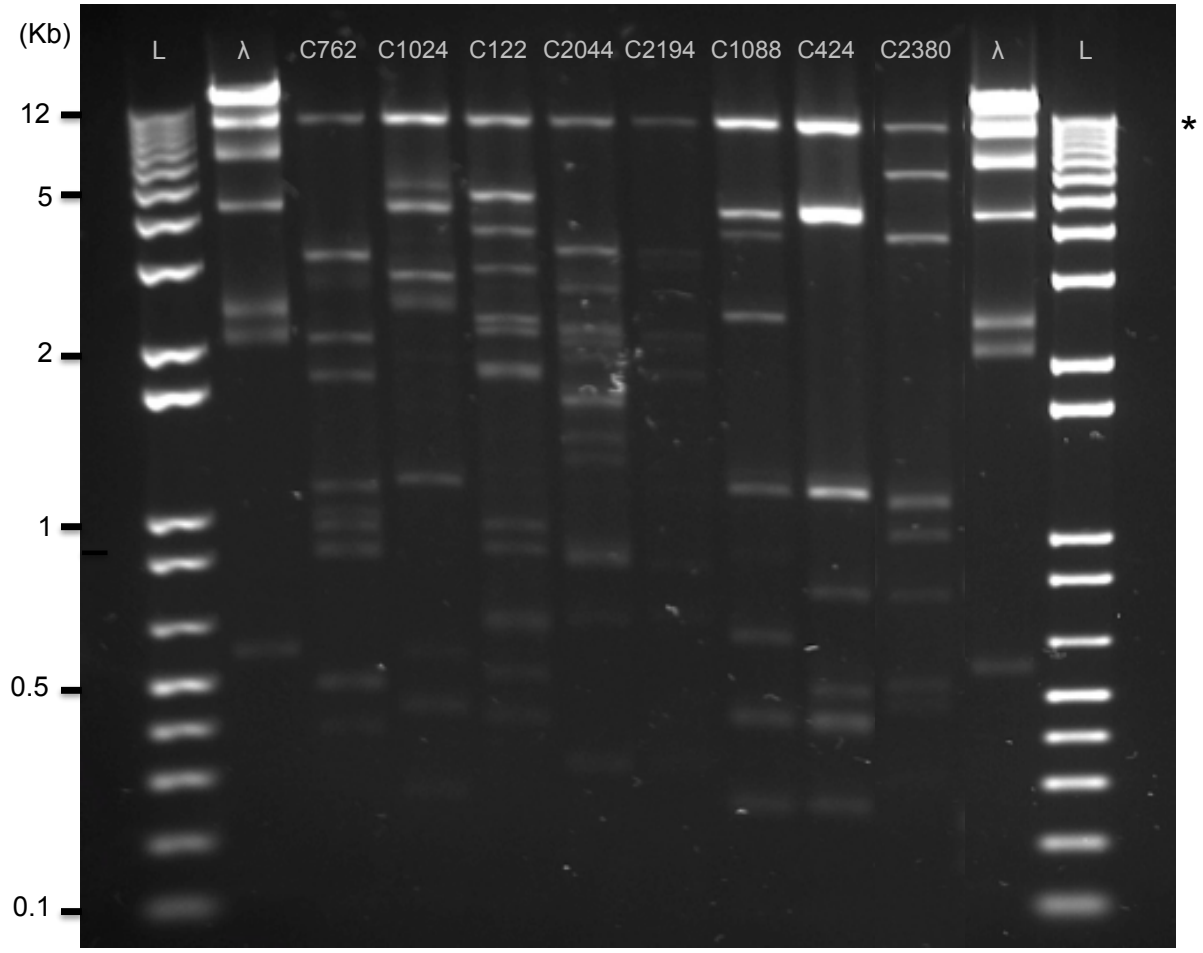

Figure 17. Size determination of inserts from eight positive clones within the ^{13}C-cellulose enriched metagenomic library. Recombinant cosmids were analyzed by *Eco*RI/*Hin*dIII/*BamHI* digestion. The upper shared band (denoted by "*") represents the pJC8 cosmid backbone common to all clones. Both the 1 kb plus (L) and λ DNA digested with *Hin*dIII (λ) ladders are shown.

Table 4. Substrate-specific activities of positive metagenomic clones from the ^{13}C-cellulose DNA-SIP library screened on methylumbelliferone-based substrates and carboxymethylcellulose.

		Activity (μM MU released)[1]					
Clone ID	Insert Size (Kb)[2]	α-L-arabino-furanoside pyranoside	β-D-cellobio-pyranoside	β-D-gluco-pyranoside	β-D-xylo-pyranoside	N-acetyl-β-D-galactos aminide	CMC[3]
C122	21.6	0.4	0.2	0.6	0.7	124.2	-
C424	8.2	0.9	57.6	109.4	1.6	0.7	-
C762	13.5	2.4	5.4	21.2	0.7	0.4	-
C1024	16.8	123.8	6.5	35.8	1.7	0.5	-
C1088	11.9	0.5	25.6	79.2	1.2	0.6	-
C2194	12.9	0.5	0.3	0.6	0.4	39.6	-
C2380	14.9	0.38	0.46	0.53	0.41	0.40	+++
C2044	14.7	0.40	0.40	0.52	0.39	0.36	++

[1]MU is methylumbelliferone units.

[2]Insert size was calculated based on the restriction digestion using *Eco*RI/*Hin*dIII/*BamHI*.

[3]CMC is carboxymethylcellulose; plate-based clearing (high "+++", medium "++" and negative "-") was detected by Congo Red stain and activity based on comparison to positive and negative controls.

The high frequency of positive clones after screening 2,876 DNA-SIP-derived clones screened compares favorably to previous soil functional metagenomic studies reporting the recovery of single positive cellulose hits from screening thousands of clones. For example, a single cellulase and two xylanases were recovered from functional screening of 13,800 clones from three fosmid metagenomic libraries derived from grassland in Germany, with an insert size range between 19-30 kb (Nacke *et al.*, 2012). Also, one cellulase was retrieved from the functional screening of 3,024 clones from a bacterial artificial chromosome metagenomic library derived from red soil in China, with insert sizes ranging from 25-165 kb (Liu *et al.*, 2011). In other case, one cellulase was recovered from functional screening of 14,000 clones with an average insert size of 5 kb from a metagenomic phage library from a forest soil in China (Wang *et al.*, 2009). Finally, a positive clone was retrieved from a metagenomic fosmid library derived from wetland soil in Korea, after screening 70,000 clones with an average insert of 40 kb (Kim *et al.*, 2008). Combining DNA-SIP and metagenomics helped recover soil glycosyl hydrolases in higher proportions than all of those reported previously via direct metagenomics, which demonstrates the power of using DNA-SIP as an activity-based pre-filter for targeted metagenomics approaches.

3.8. BLAST analysis of end-sequences

Cosmids were profiled by Sanger sequencing with forward and reverse primers flanking the site of the metagenomic DNA insertion. For each clone, two end-sequences were obtained, referred to as “reverse” and “forward” reads. The sequences were characterized by blastx and tblastx. Blastx uses all six reading frames translated nucleotide sequence and compares against the NCBI protein database. This approach is widely used to find proteins encoded in an

unknown nucleotide sequence and is more sensitive than blastn because the comparison is made at the protein level. The tblastx analysis uses all six reading frames of a translated nucleotide sequence and compares against all six reading frames of a translated nucleotide database. This is a useful tool in the identification of novel genes because it circumvents the limitations of potential frameshift errors that can prevent some open reading frames from being detected (Altschul *et al*., 1990). End sequencing of positive metagenomic library hits demonstrated a diversity of bacterial origins in the cloned metagenomic DNA that was actively expressed in *E. coli*. Most clones had at least one end sequences with a highest match to *Cellvibrio* (Table 5), the known cellulolytic *Gammaproteobacteria* (Mergaert *et al*., 2003). Also, two sequences matched with *Sorangium cellulosum*, a cellulose degrader within the *Deltaproteobacteria* (Lampky, 1971). Other important matches included *Saccharophagus degradans*, *Dyadobacter fermentans*, *Alicyclobacillus acidocaldarius* and *Chthoniobacter flavus* (Table 5). Although these bacteria were not well studied, it was reported that they use cellulose and other carbohydrates as a carbon source and/or they contained glycosyl hydrolases encoded in their genome (Taylor *et al*., 2006; Lang *et al*., 2009; Cheng *et al*., 2011, Kant *et al*, 2011). As predicted, the end sequence identities for C424 and C1088 were very similar taxonomically (i.e., *Cellvibrio* sp.). On the other hand, end-sequence data for C122 and C2194 did not suggest a similar genomic origin (Table 5). Posterior analysis of reverse and forward end-sequences of the positive clones was done by comparing end sequences to Illumina forward and reverse reads from whole genome sequencing of the three SIP libraries. The results showed that the majority of end-sequences were represented in the cellulose library, as expected, and only few sequences matches were found in other libraries using the selected threshold (Table 6).

Table 5. BLAST analysis of cosmid insert end-sequences. Cosmids were end-sequenced with M13 forward and reverse primers flanking the site of metagenomic DNA insertion. For each clone, two end-sequences were obtained referred to as "reverse" and "forward" reads. Top matches for blastx and tblastx analyses are shown

Clone	Forward tBlastx		Forward blastx		Reverse tBlastx		Reverse blastx	
	Description	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**
C 122	*Anabaena variabilis* (Glycosyl Transferase Fam 39)	$2e^{-04}$/40%/ /19 of 25	*Porphyromonas gingivalis* (4-amino-4-deoxy-L-arabinose transferase)	$4e^{-05}$/29% /40 of 139	*Cellvibrio japonicus* Ueda107 (β -xylosidase)	$5e^{-87}$/81%/ 146 of 162	*Cellvibrio japonicus* Ueda107 (β -xylosidase)	$8e^{-136}$/82% /131 of 162
C 424	*Saccharophagus degradans* (DNA-directed DNA polymerase)	$9e^{-16}$/49%/ 52 of 80	*Cellvibrio* sp. BR (DNA-directed DNA polymerase)	$1e^{-28}$/69%/ 66 of 80	*Cellvibrio japonicus* Ueda107 (Glucuronate isomerase)	$6e^{-96}$/88%/ 152 of 163	*Cellvibrio* sp. BR (Glucuronate isomerase)	$2e^{-103}$/91%/ 157 of 163
C 762	*Geobacter* sp. M18 (PAS/PAC sensor signal transduction histidine kinase)	$2e^{-71}$/68%/ 143 of 171	*Chthoniobacter flavus* (putative PAS/PAC sensor protein)	$1e^{-86}$/78%/ 151 of 171	*Sorangium cellulosum* (hypothetical protein)	$1e^{-29}$/54%/ 83 of 125	*Sorangium cellulosum* (hypothetical protein)	$2e^{-28}$/54%/ 83 of 125
C1024	*Cellvibrio japonicus* Ueda107 (Glucuronate isomerase)	$3e^{-28}$/85%/ 37 of 41	*Cellvibrio sp*. BR (Glucuronate isomerase)	$2e^{-17}$/95%/ 34 of 40	*Cellvibrio japonicus* Ueda107 (putative regucalcin)	$2e^{-39}$/69%/ 78 of 95	*Cellvibrio sp*. BR (Gluconolaconase)	$2e^{-46}$/80%/ 85 of 96
C1088	*Cellvibrio japonicus* Ueda107 (sodium/glucose cotransport protein)	$8e^{-62}$/69%/ 123 of 150	*Saccharophagus degradans* (SSS sodium solute transporter superfamily)	$6e^{-61}$/68%/ 123 of 150	*Cellvibrio japonicus* Ueda107 (putative membrane protein regucalcin)	$3e^{-27}$/60%/ 83 of 108	*Cellvibrio* sp. BR (auxin efflux carrier)	$5e^{-44}$/75%/ 101 of 114

Clone	Forward tBlastx		Forward blastx		Reverse tBlastx		Reverse blastx	
	Description	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**	**Description**	**E-value/% Identities/ Positives**
C2194	*Dyadobacter fermentans*	$6e^{-86}$/95%/ 140 of 142	*Dyadobacter fermentans* (ROK family protein)	$1e^{-91}$/95%/ 140 of 142	Failed sequencing reaction		Failed sequencing reaction	
C2380	*Alicyclobacillus acidocaldarius* (Glyoxalase /bleomycin resistance protein/dioxygenase)	$2e^{-12}$/57%/ 41 of 49	*Alicyclobacillus acidocaldarius* (Glyoxalase /bleomycin resistance protein/dioxygenase)	$2e^{-15}$/52%/ 51 of 69	*Cellvibrio japonicus* Ueda107 (glucosamine--fructose-6-phosphate aminotransferase, isomerizing)	$1e^{-88}$/87%/ 151 of 163	*Cellvibrio* sp. BR (glucosamine fructose-6-phosphate aminotransferase, isomerizing)	$3e^{-105}$/96%/ 162 of 163
C2044	*Cellvibrio japonicus* Ueda107 (DNA polymerase III subunit delta)	$1e^{-61}$/86%/ 109 of 117	*Cellvibrio sp*. BR (DNA polymerase III subunit delta)	$1e^{-71}$/96%/ 116 of 118	*Dyadobacter fermentans* (hypothetical protein)	$3e^{-124}$/97%/ 181 of 184	*Dyadobacter fermentans* (hypothetical protein)	$9e^{-129}$/97%/ 181 of 184

Positives are the number of amino acids from the query that are matching to the amino acids from the subject sequence. The total number of amino acids from the subject is shown.

Table 6. Sequence matches for forward and reverse cosmid end sequences against whole genome paired-end Illumina data for the three SIP libraries.

Clone	Low pH forward	Low pH reverse	Agricultural forward	Agricultural reverse	Cellulose forward	Cellulose reverse
C122 forward	0	0	1	4	4	1
C122 reverse	1	0	2	0	7	4
C424 forward	0	0	2	0	0	0
C424 reverse	0	0	0	0	2	1
C762 forward	0	0	0	0	3	3
C762 reverse	0	0	0	0	2	0
C1024 forward	0	0	0	0	0	1
C1024 reverse	0	0	0	0	1	0
C1088 forward	0	0	0	0	1	0
C1088 reverse	0	0	0	0	2	3
C2194 forward	1	0	1	2	0	1
C2194 reverse	Failed sequencing reaction	Failed sequencing reaction	Failed sequencing reaction	Failed sequencing reaction	Failed sequencing reaction	Failed sequencing reaction
C2380 forward	0	0	0	0	1	1
C2380 reverse	0	0	0	0	4	4
C2044 forward	0	0	0	1	7	3
C2044 reverse	0	0	0	0	0	0

4.0.Conclusions and future research

This research used DNA-SIP and metagenomics to recover novel glycosyl hydrolases from the microbial communities present in multiple Canadian soils. Using DNA-SIP, I was able to identify groups of microorganisms able to assimilate carbon from carbohydrates such as glucose, cellulose, arabinose and xylose. These organisms were studied using labeled substrates and confirmed the recovery of labeled fractions using DGGE and next-generation sequencing. Heavy DNA enrichment patterns were found in all samples incubated with ^{13}C-labelled substrates, revealing microorganisms active in plant-derived carbohydrate assimilation. The use of high throughput sequencing of the 16S rRNA gene marker confirmed that heavy DNA was distinct from light DNA in all cases. These heavy DNA samples were valuable data for proceeding with functional metagenomics for the recovery of novel glycosyl hydrolase genes for potential industrial applications.

The analysis of whole genome sequencing yielded multiple glycosyl hydrolase annotations, suggesting that cosmid library screens successfully recovered diverse GH genes from active microorganisms. Studies in metagenomics reveal that analysis of low complexity communities is ideal for acquiring knowledge about community structure, function and adaptation (Williamson & Yooseph, 2012). Soil microbial communities are considered the most complex due to the diversity of the microorganisms that inhabit enormous spatial heterogeneity (Torsvik & Øvreås, 2002; Mocali & Benedetti, 2010). Therefore, information related with soil metagenomic studies is poor compared with the information from other environments (Williamson & Yooseph, 2012). Although metagenomics has revolutionized the study of microbial ecology, there are still many challenges that need to be addressed in the future using this technology (Williamson & Yooseph, 2012).

My first hypothesis was to test whether combining metagenomics and stable-isotope probing (SIP) would lead to the identification of novel enzymes and microorganisms. The original hypothesis was confirmed by the results obtained in this research. The presence of novel genera of bacteria was shown in all soils. DNA sequence data suggested the presence of different GH families from cellulases, hemicellulases and debranching enzymes as well as a large group of uncharacterized GHs not present in CAZy database. Also, preliminary analysis in BLAST and CAZy from groups of contigs generated by ion torrent sequencing from the positive clones obtained, showed the presence not only of GHs, but glycosyl transferases (GT) and CBM enzymes involved in the metabolism of carbohydrates. Notably, many of them showed no similarities to GHs (data not shown).

My second hypothesis tested if the recovery of screened target enzymes from a combined DNA-SIP and metagenomics approach will be higher than by conventional metagenomics. This hypothesis was accepted because we recovered multiple enzymes from 2,876 clones screened, while all previous soil studies screened many more clones (10,000 to 70,000) than this study to recover single cellulases (Wang *et al*., 2009; kim *et al*., 2008; Liu *et al*., 2011; Nacke *et al*., 2012). My screening of 2,876 clones represents only an initial survey of the 83,000 clones generated in our cellulose-enriched library and I anticipate many more recovered enzymes if further screening is done on this one SIP-enriched metagenomic library.

My research has demonstrated the possibility of scaling DNA-SIP analysis for the interrogation of multiple environmental samples with multiple substrates, sampling at multiple time points. The utility of this pre-filter step prior to constructing metagenomic libraries was evident by the high proportion of positive clones with screening of a small proportion of the total clones available.

Combining SIP and metagenomics may also be useful in assigning metabolic function to both abundant microorganisms and low-relative abundance organisms that comprise the "rare biosphere" of soils. The rare biosphere comprises the microorganisms that are in low-relative abundance and can be detected by high-throughput sequencing (Lynch *et al.*, 2012). Based on the results obtained, many uncharacterized OTUs were represented in the indicator species OTUs associated with heavy DNA. Regardless of initial abundance, DNA-SIP can characterize microorganisms that are active in metabolizing specific substrates (Neufeld *et al.*, 2007b). Afterwards, metagenomics can aid with characterization or identification of the microorganisms involved in the metabolism of the substrate used. Importantly, indicator species analysis applied to heavy DNA can help circumvent the possibility that sequencing errors and chimeras may be common among detected rare biosphere members (Reeder & Knight, 2009; Lynch *et al.*, 2012).

Future research will involve sequencing and assembly of inserts, as well as transposon mutagenesis and sub-cloning experiments for identifying the specific genes encoding glycosyl hydrolases responsible for the activity detected in my screens, as well as assessing their phylogenetic placement. Also, studies of protein expression will be helpful because many of the novel sequences will not have any representation in Genbank. Therefore, protein sequence and structure characterization will complement the results obtained in this work using SIP and metagenomics.

Also, screening additional inserts from the ^{13}C-cellulose DNA-SIP library and libraries from the other pooled heavy DNA samples, will be important for comparison of these multiple heavy DNA pooled samples for maximizing novel glycosyl hydrolase gene recovery expanding the recovery of glycosyl hydrolases from these active and uncultivated soil bacteria and to

assess the bias introduced by the bacterial expression host, given that all of the positive clones recovered in this initial screen were not active in an alternative bacterial host (i.e., *Sinorhizobium meliloti*; data not shown).

This work demonstrated the power of combining functional metagenomics and DNA-SIP analysis for analyzing diverse environmental samples amended with multiple plant-derived carbon substrates, sampling at multiple time points. Using high-throughput sequencing of 16S rRNA genes of ^{13}C-enriched samples allowed the identification of multiple GHs. MDA of ^{13}C-labelled metagenomic DNA circumvented the problem of low DNA concentration. A high-quality cosmid library with an average insert 31 kb was constructed and screening of GHs from a small set of clones exhibited the value of these combined techniques for functional metagenomics research.

References

Agriculture and Agri-Food Canada. (1998). The Canadian system of soil classification. (3rd Ed). NRC-CNR. NRC Research Press. Ottawa, Canada.

Altieri, M. A. (1999). The ecological role of biodiversity in agroecosystems. Agric. Biol. Chem. 74: 19-31.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D. (1990). Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

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

Anderson, D. W. (1979). Processes of humus formation and transformation in soils of the Canadian great plains. J. Soil. Sci. 30: 77-84.

Angelov, A., Loderer, C., Pompei, S. & Liebl, W. (2011). Novel family of carbohydrate-binding modules revealed by the genome sequence of *Spirochaeta thermophila* DSM 6192. Appl. Environ. Microbiol. 77: 5483-5489.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., Yeh, L.-S. L. (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Res. 32: 115-119.

Astrow, J. D. (1996). Soil aggregate formation and the accrual of particulate and mineral associated organic matter. Soil Biol. Biochem. 28: 665-676.

Augusto, L., Ranger, J., Binkley, D. & Rothe, A. (2002). Impact of several common trees species of European temperate forest on soil fertility. Ann. For. Sci. 59: 233-253.

Bagnara, C., Toci, R., Gaudin, C. & Belaich, J. P. (1985). Isolation and characterization of a cellulolytic microorganism *Cellulomonas fermentans* sp. nov. Int. J. Syst. Bacteriol. 35: 502-507.

Baldrian, P., Kolařík, M., Štursová, M., Kopecký, J., Valášková, V., Větrovský, T., Žifčáková, L., Snajdr, J., Rídl, J., Vlček, C., Voříšková, J. (2012). Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J. 6: 248-258.

Balesdent, J., Chenu, C. & Balabane, M. (2000). Relationship of soil organic matter dynamics to physical protection and tillage. Soil Till. Res. 53: 215-230.

Bardgett, R. (2005). The Biology of Soil: A Community and Ecosystem Approach. Oxford University Press, New York. 232 p.

Bartram, A. K., Jiang, X., Lynch, M. D. J., Masella, A. P., Nicol, G. W., Dushoff, J. & Neufeld, J. D. (2014). Exploring links between pH and bacterial community composition in soils from the craibstone experimental farm. FEMS Microbiol. Ecol. 87: 403–415.

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. 77: 3846-3852.

Bauer, M. W., Driskill, L. E. & Kelly, R. M. (1998). Glycosyl hydrolases from hyperthermophilic microorganisms. Curr. Opin. Biotechnol. 9: 141-145.

Bayer, E. A., Belaich, J., Shoham ,Y. & Lamed, R. (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. Annu. Rev. Microbiol. 58: 521-554.

Bernard, L., Maron, P. A., Mougel, C., Nowak, V., Leveque, J., Marol, C., Balesdent, J., Gibiat, F. & Ranjard, L. (2009). Contamination of soil by copper affects the dynamics, diversity, and activity of soil bacterial communities involved in wheat decomposition and carbon storage. Appl. Environ. Microbiol. 75: 7565-7569.

Bernard, L., Mougel, C., Maron, P.-A., Nowak, V., Leveque, J., Henault, C., el Zahar Haichar, F., Berge, O., Marol, C., Balesdent, J., Gibiat, F., Lemanceau, P., Ranjard, L. (2007). Dynamics and identification of soil microbial populations actively assimilating carbon from ^{13}C-labelled wheat residue as estimated by DNA- and RNA-SIP techniques. Environ. Microbiol. 9: 752-764.

Billings, W. D. (1987). Carbon balance of Alaskan tundra and taiga ecosystems: past, present and future. Quat. Sci. Rev. 6: 165-177.

Boike, J., Langer, M., Lantuit, H., Muster, S., Roth, K., Sachs, T., Overduin, P., Westermann, S. & McGuire, A. D. (2012). Permafrost – physical aspects, carbon cycling, databases and uncertainties. In Recarbonization of the Biosphere: Ecosystems and Global Carbon Cycle. Lal R., Lorenz, K. & Hüttl, R. F. (Eds). Vol 8. pp. 159-180. Springer, London.

Borch ,T., Kretzschmar, R., Kappler, A., Van Capellen, P., Ginder-Vogel, M., Voegelin, A. & Campbell, K. (2010). Biogeochemical redox processes and their impact on contaminant dynamics. Environ. Sci. Technol. 44: 15-23.

Boyer, J. N. & Groffman, P. M. (1996). Bioavailability of water extractable organic carbon fractions in forest and agricultural soil profiles. Soil Biol. Biochem. 28: 783-790.

Bridgham, S. D., Updegraff, K. & Pastor, J. (1998). Carbon, nitrogen and phosphorous mineralization in northern wetlands. Ecology. 79:1545-1561.

Brussaard, L., Behan-Pelletier, V., Bignell, D. Brown, V., Didden, W., Folgarait, P., Fragoso, C., Freckman, D., Gupta, V., Hattori, T., Hawksworth, D., Kopatek, C., Lavelle, P., Malloch, D., Rusek, J., Soedestrooem, B., Tiedje, J., Virginia, R. (1997). Biodiversity and ecosystem functioning in soil. Ambio 26: 563-570.

Busk, P. K. & Lange, L. (2013). Cellulolytic potential of thermophilic species from fungal orders AMB Express 3: 1-10.

Butt, M. S., Tahir-Nadeem, M., Ahmad, Z. & Sultan, M. T. (2008). Xylanases and their applications in baking industry. Food Technol. Biotechnol. 46: 22-31.

Campanella, R. G. & Mitchell, J. K. (1968). Influence of temperature variations on soil behavior. J. soil Mech. Found. Div. 94: 709-734.

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: D233-238.

Caporaso JG, Kuczynski J, Stombaugh J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Gonzalez Peña, A., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, Jeremy, E., Ley, R. E., Lozupone, C. A. McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7: 335-336.

Chen, Y. & Murrell, J. C. (2010). When metagenomics meets stable-isotope probing: progress and perspectives. Trends Microbiol. 30:1-7.

Chen, Y., He, Y., Zhang, B., Yang, J. Li, W., Dong, Z., Hu, S. (2011). Complete genome sequence of *Alicyclobacillus acidocaldarius* strain Tc-4-1. J. Bacteriol. 193: 5602-5603.

Chen, Y., Dumont, M.G., Neufeld, J.D., Bodrossy, L., Stralis-Pavese, N., McNamara, N.P., Ostle, N., Briones, M. J. I. & Murrell, J. C. (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.

Chu, H., Fierer, N., Lauber, C. L., Caporaso, J. G., Knight, R. & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. Environ. Microbiol. 12: 2998-3006.

Cleveland, C. C., Neff, J. C., Towsend, A. R. & Hood, E. (2004). Composition, dynamics and fate of leached dissolved organic matter in terrestrial ecosystems: results from a decomposition experiment. Ecosystems. 7: 275-285.

Conen, F. & Smith, K. A. (2000). An explanation of linear increases in gas concentration under closed chambers used to measure gas exchange between soil and the atmosphere. Eur. J. Soil Sci. 51: 111-117.

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

Davidson, E. A., Savage, K. E., Trumbore, S. E. & Borken, W. (2006). Vertical partitioning of CO_2 production within a temperate forest soil. Glob. Change Biol. 12: 944-956.

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

Del Grosso, S. J., Parton, W. J., Mosier, A. R., Ojima, D. S., Kumala, A. E. & Phongpan, S. (2000). General model for NO_2 and N_2 gas emissions from soils due to denitrification. Global Biogeochem. Cy. 14: 1045-1060.

Dexter, A. R. (1988). Advances in characterization of soil structure. Soil Till. Res. 11: 199-238.

Dufrene, M. & Legendre, P. (1997). Species assemblages and indicator species: the need for a flexible assymetric aproach Ecol. Monogr. 67: 345-366.

Dumont, M. G. & Murrell, J. C. (2005). Stable isotope probing-linking microbial identity to function. Nat. Rev. Microbiol. 3: 499-504.

Dumont, M. G., Radajewski, S. M., Miguez, C. B., McDonald, I. R. & Murrell, J. C. (2006). Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. Environ. Microbiol. 8: 1240-1250.

Dunford, E., Pinnell, L. J., Verastegui, Y., Cheng, J., Charles, T. C. & Neufeld, J. D. (2010). Metagenomic characterization of active cellulolytic consortia from artic tundra. In Canadian Society of Microbiologists 60th annual conference. McMaster University, Hamilton, Ontario p.69

Dunford, E. A. & Neufeld, J. D. (2010). DNA stable-isotope probing (DNA-SIP). J. Vis. Exp. (42). e2027. doi:10.3791/2027.

Eichorst, S. A. & Kuske, C. R. (2012). Identification of cellulose-responsive bacterial and fungal communities in geographically and edaphically different soils by using stable isotope probing. Appl. Environ. Microbiol. 78: 2316-2327.

el Zahar Haichar, F., Achouak, W., Christen, R., Heulin, T., Marol, C., Marais, M. F., Mougel, C., Ranjard, L., Balesdent, J. & Berge, O. (2007). Identification of cellulolytic bacteria in soil by stable isotope probing. Environ. Microbiol. 9: 625-634.

Enebro, J., Momcilovic, D., Siika-aho, M. & Karlsson, S. (2009). Investigation of endoglucanase selectivity on carboxymethyl cellulose by mass spectrometric techniques. Cellulose. 16: 271-280.

Ferrer, M., Beloqui, A., Timmis, K. N. & Golyshin, P. N. (2009). Metagenomics for mining new genetic resources of microbial communities. J. Mol. Microbiol. Biotechnol. 16: 109-123.

Fierer, N., Bradford, M. A. & Jackson, R. B. (2007). Toward an ecological classification of soil bacteria. Ecology. 88: 1354-1364.

Fierer, N. & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U. S. A. 103: 626-631.

Finlay, B. J., Maberly, S. C. & Cooper, J. I. (1997). Microbial diversity and ecosystem function. Oikos. 80: 209-213.

Franzluebbers, A. (1999). Microbial activity in response to water-filled pore space of variably eroded southern Piedmont soils. Appl. Soil Ecol. 11: 91-101.

Gardner, T., Acosta-Martinez, V., Senwo, Z. & Dowd, S. E. (2011). Soil rhizosphere microbial communities and enzymes activities under organic farming in Alabama. Diversity. 3: 308-328.

Gaudinski, J. B., Trumbore, S. E., Davidson, E. A. & Zheng, S. (2000). Soil carbon cycling in a temperate forest: radiocarbon-based estimates of residence times, sequestration rates and partitioning fluxes. Biogeochemistry. 51: 33-69.

Gelsomino, A. & Cacco, G. (2006). Compositional shifts of bacterial groups in a solarized and amended soil as determined by denaturing gradient gel electrophoresis. Soil Biol. Biochem. 91-102.

Gil-Sotres, F., Trasar-Cepeda, C., Leiros, M. C. & Seoane, S. (2005). Different approaches to evaluating soil quality using biochemical properties. Soil Biol. Biochem. 37: 877-887.

Girvan, M. S., Bullimore, J., Pretty, J. N., Osborn, A. M. & Ball, A. S. (2003). Soil type is the primary determinant of the composition of the total and active bacterial communities in arable soils. Appl. Environ. Microbiol. 69: 1800-1809.

Gobat, J., Aragno, M. & Matthey, W. (2004). The living soil: Fundamentals of soil science and soil biology. Ed. Science Publishers, INC. USA. 585 pp.

Gowen, C. M. & Fong, S. S. (2010). Exploring biodiversity for cellulosic biofuel production. Chem. Biodivers. 7: 1086-1097.

Green, S. J., Leigh, M. B. & Neufeld, J. D. (2010). Denaturing gradient gel electrophoresis (DGGE) for microbial community analysis. In Microbiology of Hydrocarbon and Lipid Microbiology. Timmis, K. N. (Ed). pp. 4137-4158. Springer-Verlag.

Gregorich, E. G., Carter, M. R., Angers, D. A., Monreal, C. M. & Ellert, B. H. (1994). Towards a minimum data set to assess soil organic matter quality in agricultural soils. Can. J. Soil Sci. 74: 367-385.

Guazzaroni, M., Beloqui, A., Golyshin, P. N. & Ferrer, M. (2009). Metagenomics as a new technological tool to gain scientific knowledge. World J. Microbiol. Biotechnol. 25: 945-954.

Gupta, P., Samant, K. & Sahu, A. (2012). Isolation of cellulose-degrading bacteria and determination of their cellulolytic potential. Int. J. Microb. doi:10.1155/2012/578925.

Hall, M., Bansal, P., Lee, J. H., Realff, M. J. & Bommarius, A. S. (2010). Cellulose crystallinity a key predictor of the enzymatic hydrolysis rate. FEBS J. 277: 1571-1582.

Han, S. K., Nedashkovskaya, O. I., Mikhailov, V. V., Kim, S. B. & Bae, K. S. (2003). *Salinibacterium amurskyense* gen. nov., sp. nov., a novel genus of the family *Microbacteriaceae* from the marine environment. Int. J. Syst. Evol. Microbiol. 53: 2061-2066.

Hancock, S. M. & Withers, S. G. (2007). Glycosidases: functions, families and folds. In: eLS. John Wiley & Sons Ltd, Chichester. doi: 10.1002/9780470015902.a0020548

Handelsman, J. (2007). Metagenomics and microbial communities. Encyclopedia of Life Sciences. John Willey & Sons Ltd. pp.1-8.

Handelsman, J. (2005). Metagenomics or megagenomics. Nat. Rev. Microbiol. 3: 457-458.

Hatami, S., Alikhani, H. A., Besharati, H., Salehrastin, N., Afrousheh, M. & Jahromi, Z. Y. (2008). Invetigation on aerobic cellulolytic bacteria in some of north forest and farming. American-Eurasian J. Agric. Environ. Sci. 3: 713-716.

Henrissat, B. & Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. 7: 637-644.

Henrissat, B. & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. Biochem. J. 316: 695-696.

Hill, P. W., Farrar, J. F. & Jones, D. L. (2008). Decoupling of microbial glucose uptake and mineralization in soil. Soil Biol. Biochem. 40: 616-624.

Hoover, R. B., Pikuta, E. V., Bej, A. K., Marsic, D., Whitman, W. B., Tang, J. & Krader, P. (2003). *Spirochaeta americana* sp. nov., a new haloalkaliphilic, obligately anaerobic

Spirochaete isolated from soda Mono Lake in California. Int. J. Syst. Evol. Microbiol. 53: 815-821.

Horn, R. & Smucker, A. (2005). Structure formation and its consequences for gas and water transport in unsaturated arable and forest soils. Soil Till. Res. 82: 5-14.

Horn, S. J., Vaaje-Kolstad, G., Westereng, B. & Eijsink, V. G. (2012). Novel enzymes for the degradation of cellulose. Biotechnol. Biofuels 5: 1-12.

Huan, S., Shen, P. & Zhang, H. (2012). Isolation and identification of cellulolytic bacteria from the gut of *Holotrichia parallela* larvae (*Coleoptera*:*Scarabidae*). Int. J. Mol. Sci. 13: 2563-2577.

Islam, M. S., Kawasaki, H., Nakagawa, Y., Hattori, T. & Seki, T. (2007). *Labrys okinawensis* sp. nov. and *Labrys miyagiensis* sp. nov., budding bacteria isolated from rhizosphere habitats in Japan and emended descriptions of the genus Labrys and *Labrys monachus*. Int. J. Syst. Evol. Microbiol. 57: 552-557.

Johnson, M. J., Lee, K. Y. & Scow, K. M. (2003). DNA fingerprinting reveals links among agricultural crops, soil properties and the composition of soil microbial communities. Geoderma. 114: 279-303.

Jones, J. D. & Gutterson, N. (1987). An efficient mobilizable cosmid vector, pRK7813 and its use in a rapid method for marker exchange in *Pseudomonas fluorescens* strain HV37a. Gene. 61: 299-306.

Kaisermann, A., Roguet, A., Nunan, N., Maron, P., Ostle, N. & Lata, J. (2013). Agricultural management affects the response of soil bacterial community structure and respiration to water stress. Soil Biol. Biochem. 66: 69-77.

Kalyuzhnaya, M. G., Lapidus, A., Ivanova, N., Copeland, A.C., McHardy, A.C., Szeto, E., Salamov, A., Grigoriev, I. V., Suciu, D., Levine, S.R., Markowitz, V. M., Rigoutsos, I., Tringe, S.G., Bruce, D.C., Richardson, P.M., Lidstrom, M. E., Chistoserdova, L. (2008). High-resolution metagenomics targets specific functional types in complex microbial communities. Nat. Biotechnol. 26: 1029-1034.

Kant, R., van Passel, M. W. J., Palva, A., Lucas, S., Lapidus, A., del Rio, T. G., Dalin, E., Tice, H., Bruce, D., Goodwin, L., Pitluck, S. Larimer, F. W., Land, M. L., Hauser, L., Sangwan, P., de Vos, W. M., Janssen, P. H., Smidt, H. (2011). Genome sequence of *Chthoniobacter flavus* Ellin 428, an aerobic heterotrophic soil bacterium. J. Bacteriol. 193: 2902-2903.

Kennedy, J., O'Leary, N. D., Kiran, G. S., Morrissey, J. P., O'Gara, F., Selvin, J. & Dobson, A. D. W. (2011). Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. J. Appl. Microbiol. 111: 787-796.

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

Koch, O., Tscherko, D. & Kandeler, E. (2007). Temperature sensitivity of microbial respiration, nitrogen mineralization and potential soil enzyme activities in organic alpine soils. Global Biogeochem. Cy. 21: 1-11.

Koizumi, S., Yue, Z., Tomita, Y., Kondo, T., Iwase, H., Yamaguchi, D. & Hashimoto, T. (2008). Bacterium organizes hierarchical amorphous structure in microbial cellulose. Eur. Phys. J. E. 26: 137-142.

Kovatcheva-Datchary, P., Egert, M., Maathuis, A., Stojanovic, M. R., Graaf, A. A., Smidt, H., Vos, W. M. & Venema, K. (2009). Linking phylogenetic identities of bacteria to starch fermentation in an in vitro model of the large intestine by RNA-based stable isotope probing. Environ. Microbiol. 11: 914-926.

Krueger, J. P., Butz, R. G. & Cork, D. J. (1991). Aerobic and anaerobic soil metabolism of dicamba. J. Agric. Food Chem. 39: 995-999.

Ladd, J. N., Foster, R. C. & Skjemstad, J. O. (1993). Soil structure: carbon and nitrogen metabolism. Geoderma. 56: 401-434.

Lal, R. (2005). Forest soils and carbon sequestration. Forest. Ecol. Manag. 220: 242-258.

Lal, R. & Kimble, J. M. (1977). Conservation tillage for carbon sequestration. Nutr. Cycl. Agroecosys. 49: 243-253.

Lal, R. & Lorenz, K. (2012). Carbon sequentration in temperate forests. In Recarbonization of the Biosphere:Ecosystems and the Global Carbon Cycle. Lal, R., Lorenz, K. & Hüttl, R. F., (Eds). Vol. 9. pp. 187-198. Springer, London.

Lampky, J.R. (1971). Distribution of *Sorangium cellulosum*. Appl. Microbiol. 22: 937-938.

Lang, E., Lapidus, A., Chertkov, O., Brettin, T., Detter. J. C., Han, C., Copeland. A., del Rio, T. G., Nolan, M., Chen, F., Lucas, S., Tice, T., Cheng, J.-F., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C. D., Kopitz, M., Bruce, D., Goodwin, L., Pitluck, S., Ovchinnikova, G., Pati, A., Ivanova, N., Mavrommatis, K., Chen, A., Palaniappan, K., Chain, P., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Göker, M., Rohde, M., Kyrpides, N. C., Klenk, H.-P. (2009). Complete genome sequence of *Dyadobacter fermentans* type strain ($NS114^T$). Stand. Genomic Sci. 1: 133-140.

Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. 75: 5111-5120.

Lee, S. D. (2007). *Devosia subaequoris* sp. nov., isolated from beach sediment. Int. J. Syst. Evol. Microbiol. 57: 2212-2215.

Lee, T. K., Lee, J., Sul, W. J., Iwai, S., Chai, B., Tiedje, J. M. & Park, J. (2011). Novel biphenyl-oxidizing bacteria and dioxygenase genes from a Korean tidal mudflat. Appl. Environ. Microbiol. 77: 3888-3891.

Leschine, S. B. (1995). Cellulose degradation in anaerobic environments. Annu. Rev. Microbiol. 49: 399-426.

Li, X., Li, F., Zed, R., Zhan, Z. & Sing, B. (2007). Soil physical properties and their relations to organic carbon pools as affected by land use in an alpine pastureland. Geoderma. 139: 98-105.

Li, L., McCorkle, S. R., Monchy, S., Taghavi, S. & Van der Lelie, D. (2009). Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnol. Biofuels 2:1-11

Liebner, S., Harder, J. & Wagner, D. (2008). Bacterial diversity and community structure in polygonal tundra soils from Samoylov island, Lena Delta, Siberia. Int. Microbiol. 11: 195-202.

Lin, X., Smerdon, J. E., England, A. W. & Pollack, H. N. (2003). A model study of the effects of climatic precipitation changes on ground temperatures. J. Geophys. Res. 108: 1-11.

Lindroos, A., Derome, J., Derome, K. & Smolander, A. (2011). The effect of Scots pine, Norway spruce and silver birch on the chemical composition of stand throughfall and upper soil percolation water in norther Finland. Boreal Environ. Res. 16: 240-250.

Linn, D. & Doran, J. (1984). Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. Soil Sci. Soc. Am. J. 48: 1267-1272.

Liu, J., Liu, W., Zhao, X., Shen, W., Cao, H. & Cui, Z. (2011). Cloning and functional characterization of a novel endo-ß-1,4-glucanase gene from a soil-derived metagenomic library. Appl. Microbiol. Biotechnol. 89: 1083-1092.

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. (2013). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42:D490-495.

Lopuzone, C. & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71: 8228-8235.

Lorenz, K. & Lal, R. (2012). Cropland soil carbon dynamics. In Recarbonization of the Biosphere:Ecosystems and Global Carbon Cycle. Lal, R., Lorenz, K. & Hüttl, R. F. (Eds). Vol. 14. pp. 303-335. Springer, London.

Lynch, M. D., Masella, A. P., Hall, M. H., Bartram, A. K. & Neufeld, J. D. (2013). AXIOME: automated exploration of microbial diversity. GigaScience 2: 1-5.

Lynch, M. D. J., Bartram, A. K. & Neufeld, J. D. (2012). Targeted recovery of novel phylogenetic diversity from next-generation sequence data. ISME J. 6: 2067-2077.

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: 506-577.

Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., Rubin, E. M. & Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature. 480: 368-371.

Maki, M., Leung, K. T. & Qin, W. (2009). The prospects of cellulase-producing bacteria for the bioconversion of lignocellulosic biomas. Int. J. Biol. Sci. 5: 500-516.

Marschner, P., Crowley, D. & Yang, C. H. (2004). Development of specific rhizosphere bacterial communities in relation to plant species, nutrition and soil type. Plant Soil 261: 199-208.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics. 13: 31.

McCarter, J. D. & Withers, S. G. (1994). Mechanisms of enzymatic glycoside hydrolysis. Curr. Opin. Struct. Biol. 4: 885-892.

Medeiros, P. M., Fernandes, M. F., Dick, R. P. & Simoneit, B. R. T. (2006). Seasonal variations in sugar contents and microbial community in a ryegrass soil. Chemosphere. 65: 832-839.

Mergaert, J., Lednicka, D., Goris, J., Cnockaert, M. C., De Vos, P. & Swings, J. (2003). Taxonomic study of *Cellvibrio* strains and description of *Cellvibrio ostraviensis* sp. nov., *Cellvibrio fibrivorans* sp. nov. and *Cellvibrio gandavensis* sp. nov. Int. J. Syst. Evol. Mic. 53: 465-471.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R. & Wilke, A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9: 386.

Mocali, S. & Benedetti, A. (2010). Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res. Micr. 161: 497-505.

Monard, C., Vandenkoornhuyse, P., Le Bot, B. & Binet, F. (2011). Relationship between bacterial diversity and function under biotic control: the soil pesticide degraders as a case study. ISME J. 5: 1048-1056.

Muyzer, G., de Waal, E. C. & Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. 59: 695-700.

Nacke, H., Engelhaupt, M., Brady, S., Fischer, C., Tautzt, J. & Daniel, R. (2012). Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. Biotechnol. Lett. 34: 663-675.

Naik, V. N., Sharma, D. D., Kumar, P. M.P. & Yadav, R. D. (2012). Efficacy of ligno-cellulolytic fungi on recycling sericultural wastes. Acta Biol. Indica 1: 47-50.

Nannipieri, P., Ascher, J., Ceccherini, M. T., Landi, L., Pietramellara, G. & Renella, G. (2003). Microbial diversity and soil functions. Eur. J. Soil Sci. 54: 655-670.

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. Genomic Sci. 5: 203-210.

Neufeld, J. & Mohn, W. (2006). Assessment of microbial phylogenetic diversity based on environmental nucleic acids. In Molecular Identification, Systematics and Population Structure of Prokaryotes. Stackebrandt, E. (Ed.), pp. 219-260. Eidelberg: Springer-Verlag.

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., Dumont, M. G., Vohra, J. & Murrell, J. C. (2007 a). Methodological considerations for the use of stable isotope probing in microbial ecology. Microb. Ecol. 53: 435-442.

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

Nilsson, I. S., Miller, H. G. & Miller, J. D. (1982). Forest growth as a possible cause of soil and water acidification: an examination of the concepts. Oikos. 39: 40-49.

Oades, J. M. (1984). Soil organic matter and structural stability: mechanisms and implications for management. Plant Soil 76: 319-337.

Osman, K. T. (2013). Soils: Principles, Properties and Management. Springer Verlag. Dordrecht. 271 p.

Pan, Y., Birdsey, R. A, Fang, J., Houghton, R., Kauppi, P.E., Kurz, W. A., Phillips, O.L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., Hayes, D. (2011). A large and persistent carbon sink in the world's forest. Science. 333: 988-993.

Paustian, K., Andren, O., Janzen, H. H., Lal, R., Smith, P., Tian, G., Tiessen, H., Van Noordijk, M. & Woomer, P. L. (1997). Agricultural soils as a sink to mitigate CO_2 emissions. Soil Use Manage. 13: 230-244.

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

Pinnell, L. J., Dunford, E., Ronan, P., Hausner, M. & Neufeld, J. D. (2013). Metagenomic analysis of active cellulolytic microorganisms from an Arctic tundra soil. Can. J. Microbiol. Submitted.

Post, W. M., Izaurralde, R. C., West, T. O., Liebig, M. A. & King, A. W. (2012). Management opportunities for enhancing terrestrial carbon dioxide sinks. Front. Ecol. Environ. 10: 554-561.

Priha, O. & Smolander, A. (1999). Nitrogen transformations in soil under *Pinus sylvestris*, *Picea abies* and *Betula pendula* at two forest sites. Soil. Biol. Biochem. 31: 965-977.

Prusov, E. V. (2013). Total synthesis of antibiotics: recent achievements, limitations and perpectives. Appl. Microbiol. Biotechnol. 97: 2773-2795.

Quastel, J. H. (1965). Soil Metabolism. Ann. Rev. Plant Physiol. 16: 217-240.

Raaijmakers, J. M., Vlami, M. & de Souza, J. T. (2002). Antibiotic production by bacterial biocontrol agent. A. Van Leeuw. 81: 537-547.

Radajewski, S., Ineson, P., Parekh, N. R. & Murrell, J. C. (2000). Stable-isotope probing as a tool in microbial ecology. Nature. 403: 646-649.

Radajewski, S., Webster, G., Reay, D. S., Morris, S. A., Ineson, P., Nedwell, D. B., Prosser, J. I. & Murrell, J. C. (2002). Identification of active methylotroph populatiosn in an acidic forest soil by stable-isotope probing. Microbiology. 148: 2331-2342.

Reeder, J. & Knight, R. (2009). The "rare biosphere": a reallity check. Nat. Methods. 6: 636-637.

Ronan, P., Yeung, C.W., Schellenberg, J., Sparling, R., Wolfaardt, G. M. & Hausner, M. (2013). A versatile and robust aerotolerant microbial community capable of cellulosic ethanol production. Bioresour. Technol. 129: 156-163.

Sazci, A., Radford, A. & Erengler, K. (1986). Detection of cellulolytic fungi by using Congo Red as an indicator: a comparative study with the dinitrosalicyclic acid reagent method. J. Appl. Bacteriol. 61: 559-562.

Schäfer, T., Borchert, T. W., Nielsen, V. S., Skagerlind, P., Gibson, K., Wenger, K., Hatzack, F., Nilsson, L. D., Salmon, S., Pedersen, S., Heldt-Hansen, H. P., Poulsen, P. B., Lund, H., Oxenboll, K. M., Wu, G. F., Pedersen, H. H., Xu, H. (2007). Industrial Enzymes. Adv. Biochem. Eng. Biotechnol. 105: 59-131.

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.

Schimel, J. P. & Schaeffer, S. M. (2012). Microbial control over carbon cycling in soil. Front. Microb. 3: 1-11.

Schjonning, P., Munkholm, L. J., Moldrup, P. & Jacobsen, O. H. (2002). Modelling soil pore characteristics from measurements of air exchange: the long-term effects of fertilization and crop rotation. Eur. J. Soil Sci. 53: 331-339.

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.

Sessitsch, A., Weilharter, A., Gerzabek, M. H., Kirchmann, H. & Kandeler, E. (2001). Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Appl. Environ. Microbiol. 67: 4215-4224.

Seyfried, M. S. & Grant, L. E. (2007). Temperature effects on soil dielectric properties measured at 50 MHz. Vadose Zone J. 6: 759-765.

Sharma, P., Kumari, H., Kumar, M., Verma, M., Kumari, K., Malhotra, S., Khurana, J. & Lal, R. (2008). From bacterial genomics to metagenomics: concept, tools and recent advances. Indian J. Microbiol. 48: 173-194.

Sharma, S. K., Ramesh, A., Sharma, M. P., Joshi, O. P., Govaerts, B., Steenwerth, K. L. & Karlen, D. L. (2010). Microbial community structure and diversity as indicators for evaluating soil quality. In Biodiversity, Biofuels, Agroforestry and Conservation Agriculture, Sustainable Agriculture Reviews. Lichtfouse, E. (Ed.). Vol. 5. pp. 317-345. Springer, London.

Shin, S. C., Kim, S. J., Ahn, D. H., Lee, J. K., Lee, H., Lee, J., Hong, S. G., Lee, Y. M. & Parka, H. (2012). Genome Sequence of a *Salinibacterium sp*. isolated from antarctic soil. J. Bacteriol. 194: 2404. doi: 10.1128/JB.00235-12.

Simon, C. & Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. Appl. Microbiol. Biotechnol. 85: 265-276.

Sleator, R. D., Shortall, C. & Hill, C. (2008). Metagenomics. Lett. Appl. Microbiol. 47: 361-366.

Smith, K. A., Ball, T., Conen, F., Dobbie, K. E., Massheder, J. & Rey, A. (2003). Exchange of greehouse gases between soil and atmosphere:interactions of soil physical factors and biological processes. Eur. J. Soil Sci. 54: 779-791.

Stotzky, G. (1997). Soil as an environment for microbial life. In Modern Soil Microbiology. Van, J. D., Trevors, J. T. & Wellington, E., (Eds.). pp. 1-19. Marcel Dekker Inc, New York.

Subramaniyan, S. & Prema, P. (2002). Biotechnology of microbial xylanases:enzymology, molecular biology and application. Crit. Rev. Biotechnol. 22: 33-46.

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

Taylor, L. E., Henrissat, B., Coutinho, P. M., Ekborg, N. A., Hutcheson, S. W., Weiner, R. M. (2006). Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40^{T}. J. Bacteriol. 188: 3849-3861.

Teather, R. M. & Wood, P. J. (1982). Use of Congo Red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. Appl. Environ. Microbiol. 43: 777-780.

Thomsen, I. K., Schjønning, P., Jensen, B., Kristensen, K. & Christensen, B. T. (1999). Turnover of organic matter in differently textured soils: II. Microbial activity as influenced by soil water regimes. Geoderma 89: 199-218.

Tisdall, J. M. & Oades, J. M. (1982). Organic matter and water-stable aggregates in soils. J. Soil Sci. 33: 141-163.

Tkacz, J. S. & Lange, L. (2004). Advances in Fungal Biotechnology for Industry, Agriculture and Medicine. Kluwer Academic Plenum Publishers, New York. 447 p.

Torsvik, V. & Øvreås, L. (2002). Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. 5: 240-245.

Twarakavi, N. K. C., Simunek, J. & Schaap, M. G. (2010). Can texture-based classification optimally classify soils with respect to soil hydraulics. Water Resour. Res. 46: 1-11.

Urich, T., Lanzen, A., Qi, J., Huson, D. H., Schleper, C. & Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PloS ONE. 3: 1-13.

Van Veen, J. A. & Kuikman, P. J. (1990). Soil structural aspects of decomposition of organic matter by micro-organisms. Biogeochemistry. 11: 213-233.

Vasilyeva, L. V., Omelchenko, M. V., Berestovskaya, Y. Y., Lysenko, A. M., Abraham, W., Dedysh, S. N. & Zavarzin, G. A. (2006). *Asticcacaulis benevestitus* sp. nov., a psychrotolerant, dimorphic, prosthecate bacterium from tundra wetland soil. Int. J. Syst. Evol. Microbiol. 56: 2083–2088**.**

Viaud, M., Pasquier, A. & Brygoo, Y. (2000). Diversity of soil fungi studied by PCR-RFLP of ITS. Mycol. Res. 104: 1027-1032.

Wackett, L. P. (2004). Stable isotope probing in biodegradation reearch. Trends Biotechnol. 22: 153-154.

Walker, E. & Warren, F. L. (1937). Decomposition of cellulose by *Cytophaga I.* Biochem. J. 32:31-43.

Wang, F., Li, F., Chen, G. & Liu, W. (2009). Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches. Microbiol. Res. 164: 650-657.

Wang, T.-Y., Chen, H.-L., Lu, J. M.-Y., Chen, Y.-C., Sung, H.-M., Mao, C.-T., Cho, H.-Y., Ke, H.-M., Hwa, T.-Y., Ruan, S-K, Hung, K-Y., Chen, C.-K., Li, J.-Y., Wu, Y.-C., Chen, Y.-H., Chou, S.-P., Tsai, Y.-W., Chu, T.-C., Shih, A. C.-C., Li, W.-H., Shih, M.-C. (2011). Functional characterization of cellulases identified from the cow rumen fungus *Neocallimastix patriciarum* W5 by transcriptomic and secretomic analyses. Biotechnol. Biofuels 4:1-16.

Warren, R. A. J. (1996). Microbial hydrolysis of polysaccharides. Annu. Rev. Microbiol 50: 183-212.

Weber, C. F., Zak, D. R., Hungate, B. A., Jackson, R. B., Vilgalys, R., Evans, R. D., Schadt, C. W., Megonigal, J. P. & Kuske, C. R. (2011). Responses of soil cellulolytic fungal communities to elevated atmospheric CO_2 are complex and variable across five ecosystems. Environ. Microbiol. 13: 2778-2793.

Williamson, S. J. & Yooseph, S. (2012). From bacterial to microbial ecosystems (metagenomics). In Bacterial Molecular Networks: Methods and Protocols, Methods in

Molecular Biology. van Helden, J., Toussaint, A. & Thieffry, D. (Eds.). Vol. 804. pp. 35-51. Springer Science, London.

Wilson, D. B. (2011). Microbial diversity of cellulose hydrolysis. Curr. Opin. Microbiol. 14: 259-263.

Wilson, D. B. (2010). Demonstration of the importance for cellulose hydrolysis of CelS, the most abundant cellulosomal cellulase in *Clostridium thermocellum*. Proc. Natl. Acad. Sci. U.S.A. 107: 17855-17856.

Withers, S. G. (1995). Enzymatic cleavage of glycosides:how does it happen. Pure App. Chem. 67: 1673-1682.

Xu, J. (2006). Microbial ecology in the age of genomics and metagenomics:concepts, tools and recent advances. Mol. Ecol. 15: 1713-1731.

Yoo, S., Weon, H., Kim, B., Hong, S., Kwon, S., Cho, Y., Go, S. & Stackebrandt, E. (2006). *Devosia soli* sp. nov., isolated from greenhouse soil in Korea. Int. J. Syst. Evol. Microbiol. 56: 2689-2692.

Zhang, D., Liu, H., Xin, Y., Yu, Y., Zhou, P. & Zhou, Y. (2008). *Salinibacterium xinjiangense* sp. nov., a psychrophilic bacterium isolated from the China No. 1 glacier. Int. J. Syst. Evol. Microbiol. 58: 2739–2742.

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

Zhou, J., Davey, M. E., Figueras, J. B., Rivkina, E., Gilichinsky, D. & Tiedje, J. M. (1997). Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. Microbiology**.** 143: 3913-3919.

Appendix

Appendix A: Indicator species results

Tables A1 to A4 can be obtained upon request:

A1. Indicator species results comparing heavy DNA samples versus light DNA samples from all three soils (1AT, 7TR, 11AW) using an indicator value threshold of 0.7 and a 250 sequence minimum sum. The number of OTUs is represented for each treatment as well as the sum of all columns. A representative OTU sequence is shown.

A2. Indicator species results comparing heavy DNA samples versus light DNA samples from Tundra soil (1AT), Temperate rainforest soil (7TR) and Agricultural soils (7TR), using an indicator value threshold of 0.7 and a 250 sequence minimum sum. The number of OTUs is represented for each treatment as well as the sum of all columns. A representative OTU sequence is shown.

A3. Indicator species results when comparing individual substrates versus the other substrates from heavy DNA samples from all three soils (1AT, 7TR, 11AW) using an indicator value threshold of 0.7 and a 250 sequence minimum sum. The number of OTUs is represented for each treatment as well as the sum of all columns. A representative OTU sequence is shown.

A4. Indicator species results when comparing individual substrates versus the other substrates from heavy DNA samples within each soil (1AT, 7TR,11AW) using an indicator value threshold of 0.7 and a 250 sequence minimum sum. The number of OTUs is represented for each treatment as well as the sum of all columns. A representative OTU sequence is shown

Appendix B

Number of glycosyl hydrolase families present in the pooled heavy DNA of three soils.

GH Families	Low pH forward	Low pH reverse	Agricultural forward	Agricultural reverse	Cellulose forward	Cellulose reverse
GH1	116	119	106	68	107	105
GH2	68	63	63	52	63	56
GH3	99	97	100	72	93	85
GH4	11	9	14	13	8	10
GH5	76	71	85	41	87	79
GH6	13	11	11	7	13	13
GH7	18	9	18	2	28	28
GH8	13	13	12	9	12	11
GH9	61	55	49	33	59	51
GH10	40	39	38	24	39	38
GH11	28	24	25	18	30	31
GH13	136	121	132	100	127	126
GH14	8	5	9	2	8	6
GH15	12	10	8	5	11	10
GH16	20	16	17	8	21	19
GH17	26	17	15	8	14	14
GH18	10	7	7	5	9	9
GH20	26	21	21	9	16	18
GH22	10	4	1	1	8	1
GH23	197	146	191	140	149	141
GH24	9	8	14	7	7	6
GH25	4	3	4	1	4	4
GH26	8	8	7	5	8	7
GH27	30	29	24	16	23	21
GH28	61	44	35	11	48	37
GH29	14	14	13	10	14	11
GH30	6	4	4	1	4	6
GH31	41	35	39	21	39	29
GH32	32	24	41	25	26	24
GH33	11	10	13	4	18	10
GH34	2	4	1	0	2	2
GH35	60	54	40	13	47	39
GH36	10	10	8	8	7	8
GH37	49	46	45	34	46	32
GH38	18	13	10	6	17	12
GH39	3	3	2	0	2	3
GH42	6	6	6	6	6	6
GH43	1	1	1	1	1	1
GH44	1	1	1	1	1	1
GH45	3	3	2	0	4	4
GH46	3	2	3	1	4	4
GH47	23	16	19	6	26	21
GH48	5	5	5	5	5	5

GH Families	Low pH forward	Low pH reverse	Agricultural forward	Agricultural reverse	Cellulose forward	Cellulose reverse
GH1	116	119	106	68	107	105
GH2	68	63	63	52	63	56
GH3	99	97	100	72	93	85
GH4	11	9	14	13	8	10
GH5	76	71	85	41	87	79
GH6	13	11	11	7	13	13
GH7	18	9	18	2	28	28
GH8	13	13	12	9	12	11
GH49	3	3	3	2	4	3
GH50	2	2	2	1	2	1
GH51	16	15	12	9	14	14
GH52	0	0	0	0	0	0
GH53	8	7	4	3	7	6
GH54	8	7	7	3	8	8
GH55	2	1	2	0	2	1
GH56	1	2	4	2	7	3
GH57	7	6	7	6	7	6
GH58	2	2	2	1	1	1
GH61	2	2	2	0	3	4
GH62	6	6	7	4	7	7
GH63	6	3	2	1	3	3
GH64	1	1	1	1	2	1
GH65	7	7	4	2	7	5
GH66	5	1	4	4	4	2
GH68	4	4	10	10	2	3
GH71	1	1	0	0	2	0
GH73	4	5	3	1	5	5
GH74	10	10	10	10	10	10
GH76	4	4	3	0	4	4
GH77	11	11	10	7	9	8
GH79	5	6	2	4	2	3
GH81	2	2	4	0	4	4
GH82	2	1	0	0	1	1
GH83	2	2	3	1	2	3
GH84	1	0	1	0	1	1
GH85	0	1	1	0	1	1
GH88	1	0	1	1	1	1
GH89	1	1	1	0	1	1
GH95	3	3	3	2	3	3
GH96	2	1	2	2	2	2
GH99	6	4	4	1	5	3
GH102	0	2	2	0	0	2
GH103	2	2	2	2	2	2
GH104	2	2	2	2	2	0
GH105	1	1	1	1	1	0
GH109	4	4	4	3	3	4
GH110	7	7	5	3	3	3

GH Families	Low pH forward	Low pH reverse	Agricultural forward	Agricultural reverse	Cellulose forward	Cellulose reverse
GH1	116	119	106	68	107	105
GH2	68	63	63	52	63	56
GH3	99	97	100	72	93	85
GH4	11	9	14	13	8	10
GH5	76	71	85	41	87	79
GH6	13	11	11	7	13	13
GH7	18	9	18	2	28	28
GH8	13	13	12	9	12	11
GH116	4	3	3	1	3	3
Others*	1,590	1,454	1,557	1,071	1,441	1,257
Total	3,133	2,796	2,956	1,960	2,849	2,539

* Others represent the sequences that annotated for GHs but are not listed in CAZY.

Appendix C. Functional screening of positive clones with carboxymethylcellulose (CMC) as substrate. Clone C2380 (left) and C2044 (right) showed a clearing zone after staining with Congo Red.