Functional metagenomic analysis of carbohydrate degrading enzymes from the human gut microbiota

A thesis presented for the degree of Doctor of Philosophy
at the University of East Anglia
(Institute of Food Research and Rowett Institute of Nutrition and Health)

by
Anna Maria Szczepańska
MSc (University of Łódź, Poland)

October 2011

(73,761 words)

Abstract

The gut microbiota is a complex and diverse microbial community that is adapted to a carbohydrate-rich ecosystem. Plant cell wall components (cellulose, hemicelluloses and pectins), resistant starch and various oligosaccharides reach the colon by escaping digestion in the upper gastrointestinal tract. Fermentation of these dietary carbohydrates by the gut microbiota has well-recognised beneficial effects on host health. The microbial community in the human gut requires specific enzymes to efficiently degrade these carbohydrates. In this project, a culture-independent approach based on functional screening of genomic and metagenomic libraries using *Escherichia coli* and *Lactococcus lactis* as heterologous expression hosts, was used to isolate novel genes encoding glycoside hydrolase (GH) enzymes. The study identified several active GH enzymes involved in the breakdown of dietary polysaccharides such as starch, cellulose, xylan and β-glucan, recovered from the *E. coli* metagenomic library. The bioinformatic analysis of the insert from positive clones showed the presence of ORFs with the similarity to enzymes from GH families 13, 43 and 51 encoded by dominant bacterial genera from the human colon (*Bacteroides* sp., *Roseburia* sp., *Ruminococcus* sp.). A group of clones encoding potentially novel GH enzymes was also identified, emphasising the importance of functional-based study. One highly active clone was detected during screening of the *L. lactis* metagenomic library and showed fibrolytic activity on cellulose-, lichenan- and xylan-containing plates. The insert contained a partial gene with the GH9 catalytic domain and identity to the protein from *Coprococcus eutactus* ART55/1. Further functional analysis established the fibrolytic activity of selected *Coprococcus* species. Moreover, several active clones were isolated from the *Ruminococcus* sp. 80/3 genomic library which encoded protein with the similarity to enzymes from GH families 2, 3 and 5. In this work, the traditional approach of expression in *E. coli* was complemented by using an alternative host – *L. lactis.* While this did not improve the screening efficiency in terms of number of recovered clones, differences in gene expression and protein export between *E. coli* and *L. lactis* were noted during this study which highlights the benefits of using different heterologous hosts in functional metagenomic approaches.

List of contents

List of tables

List of figures

Verses on I Know Not What

My latest tribute here I send,
With this let your collection end.
Thus I consign you down to fame,
A character to praise and blame,
And, if the whole may pass for true,
Contented rest; you have your due –
Give future times the satisfaction
To leave one handle for detraction.

Jonathan Swift (1765)

Acknowledgements

This thesis would not have been accomplished without the help and support of people I have met and worked with during this project.

I would like to express my gratitude for this sponsorship from organizations where I had the privilege to work: the **Rowett Institute of Nutrition and Health** and the **Institute of Food Research**. I would like to thank **Dr. Petra Louis** who always inspired and challenged me, whose knowledge and expertise helped me to go through many difficulties. I would like to give special thanks to **Professor Harry Flint** who served as a mentor and advisor to this project. I appreciate his kindness and comprehensive guidance through the period of my study. To my fellow researchers from the Rowett Institute of Nutrition and Health: **Dr. Sylvia Duncan** and **Dr. Karen Scott**, for their advice and everyday smile. **Freda Farquharson** and **Jenny Martin** for their practical assistance in everyday lab work. I would like to acknowledge **Pauline Young** and **Gill Campbell** for their assistance with colony picking, library screening and sequencing. From the Institute of Food Research I would like to thank my former supervisors **Dr. Claire Shearman** and **Professor Mike Gasson** who initiated this project. I appreciate their help and support during first years of my PhD. I would like to thank **Professor Mike Peck** and **Dr. Bruce Pearson** for taking over the position of my supervisors and rescuing the situation. I would like to acknowledge and extend my heartfelt gratitude to **Dr. Udo Wegmann** for all the practical advice and unlimited help in the lab.

A special thanks to all **fellow students** who I have met during last four years. Thank you all the joyful moments we have sheared together.

Most especially, I would like to thank my **family, friends** in Poland and **Julien**. There are not words that can express what I own you, thank you for all the support and patience for all these years, thanks for always being with me and for me, no matter what happened and where we are.

Finally, I would like to thank my examiners: **Dr. Paul O'Toole** and **Dr. Arjan Narbad** for reading this thesis and any future readers who might find this work inspiring.

Declaration

I hereby declare that this thesis and the work described herein are original, except where indicated by reference or otherwise, and has not previously been submitted for any degree at this or any other university.

Anna Maria Szczepańska
October, 2011

Conference contribution

Szczepanska A., Louis P., Flint H.J. Identification of genes involved in degradation of dietary polysaccharides from the human gut metagenome. *9th Carbohydrate Engineering Meeting*, Lisbon, Portugal, May 2011

Szczepanska A., Wegmann U., Louis P., Flint H., Gasson M., Shearman C. Construction of a genomic library of *Ruminococcus*-related human gut isolate 80/3 and functional screening in *Escherichia coli* and *Lactococcus lactis*. *RINH – INRA* conference, Aberdeen, June 2010

Szczepanska A., Wegmann U., Louis P., Flint H., Gasson M., Shearman C. Functional screening of a genomic library from a human gut anaerobe in *Escherichia coli* and *Lactococcus lactis* using a novel shuttle vector. *SGM Meeting*, Edinburgh, March 2010

Abbreviations

µm – micrometre
µF – micro farad
µg – microgram
aa – amino acid
BLAST - Basic Local Alignment Search Tool
CBM – carbohydrate binding module
CFU – colony forming units
Cm - chloramphenicol
CMC - carboxymethyl cellulose
DNA - deoxyribonucleic acid
Ery – erythromycin
F – Forward
g – gram
g – gravity force
GH – glycoside hydrolase
h – hour
kb - kilobase
M – molar
mg – milligram
min – minute
ml – milliliter
mM – millimolar
MSC – multiple cloning site
NCBI – National Centre for Biotechnology Information
ng - nanogram
ORF - open reading frame
PCR – polymerase chain reaction
R – reverse
RBS – ribosomal binding site
RNA – ribonucleic acid
SCFA – short chain fatty acids
SDS-PAGE - sodium dodecyl sulfate polyacrylamide gel electrophoresis
sec – second
U – units
V – Volt
v/v – volume per volume
w/v – weight per volume
Ω - ohm

Chapter 1

Introduction

Background

The human gut microbiota is a complex and diverse community which plays an important role in maintaining human health. The main function of the gut bacterial community is the degradation of dietary carbohydrates that escaped digestion in the upper gastrointestinal tract (GI tract). The efficient carbohydrate utilisation relies on the production of a variety of enzymes known as glycoside hydrolases that catalyse the cleavage of glycosidic bonds between two carbohydrate moieties (Henrissat 1991). During bacterial fermentation a number of metabolites are formed which have a beneficial or detrimental effect on human health. Homeostasis depends on a number of factors, particularly diet, which was shown to influence microbial composition and their metabolic activities.

Modern microbiology has developed many novel approaches to study microorganisms from various, often highly inaccessible environments (e.g. Arctic soil, deep oceans, thermal springs). Nowadays, an increasing number of microbiology laboratory is equipped not only with Petri dishes in order to culture microorganisms but is also populated with high-technology machines which sequence millions of nucleotides of cultured and uncultured microbes. A significant proportion of microorganisms occupying various ecosystems remain uncultured and therefore the vast majority of information on these microbial communities is derived through metagenomic analysis. Metagenomics (environmental genomics) provides an insight into the genetic potential of various microbial communities and may identify novel biomolecules which can find application in industry, medicine and science (Zoetendal *et al.,* 2008). To date, metagenomics has been used to study microbial worlds of the aquatic environment (Venter *et al.,* 2004, Kerkhof and Goodman 2009), soil (Berlemont *et al.,* 2011, Liu *et al.,* 2011), ancient bones and skeletons (Noonan *et al.,* 2005, Noonan *et al.,* 2006), animal rumen (Hess *et al.,* 2011), human faeces (Qin *et al.,* 2010, Arumugam *et al.,* 2011) and many more.

The contribution of the human gut microbiota to fermentation of dietary polysaccharides can be examined by using functional metagenomic approaches, which can provide new information on the bacterial enzymatic activities involved. Therefore, this thesis is devoted to investigating the enzymes required for dietary carbohydrate breakdown in the human gut microbiota by using functional metagenomics.

1.1 Anatomy and physiology of the human gastrointestinal tract

The human gastrointestinal tract is a system of organs responsible for digestion of consumed food and liquids (Figure 1.1). It starts with the oral cavity where the consumed food is chewed and moistened by secreted saliva and where mechanical digestion occurs. The enzymatic breakdown of dietary components is initiated by enzymes (amylase and lipase) secreted in the saliva by salivary glands. The broken pieces of food are swallowed and pushed by the tongue to the oesophagus and down to the stomach through peristaltic contraction of muscles. The stomach is a muscular "bag" where the masticated food is retained and further digested before sending it to the small intestine. Protein-degrading enzymes and hydrochloric acid are produced in the stomach. Proteins are digested by pepsin which is activated at the low pH created by HCl. Moreover, HCl inhibits and kills microorganisms ingested with the food. The stomach is connected with the small intestine via the duodenum, which is linked to the pancreas and the liver via the biliary tract. The pancreas produces precursors of digestive enzymes such as trypsinogen, chymotrypsinogen, pancreatic lipase, and amylase, collectively known as pancreatic juice. Bile is produced by the liver which allows emulsification of lipids. The next part of the small intestine is the jejunum where the majority of digestion and absorption occurs. The last part, which is connected to the large intestine, is called the ileum. The small intestine is a long tube with an average length of five meters in adults and it is lined with villi which enhance absorption of digested food. The last segment of GI tract is the large intestine which is divided into the caecum, colon, rectum, and anal canal. The colon consists of four sections: the ascending colon, the transverse colon, the descending colon, and the sigmoid colon. The colon is the main site of microbial colonisation and microbial activity including digestion of dietary components. The colon is also the place where further absorption occurs. Undigested and unabsorbed food residue is removed from the body by defecation (DeBruyne *et al.*, 2008).

An important protective component of the GI tract is the mucus layer which covers the epithelial cells of the stomach, small intestine and colon. The mucus is a stratified barrier composed of a well-defined outer 'loose' layer and an inner layer which is firmly attached to the epithelium of the stomach and large intestine. In contrast, the mucus in the small intestine is rather irregular and the stratification is not well

defined (Johansson *et al.*, 2011). The outer layer contains a large number of bacteria, but the inner layer is resistant to bacterial penetration and protects epithelial cells from direct contact with bacteria (Johansson *et al.*, 2008). Mice lacking the main mucus component (glycoprotein - mucin (MUC2) secreted by specialised epithelial cells known as goblet cells) suffer spontaneous inflammation, emphasising the importance of the mucosal barrier in the host/ bacterial homeostasis (Johansson *et al.*, 2011, Hooper and Gordon 2001).

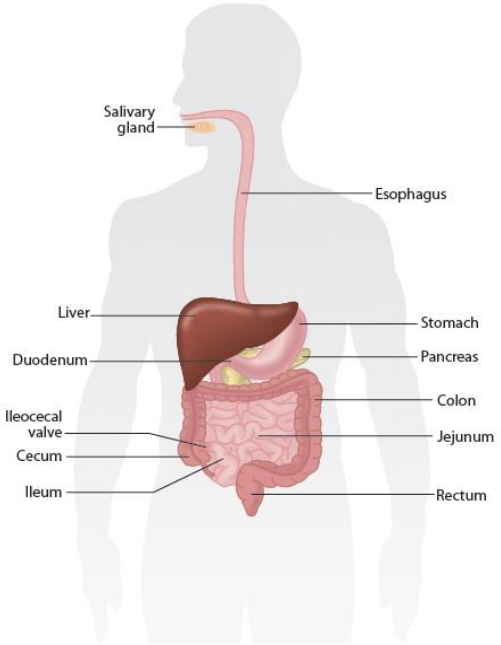


Figure 1.1 Overview of the human GI tract.
Adapted from Walter and Ley (2010)

1.2 Bacterial colonisation of the GI tract and its diversity

The human GI tract is sterile at birth, but microbial colonisation occurs immediately after birth. Initially, the microbiota is highly heterogeneous and influenced by various factors such as mode of delivery, infant feeding, and infant hospitalisation (Reid *et al.*, 2011). Babies delivered vaginally acquire a microbiota similar to their mother's vaginal microbial community. The microbiota of babies born by Caesarean section resembled the general skin microbial population of their mothers (Dominguez-Bello *et al.*, 2010). Microbial colonisation of infants' GI tract is also affected by the infant feeding regime (Reid *et al.*, 2011) and differs between breast-fed and formula-fed babies (Nakamura *et al.*, 2009). After weaning, a more diverse

and complex population becomes established, with the bacteria more characteristic of adult individuals (Kelly *et al.*, 2007, Spor *et al.*, 2011). The bacterial distribution along the human GI tract increases from the upper to the lower parts. The stomach is lightly populated by microbes due to its highly acidic environment. The jejunum is occupied by approximately 10^5 CFU.ml^{-1}. Bacterial overgrowth is restricted here by rapid luminal flow, the presence of bactericidal bile salts and a highly potent immune defence system. The bacterial population thrives in the large intestine and can reach up to 10^{11} CFU.ml^{-1}. This proliferation is facilitated by a higher pH, a larger volume and a longer retention time due to slow peristaltic movements (Walter and Ley 2010).

Studies on the gut microbiota reported substantial bacterial diversity with as many as a 1,000 different species (Hooper and Macpherson 2010). The enumeration and characterisation of cultured organisms is nowadays complemented with molecular profiling methods based on microbial 16S rRNA, including high-throughput sequencing, quantitative PCR, fluorescence *in situ* hybridization (FISH) and microarrays. These techniques have provided information on the composition and diversity of the predominant gut bacteria (Qin *et al.*, 2010, Tap *et al.*, 2009). Other molecular techniques such as deep-sequencing metagenomic analyses identified the phylogenetic and functional core of the human gut microbiota (Qin *et al.*, 2010, Arumugam *et al.*, 2011). Each individual carries at least 160 different bacterial species (Qin *et al.*, 2010). The predominant bacteria from the human gut belong to the phylum Bacteroidetes and to the low % G+C Firmicutes (Tap *et al.*, 2009, Duncan *et al.*, 2007, Walker *et al.*, 2011).

Previously, the gut microbiota was considered as difficult to culture, and it was reported that 93% of the human gut bacterial 16S rRNA sequences correspond to uncultured bacteria (Backhed *et al.*, 2005). The most recent study by Walker *et al.* (2011) however showed that the most abundant phylotypes (>2%) are cultured at nearly 100%, which clearly suggests that the majority of gut bacteria can be grown under laboratory condition. A similar conclusion was drawn by Goodman *et al.* (2011). The abundance of readily cultured bacterial phylotypes was estimated, followed by 16S rRNA analysis of complete faecal samples from healthy volunteers and was compared to the data derived from cultured samples. The results showed

that the culturability was correlated with the taxonomic level. At the family-level 89% phylotypes were readily cultured but at the species-level the proportion of cultured bacteria decreased to 56%.

1.2.1 Bacteroidetes

The Bacteroidetes phylum has been partitioned into three classes: Bacteroidia, Cytophagia and Flavobacteria, which together are called the CFB-group. The Bacteroidia class is associated with the human gut microbiota and comprises several families. The members of Bacteroidaceae are most frequently represented as part of the human gut microbiome. They are Gram-negative, pleomorphic, anaerobic bacteria that make up around 25% of the human colonic microbiota. They are well known for their metabolism of carbohydrate substrates (Xu *et al.,* 2003, Robert *et al.,* 2007, Chassard *et al.,* 2008, Mirande *et al.,* 2010) and formation of short chain fatty acids (SCFA) including succinate, acetate, lactate, formate and propionate as the end products of fermentation. Some *Bacteroides* species have also been reported to convert bile to metabolites, which have been considered as co-carcinogens or mutagens (Narushima *et al.,* 2006). Hence, some Bacteroidetes may be associated with a higher risk of colon cancer in particular *B. vulgatus* and *B. stercoris* (Narushima *et al.,* 2006, Moore and Moore 1995, Guarner and Malagelada 2003, Sobhani *et al.,* 2011). A recent study on the bacterial diversity of colorectal cancer patients reported significantly higher level of the *Bacteroides*/ *Prevotella* group than in controls (Sobhani *et al.,* 2011). The best studied member of this phylum is *B. thetaiotaomicron* whose genome was sequenced by Xu and co-workers (2003). *B. thetaiotaomicron* is a prominent human gut isolate which is able to degrade dietary glycan. The adaptation of this bacterium to an environment rich in carbohydrates is enabled by possessing multiple gene clusters that include a cell-associated multi-protein SUS system (starch utilisation system) (Xu *et al.,* 2003). This allows the bacterium to efficiently bind and degrade the substrate (Martens *et al.,* 2009). Genome sequencing and comparative analysis of other *Bacteroides* species has provided a better understanding of their metabolic potential (Xu *et al.,* 2007, Karlsson 2011). Other common species isolated from human faecal samples are *B. dorei, B. uniformis, B. distasonis* and *B. vulgatus* (Xu *et al.,* 2007). The other families of the Bacteroidia class, retrieved from human faecal samples, are

Porphyromonadaceae, Prevotellaceae and Rikenellaceae; however they are low prevalence bacterial taxa in the human gut microbiota (Peris-Bondia *et al.*, 2011). The Prevotellaceae and Porphyromonadaceae are generally associated with the oral cavity and the rumen (Karlsson 2011, Flint and Bayer 2008). A recent study showed that the genus *Prevotella* was exclusively present in African children consuming a fibre-rich diet compared to the counterparts on a Western-diet (De Filippo *et al.,* 2010). Sequence-based data analysis indicated the importance of *Prevotella* species in dietary fibre degradation by encoding a fundamentally novel xylose utilisation gene cluster (Dodd *et al.,* 2011). It has also been reported that the gut microbiota can be enriched in *Prevotella* species in a group of individuals classified to Qin's enterotype 2 (Qin *et al.,* 2010). Representatives of the family Rikenellaceae were identified in several studies; in particular *Alistipes* species are readily detectable in human faecal samples (Tap *et al.*, 2009).

1.2.2 Firmicutes

Firmicutes are the most abundant group of the human gut microbiota. They are Gram-positive, low %G+C bacteria and make up around 70% of the colonic microbiota (Tap *et al.,* 2009, Walker *et al.,* 2011, Peris-Bondia *et al.,* 2011). Based on 16S rRNA analysis, Clostridia, Bacilli, Erysipelotrichi and Negativicutes classes are present in human faecal samples (Peris-Bondia *et al.*, 2011). The Clostridia class is the most abundant and contains the order the Clostridiales with the families Ruminococcaceae, Clostridiaceae, Lachnospiraceae and Eubacteriaceae (Peris-Bondia *et al.*, 2011). The order Clostridiales has been divided into several clostridial clusters on the basis of 16S rRNA sequencing (Collins *et al.*, 1994). The members of clostridial clusters IV and XIVa are the dominant groups in the human GI tract.

Clostridium cluster IV is referred to as the *Clostridium leptum* group or Ruminococcaceae family with species such as: *Clostridium leptum, Cl. sporosphaeroides, Faecalibacterium prausnitzii, Ruminococcus bromii, R. champanellensis, R. flavefaciens,* and *R. albus*. *F. prausnitzii* is a predominant species in this group which is able to metabolize starch and inulin and form butyrate and D-lactate (Duncan *et al.,* 2007) and has been proposed to have an anti-inflammatory properties based on studies in a colitis mouse model (Sokol *et al.,*

2008). Ruminococci in cluster IV metabolize complex carbohydrates including starch (*R. bromii*) and cellulose (*R. champanellensis*) (Abell *et al.,* 2008, Chassard *et al.,* 2011). In the present study, a novel human gut isolate *Ruminococcus* sp. 80/3, (Walker *et al.,* 2008) was investigated which has an ability to hydrolyze sucrose, cellobiose, lactose, mannose, arabinose, rhamnose and trehalose (S. Duncan and M Pudenz, unpublished data). The strain possesses β-glucosidase activity which acts on glycosidic bonds present in breakdown products of plant cell wall compounds such as cellulose, β-glucans and xyloglucans (Dabek *et al.,* 2008). It also showed cellulase and weak xylanase activity. No activity was detected on starch or pectin (S. Duncan and M. Pudenz, unpublished data). Ruminococci are closely associated with particulate material from faecal samples (Walker *et al.,* 2008). Within cluster IV, highly-specialized cellulolytic ruminal species such as *R. flavefaciens* have been widely studied due to their production of a complex and sophisticated apparatus called the cellulosome (Bayer *et al.,* 2008, Flint *et al.,* 2008). Related bacteria have been detected in the large intestine of humans which are believed to be involved in the degradation of recalcitrant plant cell wall cellulose (Flint and Bayer 2008, Chassard *et al.,* 2011, Robert and Bernalier-Donadille 2003).

Clostridium cluster XIVa is referred to as the *Cl. coccoides* group or Lachnospiraceae family and consists of highly active short chain fatty acids (SCFA) producers (Louis and Flint 2009) and carbohydrate fermenters (Scott *et al.,* 2011). A recent study showed that this cluster is most abundant in the active fraction compared to the whole microbial community from human faecal samples. Among the Lachnospiraceae family the *Coprococcus* members were prevalent active bacteria in the faecal samples (Peris-Bondia *et al.,* 2011). The *Coprococcus* genus comprises Gram-positive anaerobic cocci that actively ferment carbohydrates and produce butyric and acetic acids together with formic or propionic acids (Holdeman and Moore 1974, Pryde *et al.,* 2002). The population of *Coprococcus* in healthy and IBS (irritable bowel syndrome) subjects was shown to be different (Kassinen *et al.,* 2007) and stress-induced changes were observed within its group in the mouse model (Bailey *et al.,* 2011). The other well-studied bacteria in clostridial cluster XIVa are members of *Roseburia/ Eubacterium rectale* group which have the ability to degrade starch by using cell-associated amylase enzymes (Ramsay *et al.,* 2006). The *Roseburia* species are motile, due to the presence of flagella which could possibly

enhance substrate acquisition and recognition by the host immune system (Scott *et al.,* 2011). These bacteria also contribute to the production of butyrate and lactate (Duncan *et al.,* 2007, Duncan *et al.,* 2004) and were reported as highly active linoleic acid metabolisers (McIntosh *et al.,* 2009). The bacterial members of other low %G+C clostridial clusters (clusters XI, XII, XV and XVI) are recovered from human faecal samples but are less abundant (Peris-Bondia *et al.,* 2011).

The Bacillus class of the phylum Firmicutes is mainly represented in human faecal samples by the order Lactobacillales. They are Gram-positive facultative anaerobic or microaerophilic bacteria that convert lactose and other sugars to lactic acid. Comparative genomics of the *Lactobacillus* species showed a wide range of adaptations to the human gut ecosystem (Ventura *et al.,* 2009, Frese *et al.,* 2011). Lactobacilli have been used widely as probiotics and a number of studies describe their beneficial effect against a range of GI tract conditions and infections (Ventura *et al.,* 2009). Antagonistic activity against common pathogens like *E. coli* or *Listeria monocytogenes* has also been demonstrated (Vaughan *et al.,* 2002, Servin 2004).

1.2.3 Actinobacteria

Actinobacteria are high %G+C Gram-positive bacteria that form around 2-5% of the colonic microbiota (Tap *et al.,* 2009, Peris-Bondia *et al.,* 2011), and are mainly represented by the order Bifidobacteriales and Coriobacteriales (Duncan *et al.,* 2007). Bifidobacteria ferment sugar to lactic acid and acetic acid and were shown to actively metabolise plant-derived dietary fibres or complex carbohydrate structures (Schell *et al.,* 2002, Van Den Broek and Voragen 2008). They dominate the early GI tract microbiota in breast-fed infants (Vaishampayan *et al.,* 2010). They are used as probiotics, and showed a beneficial effect towards restoration of microbial homeostasis in a number of studies (Reid *et al.,* 2011). The *Collinsella* genus from the order Coriobacteriales is found in more than 90% of human intestines (Kageyama and Benno 2000). They utilise a wide range of carbohydrates to produce acid. A reduction of a *Collinsella aerofaciens*-like phylotype was observed in the faecal samples of IBS patients compared to their healthy counterparts (Malinen *et al.,* 2010). The number of Actinobacteria in the human gut varies over time and was found to decline in elderly compared to younger subjects (Claesson *et al.,* 2011).

1.2.4 Proteobacteria

The Proteobacteria phylum is weakly represented in the normal human gut microbiome (less than 2%) (Tap *et al.,* 2009, Peris-Bondia *et al.,* 2011). It was shown that the abundance of γ-proteobacteria including *E. coli* and *Shigella* sp. was significantly higher in European children than in their counterparts from Burkina Faso and was influenced by a diet (De Filippo *et al.,* 2010). The latter bacteria are potentially pathogenic and have been associated with aetiology of IBS (Rajilic-Stojanovic *et al.,* 2011). Sulphate-reducing bacteria (SRB) belong to phylum Proteobacteria and include *Desulfovibrio* species which appear to be the dominant SRB in the human gut microbiota. SRB use H_2 as an electron donor to generate hydrogen sulphide (H_2S). The production of H_2S is potentially a major toxin to the gut epithelial cells and can be associated with chronic gastrointestinal disorders (Scanlan *et al.,* 2009).

1.2.5 Verrucomicrobia

This recently described phylum of bacteria makes up around 1-2% of the colonic microbiota (Duncan *et al.,* 2007). A member of this phylum is the novel gut isolate *Akkermansia muciniphila,* which specializes in mucin- degradation and converts it to acetate and propionate (Derrien *et al.,* 2004)

1.2.6 Archaea

The archaeal methanogens are represented in the human gut microbiota by *Methanobrevibacter smithii* and *Methanosphaera stadtmanae*. *M. smithii* uses H_2 and CO_2 or formate to produce methane whilst *M. stadtmanae* converts methanol to methane (Zoetendal *et al.,* 2008, Oxley *et al.,* 2010). The adult population is divided into methane excretors (30-40%, CH_4^+) and non-methane excretors (CH_4^-) based on detection of methane in the breath (Chassard, *et al.* 2008).

1.3 Factors affecting the GI tract bacterial population

The microbial population can be influenced by changes in diet and life events (stress, ageing, disease). The composition of the gut microbiota changes from infancy through the adulthood and in the elderly. The gradual changes in gut microbiota occur during early life with a decrease in number of aerobes and facultative anaerobes and increase of obligate anaerobic populations. After weaning, the gut microbiota starts resembling the adult's microbial community with the dominant bacteria from phyla Firmicutes and Bacteroidetes (Spor *et al.,* 2011). The higher abundance of *Bifidobacterium* and Clostridia species was reported for adolescent volunteers compare to adults (Agans *et al.,* 2011). The significant changes in gut microbiota of elderly people are affected by changing dietary habits and lifestyle; reduced intestinal motility, illness and medication treatment (Biagi *et al.,* 2010). The ratio of Firmicutes vs. Bacteroidetes was found to be atypical in elderly individuals, with a higher proportion of Bacteroidetes than in adults (Claesson *et al.,* 2011). This was in agreement with previous observations (Mariat *et al.,* 2009). A significantly higher proportion of enterobacteria was also found in elderly individuals compared to their younger counterparts. (Mueller *et al.,* 2006). The increase of intestinal permeability in elderly people is correlated with reduction of bacteria from clostridial cluster IV and XIVa and subsequently the decrease in SCFA production (Biagi *et al.,* 2010).

Diet is one of the most important environmental factors that has an impact on the microbial population in the GI tract. The dietary carbohydrates that escape digestion in the upper GI tract reach the colon and affect bacterial growth and their metabolic function. A study by Duncan *et al*. (2007) established that variations in carbohydrate intake influenced the microbial composition and SCFA production. A reduced carbohydrate intake led to a decrease in the number of butyrate-producing bacteria from the *Roseburia*/ *E. rectale* group. The same effect was observed by Walker *et al*. (2011). Reduced carbohydrate consumption may also influence bacterial homeostasis by increasing the pH of the colon. A higher colonic pH was reported to stimulate growth of *Bacteroides in vitro* and decrease the population of butyrate-producers (Walker *et al.,* 2005). Furthermore, a significant increase of cluster IV ruminococci was correlated with the proportion of resistant starch in the diet (Walker *et al.,* 2011).

The enrichment of *Bifidobacterium adolescentis*, *Eubacterium rectale* and *Ruminococcus bromii* was reported on resistant starch diet by Martinez *et al.* (2010). The substrate supplementation also changed the Firmicutes: Bacteroidetes ratio with an increase of the latter phylum (Martinez *et al.,* 2010). The possible effect of diet on the proportion of Bacteroidetes was shown for Burkina Faso and European children. A higher ratio of Bacteroidetes was observed in African children who consumed a fibre-rich diet (De Filippo *et al.,* 2010). Previously the proportion of Bacteroidetes in the faecal sample was shown to vary between lean and obese individuals, and was reported to be significantly lower in obese subjects. The Bacteroidetes fraction increased when the obese humans were on a weight loss diet (Turnbaugh *et al.,* 2009). It was proposed that the ratio of Bacteroidetes in the faecal sample might serve as an obesity biomarker in the future. However, other studies did not report a positive correlation between BMI and the occurrence of Bacteroidetes (Qin *et al.,* 2010, Duncan *et al.,* 2008).

Non-digestible food ingredients known as prebiotics such as inulin and fructo-oligosacchrides (FOS) are known to stimulate the growth of specific groups of gut bacteria. Diet supplementation with prebiotics was shown to have a bifidogenic effect on the microbiota of infants, adult participants and elderly individuals (Meyer and Stasse-Wolthuis 2009). Other study reported that an inulin supplementation also stimulated *F. prausnitzii* species which are among the main butyrate-producers in the GI tract (Ramirez-Farias *et al.,* 2009). The modulation of energy metabolism and satiety was correlated with prebiotic supplementation and provided evidence that food intake and glucose homeostasis could possibly be controlled (Cani *et al.,* 2009).

The gut microbiota varies between individuals, hence host genetics are believed to influence the microbial succession (Spor *et al.,* 2011). Similar bacterial community structures were observed between related individuals (twin siblings and their mother), but differences between monozygotic and dizygotic twins were not significant, therefore the heritability of the gut microbiota is still to be demonstrated (Turnbaugh *et al.,* 2009, Turnbaugh and Gordon 2009). Changes within the gut community were observed for each individual but they were only short-term and temporal, indicating that each person has a stable and well-defined microbial core (Turnbaugh *et al.,* 2009). Analysis of person-to-person variation of the gut

microbiota allowed clustering of individuals into three groups named enterotypes (Arumugam *et al.,* 2011). Sequencing data from human faecal samples showed variation in the level of predominant gut species between enterotypes. The presence of specific groups of bacteria was correlated to the functional differences that reflect various combinations of microbial activities between enterotypes.

1.4 Microbial activities in the GI tract

The functional potential of microbial activities in the human GI tract has been studied widely using various techniques including culture-based studies and culture-independent approaches such as metagenomics (Qin *et al.,* 2010, Arumugam *et al.,* 2011, Gill *et al.,* 2006, Kurokawa *et al.,* 2007) and metatranscriptomics (Booijink *et al.,* 2010, Gosalbes *et al.,* 2011). The outcome of these studies has revealed that the main metabolic role of the gut bacteria are carbohydrate metabolism and transport, amino acid and lipid metabolism, xenobiotic biodegradation and metabolism.

Non-digestible dietary carbohydrates (starches, cellulose, hemicelluloses, pectins and oligosaccharides) that escape digestion in the upper GI tract are efficiently degraded by gut bacteria. The ability to ferment these recalcitrant compounds is necessary for bacteria to thrive in the gut. An array of enzymes are produced which cleave the glycosidic bonds of complex carbohydrates (Flint *et al.,* 2008). The metagenomic study of the human gut microbiota showed a significant presence of genes involved in dietary carbohydrate utilisation (Qin *et al.,* 2010, Arumugam *et al.,* 2011, Kurokawa *et al.,* 2007). Host-derived glycoconjugates present in mucin are also substrates for microbes. In particularly, fucose was shown to serve as a source of energy for *Bacteroides* (Coyne *et al.,* 2005) and *Bifidobacterium* (Crociani *et al.,* 1994) species.

The end-products of carbohydrates fermentation are short-chain fatty acids (SCFAs), mainly acetate, propionate and butyrate, occurring roughly in a molar ratio of 3:1:1 (Wong and Jenkins 2007). These SCFAs supply energy to the host and play an important role in host health maintenance (Scheppach 1994). Acetate is used as a substrate for the synthesis of cholesterol and fatty acids; it increases colonic blood flow and oxygen uptake and affects ileal motility. Propionate is an anti-lipogenic agent with a cholesterol-lowering effect (Hosseini *et al.,* 2011). Propionate was

proposed as a satiety-inducing agent and its effect was studied using animal models and human studies. It affected production of satiety-stimulating hormones GLP-1 and PYY, and decreased the level of ghrelin (produced by the stomach), which stimulates appetite and therefore controls food intake (Cani *et al.*, 2009, Delzenne *et al.*, 2005). Butyrate is a major source of energy for gut epithelial growth and proliferation (Louis and Flint 2009). It plays a protective role against colitis and colorectal cancer by inhibiting the growth of colon cancer cells, inducing differentiation and apoptosis (Scharlau *et al.*, 2009). It has also been demonstrated that butyrate protects the gut by possessing anti-inflammatory properties. It inhibits the activation of the transcription factor NF-κB and reduces the formation of proinflammatory cytokines (Inan *et al.*, 2000, Segain *et al.*, 2000). A few studies have revealed pathways and specific enzymes for SCFA formation (Louis and Flint 2009, Miller and Wolin 1996, Charrier *et al.*, 2006).

Analysis of digestive material showed that in addition to carbohydrate, some soluble proteins and peptides reach the colon and are readily degraded by microbial enzymes (Macfarlane *et al.*, 1986). The degradation of proteins leads to the production of compounds such as ammonia, amines, indoles and phenols which can be potentially carcinogenic (Smith and Macfarlane 1997). The gut microbiota also contributes to the synthesis of essential amino acids and vitamins (Qin *et al.*, 2010). Enzymes that convert bile acids such as bile salt hydrolases were abundant and well conserved within the human gut microbiota. They play an important role in lipid metabolism which can influence the risk of metabolic disorders such as diabetes and obesity (Jones *et al.*, 2008). The conversion of diet-derived lipids to trimethylamine N-oxide (TMAO) by intestinal bacteria has been associated with atherosclerosis (Wang *et al.*, 2011). The metabolism of phytochemicals by human gut bacteria has been of interest in view of their potential role in health and disease prevention (Laparra and Sanz 2010). Gut bacteria are involved in hydrolysis of phenolic glycoside or ester linkages by the production of β-glucosidases, rhamnosidases and esterases (Gill *et al.*, 2006). It was reported that β-glucosidase activity is highly prevalent amongst gut microbiota (Dabek *et al.*, 2008). Microbial β-glucuronidase activity has a potentially beneficial effect on the bioavailability of dietary-derived compounds such as phytochemicals. However, production of β-glucuronidase was also connected with the generation of toxic and carcinogenic metabolites in the large intestine (McBain and Macfarlane 1998). A high frequency of β-glucuronidases in the human gut microbiota was

reported by Gloux *et al.* (2011) and is associated with the abundance of glucuronide compounds that reach the colon.

The gut microbiota plays an important protective role against pathogenesis by competing for dietary nutrients and colonization niches. They stimulate the development of gut-associated lymphoid tissue (GALT) which was shown to be under-developed in germ-free mice (raised under sterile condition) (Hooper 2004). A low density of lymphoid cells in the gut mucosa was observed in gnotobiotic mice. Specialised follicle structures were small and the immunoglobulin concentration was lower than in control animals (Kelly *et al.,* 2005). Restoration of the gut microbiota provided the signal to GALT development, including epithelial cell maturation, angiogenesis and lymphocyte development. The immunological tolerance between host and commensal bacteria is a very important aspect for the functional stability of the GI tract ecosystem. Intestinal homeostasis is maintained by several barriers, such as an adaptive IgA system, intestinal mucus, tight junctions and anti-microbial peptides. The disruption of these barriers can lead to aberrant immune responses which underlie the manifestations of inflammation, and can progress to chronic disease such as inflammatory bowel disease (IBD) or even colorectal cancer (Hooper and Macpherson 2010, O'Hara and Shanahan 2006).

1.5 Dietary polysaccharides

A healthy well-balanced human diet should provide all the macro- and micronutrients required to maintain the health of the person. Carbohydrates are key components in the diet, comprising sugars, oligosaccharides, starchy carbohydrates and non-starch polysaccharides. The carbohydrates provide an important source of energy, control blood glucose and insulin production, lipid metabolism and by reaching the colon they influence microbial fermentation and health of the GI tract (Cummings and Stephen 2007). According to the British Nutrition Foundation, the daily intake of carbohydrates for a man should be 300 g and a woman 230 g. Most of the consumed carbohydrates will be digested by enzymes produced in the upper GI tract. The remaining recalcitrant carbohydrates, known as dietary fibre, will escape digestion and reach the colon to serve as a substrate for microbial fermentation. The discussion on the term dietary fibre is still ongoing; however a recent definition was

proposed and refers to ‘intrinsic plant cell wall polysaccharides or non-starch polysaccharides’ (NSP). Other definition of dietary fibre refers to resistant starch (RS), oligosaccharides and non-starch polysaccharides (NSP). NSPs are cell wall components of plants - cellulose, hemicelluloses, pectins and lignin. Chemically related to cell wall NSP are plants gums and mucilages (Cummings and Stephen 2007). Dietary fibre can be found in foods such as cereals, wholegrain products, beans, lentils, fruits and vegetables.

1.5.1 Starch

Starch is a carbohydrate that consists of a mixture of amylose (non-branched chains of glucose residues linked by α-1,4-glycosidic bonds) and amylopectin (high-molecular weight, highly branched polymer, containing α-1,4 and α-1,6 linkages) (Figure 1.3). It is produced by green plants as an energy store. It is a major carbohydrate of the human diet, present in cereals (rice, wheat, and maize - type A), tuber, roots and high amylose starches (potatoes, banana -type B) and legumes (type C). The types of starch are defined based on the crystalline structure which confers on them distinct X-ray diffraction pattern (Hoover 2010). Starch is degraded by small intestinal amylases which act on α-1,4-glycoside linkages, but it can escape digestion in the upper GI tract. A starch that arrives in the colon as a fermentable carbohydrate source for intestinal bacteria is known as resistant starch. There are several types of resistant starch based on their chemical and physical properties. Starch which is protected by cell wall polymers and is therefore physically inaccessible for enzymatic degradation, is classified as type I (RS I). This starch is mostly present in whole grains, cereals, seeds and legumes. Food processing such as milling can reduce or eliminate resistance. Type II resistant starch is present in granules similar to type B starches observed in potatoes, whose crystalline structure protects them from enzymatic degradation. Retrograded starch (achieved after food processing e.g. cooking and cooling) is RS type III. Type IV resistant starch comes from a modification of the amylose: amylopectin ratio which can be accomplished during plant breeding. The other technique to achieve modified starch is chemical substitution such as esterification or etherification. The cross-linking introduces a limited number of linkages between chains of amylose and amylopectin and makes starch more resistant to digestion (Cummings and Stephen 2007).

1.5.2 Non-starch polysaccharides (NSP)

Non-starch polysaccharides are a group of large dietary carbohydrates consisting of monosaccharide (e.g. glucose, xylose) residues joined to each other by glycosidic linkages and are predominantly found in the plant cell wall (Cummings and Stephen 2007). It is a diverse and complex group of dietary carbohydrates which comprise cellulose, hemicelluloses and pectins.

Cellulose is a polysaccharide composed of units of glucose connected by β-1,4 glycosidic linkages (Figure 1.4). It is the most abundant organic compound in the world, building the cell wall of all plants. It is a recalcitrant, mainly crystalline carbohydrate that is not soluble in water (Cummings and Stephen 2007).

Hemicelluloses are homo- and hetero-polymers which include a mixture of hexose and pentose sugars, often in highly branched chains, comprising xylan, xyloglucan, β-glucans, arabinoxylans and glucomannans (Figure 1.5). Xylan is a cell wall hemicellulose present in cereals (millet grain, sorghum), legumes and vegetables. Xylans have a backbone of β-1,4-linked xylose residues with a varying degree of polymerisation by short side chains of residues such as α-D-glucosyluronic acid (GlcA), 4-O-methyl-α-D-glucosyluronic acid (MeGlcA), α-L-arabinose, acetic acid, ferulic acid, or coumaric acid (Dodd *et al.*, 2011) (Figure 1.5). β-glucans are present mainly in grains, fruits and vegetables and are composed of glucopyranoside residues linked by β-1,4-glycosidic bonds and β-1,3-glycosidic bonds. Xyloglucan is the hemicellulosic polysaccharide found mainly is fruits and vegetables. It is the main hemicellulosic polysaccharide in the primary walls of higher plants (up to 20% of the dry weight of the primary cell wall). The backbone of xyloglucan is composed of glucose residues with β-1,4 bonds, most of which are substituted with α-1,6 linked xylose side chains. Xyloglucans are classified as XXXG type and XXGG-type. The first type is composed of a typical glucopyranoside backbone with side chains of xylose connected to the three consecutive glucose residues of the main chain via α-1,6-branching. The xylose residues are often substituted with galactose and fucose residues. The second type of xyloglucan has two consecutive branched backbone residues and two unbranched backbone residues (Ariza *et al.*, 2011). Arabinoxylans are hemicelluloses typically found in cereal grains. The main backbone consists of β-

1,4-linked xylose residues substituted with arabinose residues by α-1,2- and/ or α-1,3 glycosidic bonds (Dodd *et al.,* 2011, Van den Abbeele *et al.,* 2011). Glucomannans are mainly linear polymers of glucose and mannose residues which can be substituted at the C-6 by α-D-galactopyranose units or acetyl groups (Várnai *et al.,* 2011).

Pectins are common to all plant cell walls and their major source in the human diet is a variety of fruits (apples, plums, oranges). They are highly heterogeneous and complex polymers composed of α-1,4-galacturonic acid residues in the backbone chain (Figure 1.5). Pectins are divided into different groups: homogalacturonan (HG), xylogalacturonan (XGA), rhamnogalacturonan I (RG-I) and rhamnogalacturonan II (RG-II). Homogalacturonan is a linear chain of α-1,4-linked galacturonic acid (GalA) residues in which some of the carboxyl groups (C6) are substituted with methyl groups. Acetyl esterification can occur at O2 or O3 of the galacturonic acid residue. Xylogalacturonan is a polymer of GalA units substituted with β-xylose at C3. XGA is found in pectins of cell wall from various sources such as apples, watermelon, soybeans and peas. Rhamnogalacturonan I (RG-I) comprises of backbone of galacturonic acid residues which are interrupted by α-1,2 linked rhamnose residues, to which long arabinan and galactan chains can be attached at the C4 position. Further substitution can occur to the side chains with terminal residues of fucopyranoside (α-L-Fucp) or glucosyluronic acid (β-D-GlcpA). Rhamnogalacturonan II (RG-II) has a backbone of homogalacturonan and consists of α-1,4-linked galacturonic residues which are highly substituted with different disaccharide and oligosaccharide chains (arabinose, rhamnose, fucose and other modified sugars) joined to C2 or C3 of the main chain (Wong 2008, Seveno *et al.,* 2009).

1.6 Carbohydrate active enzymes from the human gut microbiota

1.6.1 Glycoside hydrolases (GH)

The enzymes that hydrolyse the glycosidic bonds between two sugars or between a carbohydrate and non-carbohydrate moiety are named glycoside hydrolases (GH) (EC 3.2.1.x). GH enzymes are widespread (present in archaea, bacteria and

eukaryotes) and have significant importance in industrial, medical and biochemical applications (Lynd *et al.*, 2002). The hydrolysis step occurs via general acid catalysis, which requires a proton donor and a nucleophile or base (Henrissat 1991). The reaction can be based on retaining or inverting mechanisms of the anomeric carbon (Figure 1.2). The retaining mechanism of GH enzymes is a two-step reaction in which two amino acid residues are involved. Firstly, the nucleophilic residue acts on the anomeric centre to form a glycosyl enzyme intermediate. The second residue – the acid catalyst – participates in cleavage of the glycosidic bond. In the second step (known as the deglycosylation step), the glycosyl bond is hydrolyzed by water, with the other residue now acting as a base catalyst, deprotonating the water molecule as it attacks and cleaves the bond. During the hydrolysis the anomeric carbon configuration is retained. In contrast, inverting GH enzymes change the configuration of the anomeric carbon via a single nucleophilic displacement (Henrissat 1991). The classification of GH enzymes is based on the type of reaction the enzyme catalyses and on substrate specificity according to IUB Enzyme Nomenclature. An alternative classification is based on the enzyme sequence, which reflects structural features of the protein. Members of GH families have a similar three-dimensional structure, common mechanism of hydrolytic reaction and the catalytic residues identified for all the enzymes within GH family. The classification of GH enzymes into families was initiated by Bernard Henrissat, and led to the creation of 35 families (Henrissat 1991). The well-maintained database named CAZy (Carbohydrate Active Enzyme, http://www.cazy.org/) reports close to 300 families of catalytic and ancillary modules with over 100,000 non-redundant entries (Cantarel *et al.*, 2009).

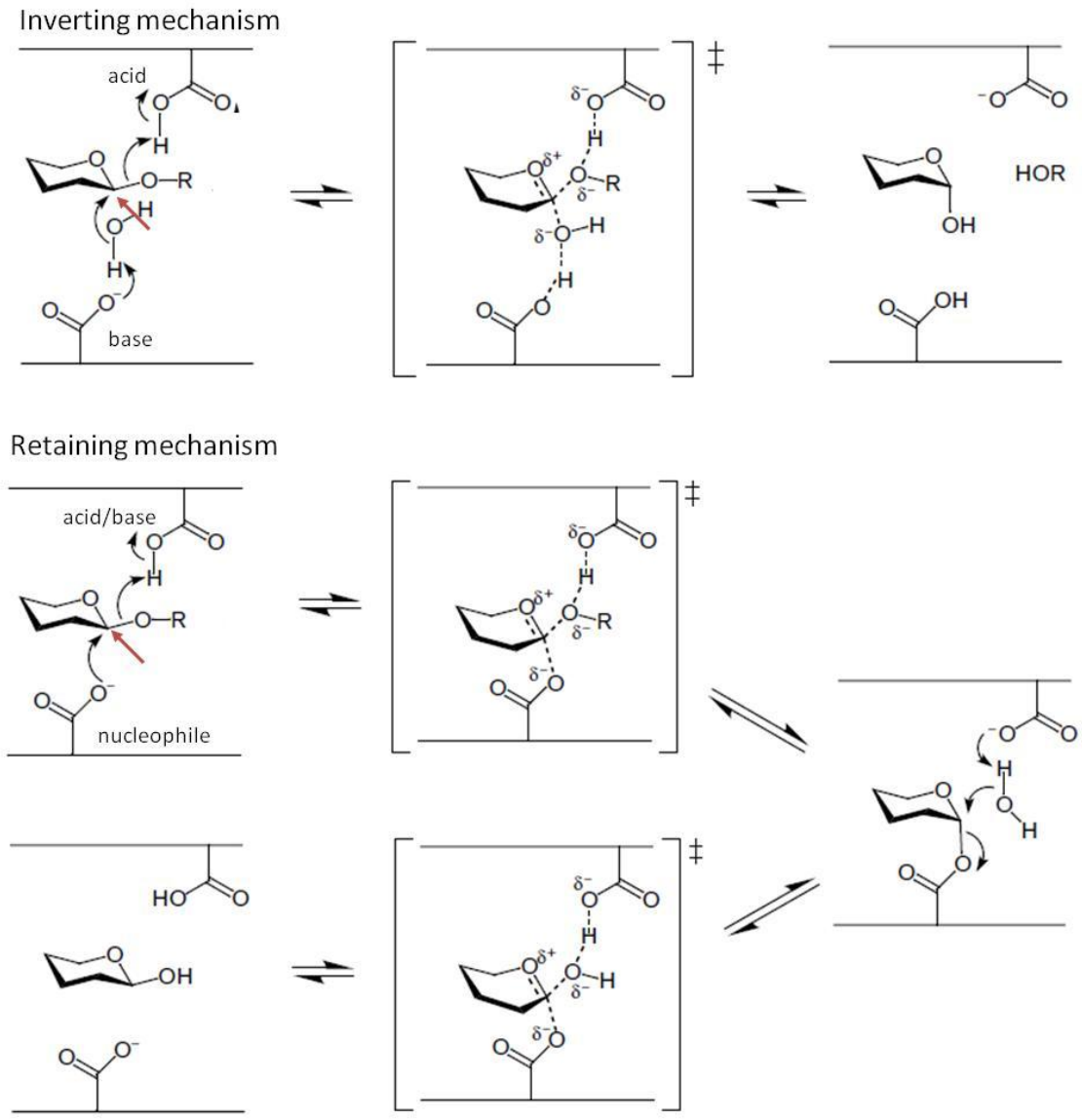


Figure 1.2 Catalytic mechanism for glycoside hydrolases.

The anomeric centre of reaction is shown with the red arrow. Adapted from Rye and Withers (2000).

1.6.2 Microbial hydrolysis of starch

The microbial degradation of starch (Figure 1.3, Table 1.1) relies on the production of several enzymes, including amylases (cleave α-1,4 bonds), pullulanases (cleave 1,6-linkages) and amylopullulanases (cleave α-1,4 and/or α-1,6 bonds). The degradation of starch yields glucose, maltose, maltotriose and other oligosaccharides (MacGregor *et al.,* 2001). The starch-degrading enzymes have been classified to several glycoside hydrolase families including family 13, 14, 15, 57 and 119. The majority of enzymes from amylolytic bacteria fall into glycoside hydrolase family 13 (CAZy database) with more than 8,000 entries. The α-amylases catalyse the hydrolysis of α-glucans with the retaining mechanism of reaction. Within the GH13 family numerous substrate specificities (enzymes acting on maltose, isomaltose, sucrose and trehalose) have emerged but well-conserved sequence stretches are recognised in all GH13 family members. There are four conserved regions and three catalytic residues corresponding to Asp206, Glu230 and Asp297 of Taka-amylase A from *Aspergillus oryzae* – the first amylase examined by X-ray crystallography (Oslancova and Janecek 2002). The first aspartate acts as a catalytic nucleophile, the glutamate is a general acid/base residue and the second aspartate stabilizes the transition state during the retaining mechanism of the enzymatic reaction (MacGregor *et al.,* 2001). Bacterial amylase enzymes show a multi-domain architecture with a conserved catalytic domain and an N- or C-terminal carbohydrate binding module (CBM). The main role of the CBM is to bind starch, deliver the substrate to the catalytic domain and disrupt the surface of the starch granules (Machovic and Janecek 2006).

The known amylolytic bacteria from the human gut belong to the genera *Bifidobacterium*, *Bacteroides*, *Roseburia* and *Butyrivibrio.* The genus *Bifidobacterium* was shown to be a dominant amylose-degrading group with several novel strains isolated by Ryan *et al*. (2006). Bifdobacteria were recovered from starch (*B. adolescentis*) during a study conducted by Leitch *et al*. (2007) which showed that primary colonization of insoluble substrates is restricted to certain groups of gut bacteria. A strong response of *Bifidobacterium* species in one individual fed a diet high in resistant starch was also observed by Walker *et al.* (2011). The *in vitro* study of *Roseburia intestinalis* and *Butyrivibrio fibrisolvens*

from cluster XIVa showed the efficient degradation of starch compounds. Both bacteria produced cell-associated amylase enzymes that enable utilization of a variety of starches (Ramsay *et al.,* 2006). This microbial group was also stimulated by a resistant starch (RS) diet in several human volunteers (Walker *et al.,* 2011). This study also showed that the proportions of *R. bromii* from cluster IV ruminococci group increased dramatically in the majority of volunteers fed on the RS diet. This was in agreement with a previous observation, which also reported *R. bromii*-related bacteria in faecal samples from humans whose diet was supplemented with RS (Abell *et al.,* 2008). Both studies propose a crucial role of this group of bacteria in resistant starch utilisation.

The Gram-negative phylum Bacteroidetes is represented by several highly specialised starch degraders like *B. thetaiotaomicron*, *B. ovatus* and *B. distasonis*. In addition to *in vitro* studies on *Bacteroides* starch metabolism initiated by Salyers *et al.* (1977), sequencing data emphasised their important position in polysaccharide degradation in the distal GI tract (Xu *et al.*, 2007). The starch utilisation system (sus) of *B. thetaiotaomicron* VPI-5482 is a well-studied example of microbial adaptation to carbohydrate metabolism. Sus is a multi-protein cell-associated complex encoded by eight adjacent genes, forming the cluster SusRABCDEFG. The outer membrane proteins encoded by SusC and SusD are involved in substrate binding and uptake. The roles of SusE and SusF are not fully understood, but they are predicted lipoproteins, exposed to the external environment, possibly involved in glucan binding and rendering it less accessible to other gut bacteria. SusG is a outer-membrane neopullulanase with α-1,4-glycosidic activities against amylose, amylopectin and pullulan. Internal cleavage by SusG produces molecules which are subsequently transported to the periplasmic compartment by SusC. In the periplasm oligosaccharides are degraded further by glycoside hydrolases SusA and SusB. The transcriptional regulation of the Sus cluster is accomplished by the SusR protein and occurs in the presence of starch (Martens *et al.,* 2009, Flint *et al.*, 2008).

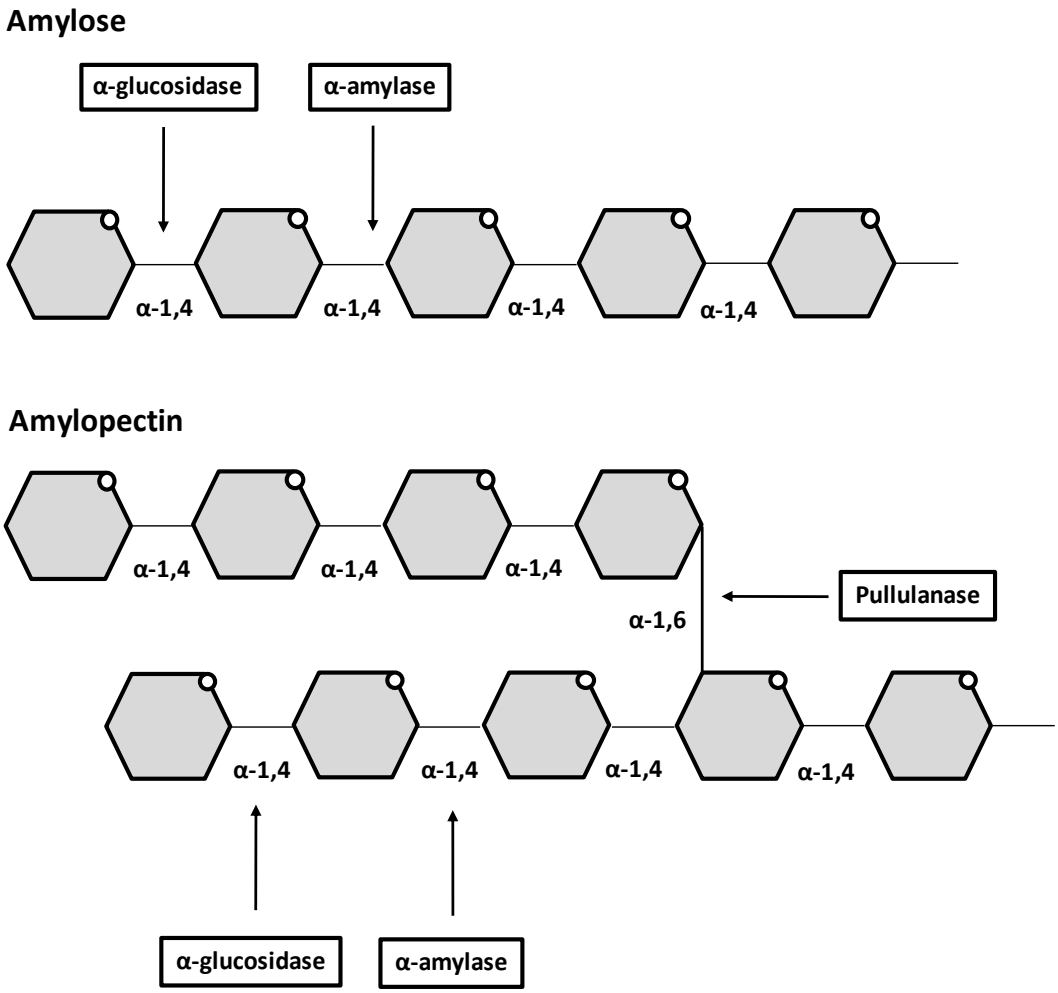


Figure 1.3 Schematic overview of structure and enzymes involved in hydrolysis of amylose and amylopectin.

Ring: glucose

1.6.3 Hydrolysis of non-starch polysaccharides

The breakdown of cellulose fibres requires the action of several GH enzymes (Figure 1.4, Table 1.1): endoglucanase (endo β-1,4-D-glucan hydrolase EC 3.2.1.4), exoglucanase (exo β-1,4-D-glucan cellobiohydrolase, EC 3.2.1.91) and β-glucosidase (EC 3.2.1.21). Endoglucanases cut at random internal sites in the cellulose chain, yielding oligosaccharides of various lengths. Exoglucanases act at the reducing or non-reducing end of a cellulose chain, liberating cellobiose or glucose (Xu *et al.,* 2003, Xu *et al.,* 2007, Lynd *et al.,* 2002). The hydrolysis of cellobiose to glucose residues is catalysed by β-glucosidases (Lynd *et al.,* 2002). The cellulases can have a processive or non-processive mode of action. The former enzymes remain bound to the cellulose substrate and continue cleaving down the polysaccharide. The latter enzymes are detached from the substrate after one round of hydrolysis (Pereira *et al.,* 2009).

Microbial cellulases are classified into various GH families (Table 1.1). The most abundant families are GH5 and GH9. Breakdown of the β-1,4 glycosidic linkage is processed through the retaining (GH5) or inverting (GH9) mechanism of the anomeric carbon. Two conserved catalytic residues are found for the active site of cellulases. Two glutamate residues are characteristic for cellulases from GH5 family (Posta *et al.,* 2004). The GH9 enzymes display the presence of a conserved glutamate and aspartate residues (Pereira *et al.,* 2009).

Cellulases exhibit a complex multi-modular architecture (Gaudin *et al.,* 2000, Kurokawa *et al.,* 2002, Devillard *et al.,* 2004). The most common arrangement consists of a catalytic domain joined by a short linker to a carbohydrate binding module (CBM) which can be at the N-terminus or C-terminus of the protein (Guillen *et al.,* 2010). They are present in cellulases of different families and were shown to enhance the hydrolytic activity of the enzyme by prolonged contact with the complex substrate. The removal of CBMs decreases the binding capacity of the enzymes, resulting in a partial or complete loss of catalytic activities. The disruption of the cell wall polysaccharides mediated by CBMs was also observed (Herve *et al.,* 2010). The CBMs present in cellulases have a wide range of binding specificities and are often combined with the catalytic domains of xylanases, mannanases and pectinases (Herve *et al.,* 2010). Some of the cellulases contain several CBMs and other accessory modules such as fibronectin type III domains (Fn3) or Ig-like domains.

The bacterial fibronectin type III domains are found in extracellular glycohydrolases displaying different enzymatic activities such as cellobiohydrolase (Shen *et al.*, 1995, Kataeva *et al.*, 2002), pullulanase (Matuschek *et al.*, 1994) or endoglucanase (Chiriac *et al.*, 2010). The main function of Fn3 domain in GH enzymes is to facilitate interaction between the catalytic domain and the substrate by separating cellulose chains and exposing the substrate for the cleavage (Kataeva *et al.*, 2002). Ig-like domains were shown to be essential for the activity of the enzymes (Liu *et al.*, 2010). Deletion of an Ig-like domain led to complete inactivation of CbhA – a cellobiohydrolase from *Clostridium thermocellum* (Kataeva *et al.*, 2004). In addition to carbohydrate binding modules, cellulases possess dockerin modules at the C-terminus of the protein. The dockerins are non-catalytic modules which interact with cohesin modules on the scaffoldin part of the cellulosome and are essential for its assembly (Fontes and Gilbert 2010) (see section 1.6.6).

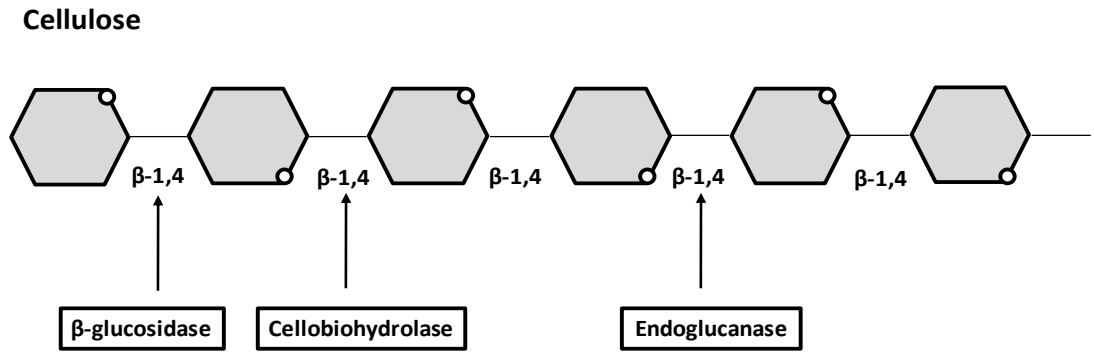


Figure 1.4 Structure of cellulose chain and the main enzymes involved in its hydrolysis.

Ring - glucose

Similarly to cellulose breakdown, hemicellulose degradation requires the action of several enzymes with different activities (Figure 1.5, Table 1.1). The main backbone of xylan is hydrolysed internally by endoxylanases (β-1,4-xylan xylanohydrolase EC 3.2.1.8) which cleaves β-1,4-xylosidic bonds. The xylo-oligosaccharides produced by endoxylanases are cleaved by β-xylosidase (EC 3.2.1.37) which removes xylose from the non-reducing ends of the xylan chain. The further degradation of xylan is mediated by enzymes which remove side chain substitutes such as acetyl- and glucuronyl-groups and arabinose. The cleavage of acetyl groups is catalysed by

acetylxylan esterase (EC 3.2.1.72). α-glucuronidases act on α-1,2-links between the main chain of xylan and removes α-D-(4-*O*-methyl)glucuronosyl residues. The arabinofuranosyl groups are removed from the main chain by action of α-1,3-arabinofuranosidase (Dodd *et al.,* 2011). The specific enzymes are required to hydrolyse xyloglucan, glucomannan and arabinoxylan (Table 1.1). The enzymes involved in hemicellulose degradation are classified as belonging to many different GH families with various modes of action (retaining and inverting) (Dodd *et al.,* 2011, Gilbert *et al.,* 2008). The xylanases display a modular structure with carbohydrate binding modules append to catalytic domains (Selvaraj *et al.,* 2010). The accessory enzymes involved in the removal of single sugars from non-reducing ends (glucosidases, xylosidases etc) usually do not require the assistance of CBMs (Harvey *et al.,* 2000).

Pectinolytic activity depends on the production of a variety of enzymes. The chain is hydrolysed by polygalacturonase via random hydrolysis of α-1,4-D-galactosiduronic linkages in homogalacturonan and other galacturonans (EC 3.2.1.15). The breakdown of rhamnogalacturonan requires the action of rhamnogalacturonase (RG-hydrolase), which is an endo-acting hydrolase cleaving α-1,2 bonds between galacturonic acid and rhamnose residues. The RG-hydrolase exhibits a single displacement mechanism resulting in the inversion of the anomeric configuration. The cleavage of α-1,4-linkages between rhamnose and galacturonic acid in rhamnogalacturonan requires the action of RG-lyase by β-elimination. The non-reducing rhamnose residues are removed by α-rhamnosidase (EC 3.2.1.40). The side chains of arabinan are cleaved by α-arabinofuranosidase (EC 3.2.1.55) and endo-α-1,5-arabinosidase (EC 3.2.1.99). Endo-β-1,4- galactanases and β-galactosidases efficiently hydrolyse the β-1,4-glucosidic bonds in the highly branched side chains of pectin. Pectin esterase releases the methyl residue linked to the galacturonic acid while pectin acetylesterase releases the acetyl residue linked to the galacturonic acid (Wong 2008).

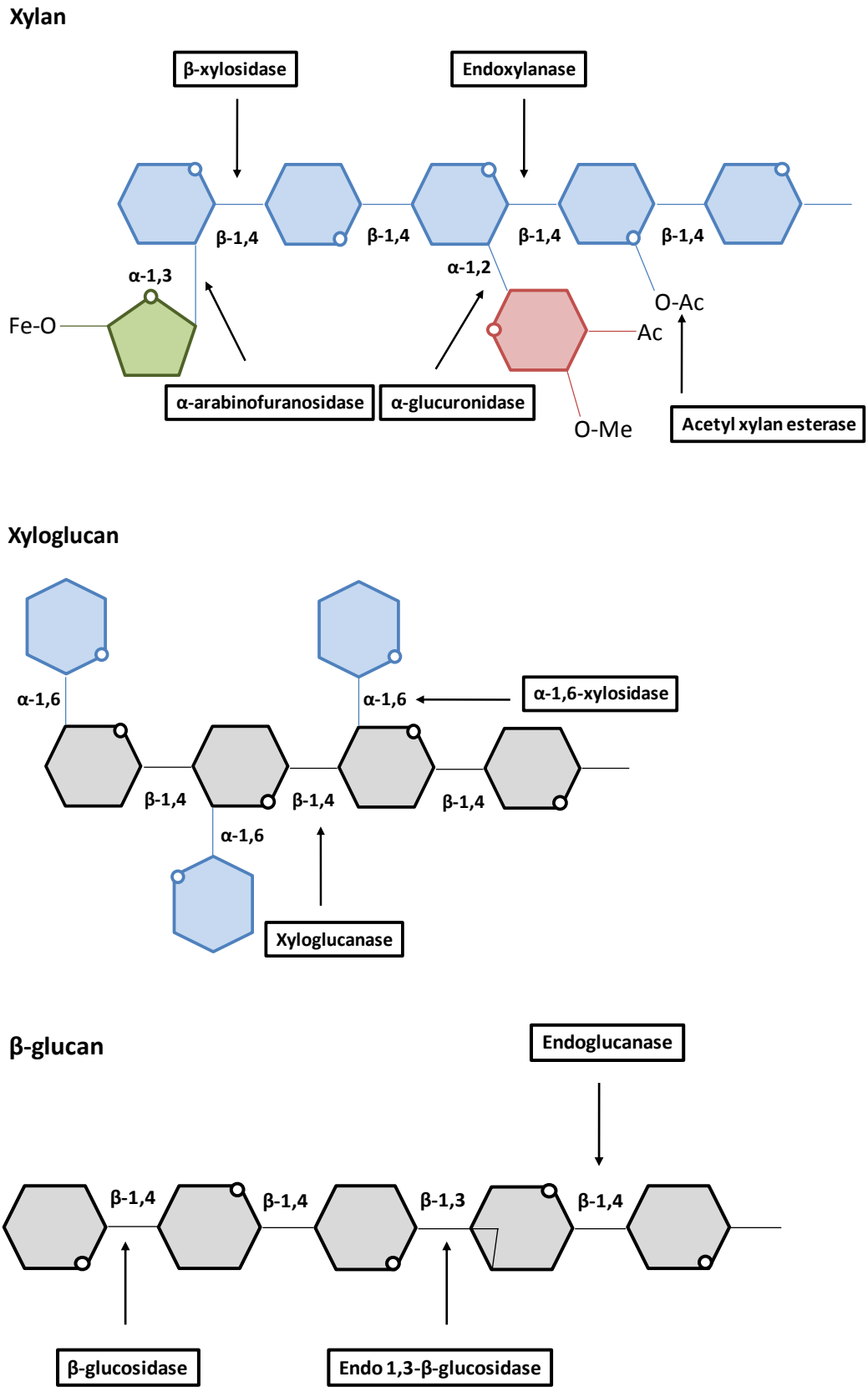


Figure 1.5 Chemical structure and enzymatic hydrolysis of different hemicelluloses.

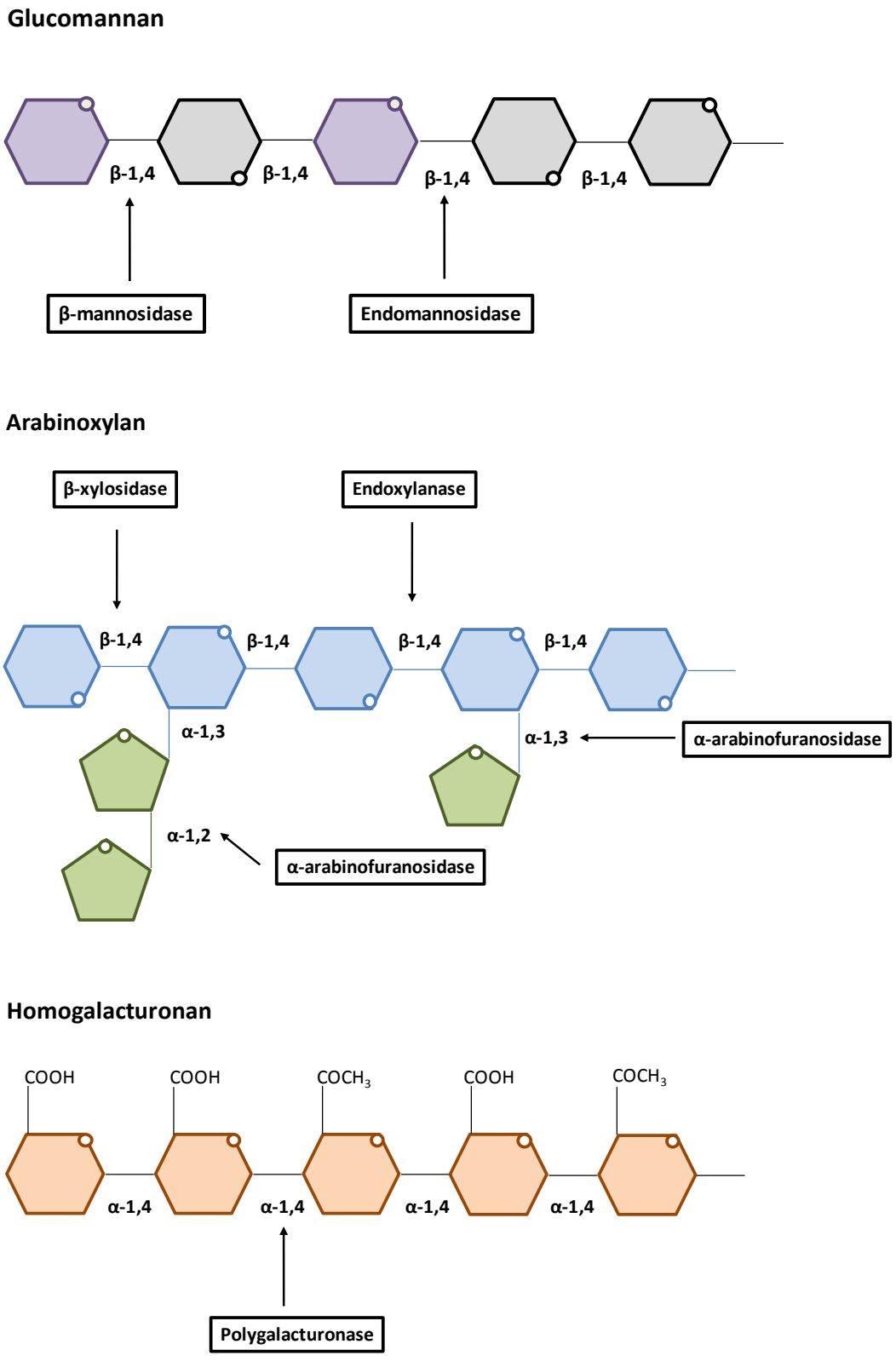


Figure 1.5 Chemical structure and enzymatic hydrolysis of different hemicelluloses.

Grey ring – glucose, blue ring – xylose, green ring – arabinose, red ring – glucuronic acid, brown ring – galacturonic acid, Fe - ferulic acid, Ac – acetyl group, Me – methyl group

1.6.4 Microbial degradation of non-starch polysaccharides

The bacterial degradation of cellulose and hemicellulose has received a lot of attention because of its importance in biomass conversion and biofuel production (Fontes and Gilbert 2010, Morrison *et al.,* 2009). Cellulolytic and hemicellulolytic bacteria comprise several diverse groups occupying different environments; fermentative Gram-positive anaerobes such as *Clostridium, Ruminococcus*, *Butyrivibrio* and aerobic Gram-positive bacteria such as *Cellulomonas* and *Thermobifida* (Lynd *et al.,* 2002). Recent advances in the metagenomics and genome sequencing of human gut bacteria have provided more information on their GH enzymes repertoire involved in non-starch polysaccharides breakdown.

Cellulose degradation in the human gut appears to be restricted to a number of isolates belonging to the genera *Ruminococcus* (Robert and Bernalier-Donadille 2003) and *Bacteroides* (Robert *et al.,* 2007). Robert *et al*. (2003) reported several novel isolates from clostridial cluster IV involved in cellulose utilisation in the human gut. This group was shown to be mainly associated with fibre particles (Walker *et al.,* 2008) and closely related to *R. flavefaciens* – one of the main degraders of cellulose in the GI tract of herbivores. *R. flavefaciens* produces an extracellular, multi-component and multi-enzyme complex known as the cellulosome, whose synergistic activity enables the bacterium to efficiently degrade cellulose fibres (Figure 1.6). The main non-catalytic part of the cellulosome is called scaffoldin, which consists of several cohesin modules. Cohesin modules interact with dockerin domains, which are part of catalytic cellulosomal enzymes. The structure is bound to the bacterial cell surface by a cohesin-containing scaffoldin. The structural components of the *R. flavefaciens* cellulosome are encoded by the *sca* gene cluster. The anchoring scaffoldin ScaE contains a single cohesin domain and is attached to the bacterial surface by a sortase-mediated reaction. The ScaE cohesin binds to the C-terminal dockerin of the primary scaffoldin ScaB, which consists of nine cohesins which either bind to the dockerins from a catalytic subunit or bind to a scaffoldin ScaA. ScaA contains two cohesins which recognise dockerins of catalytic enzymes. Additionally, *R. flavefaciens* contains a single cohesin and a single dockerin-bearing scaffoldin called ScaC, which interacts with the ScaA protein. The cellulose binding occurs via two carbohydrate binding modules present in the protein CttA, which is attached to the cell surface by ScaE. This highly complex system is essential for host

survival, since 70% of herbivore energy intake is derived from bacterial fermentation (Flint *et al.,* 2008, Fontes and Gilbert 2010). To date there are no reports on such an efficient way to utilize cellulose in the human gut. However, as mentioned above, novel cellulolytic strains related to *R. flavefaciens* have been isolated from human subjects (Robert and Bernalier-Donadille 2003). *Ruminococcus champanellensis* was reported to utilise a variety of celluloses and xylan and possesses an array of carbohydrate active enzymes (Chassard *et al.,* 2011). Future studies should provide more information on the metabolism of this bacterium and its role in polysaccharides degradation in the human gut.

The other known highly potent fibrolytic human gut isolates belong to the *Bacteroides* genus including *B. cellulosilyticus.* (Robert *et al.,* 2007), *B. xylanisolvens* (Mirande *et al.,* 2010), *B. intestinalis, B. ovatus* and *B. fragilis* (Dodd *et al.,* 2011). Bioinformatic analysis of a gene cluster from the human isolate *B. ovatus* ATCC 8483 (Whitehead 1995) showed similarity to a xylan utilisation system from the ruminal bacterium *Prevotella bryantii* $B_1$4 (Gasparic *et al.,* 1995, Flint *et al.,* 1997). The xylan utilisation system (Xus) consists of genes homologous to the SUS gene cluster of *B. thetaiotaomicron*. The cluster XusA, XusB, XusC and XusD showed similarities to the SusC and SusD outer membrane proteins. The endoxylanase encoded by gene *xyn10C* possibly represents a functional homologue of the SusG protein catalysing cleavage of xylan polymers to shorter oligosaccharides. Upstream of *xyn10C* is the *xusE* gene encoding a hypothetical protein of unknown function. A study conducted by Dodd *et al.* (2011) showed high conservation of this cluster within the species of *Bacteroides* present in the human gut. Enzymes, involved in the cleavage of sugars such as glucose, galactose or arabinose from the non-reducing end of the polysaccharide chain, are well represented within the human gut microbiota. A wide range of bacterial species from human gut was shown to possess β-glucosidase, β-glucuronidase (Dabek *et al.,* 2008, Gloux *et al.,* 2011), β-galactosidase (Goulas *et al.,* 2007), α-glucosidase (Gloster *et al.,* 2008) and α-arabinofuranosidase activity (Margolles and de los Reyes-Gavilan 2003).

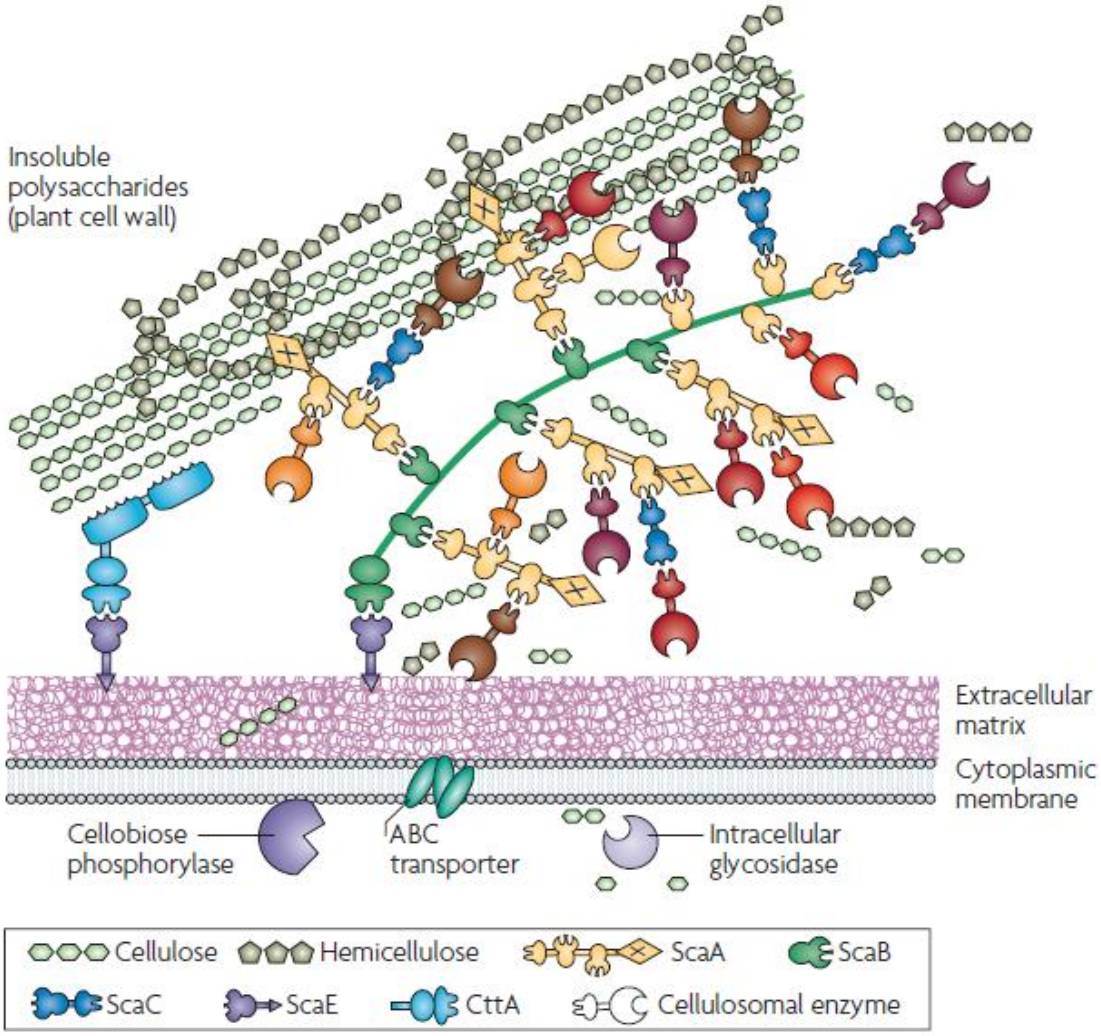


Figure 1.6 Schematic overview of the cellulosome system of *Ruminococcus flavefaciens* involved in plant cell-wall degradation.

Figure is reproduced from Flint *et al.*, (2008) with author's permission.

Enzyme	Name	Action	GH family	Substrate
3.2.1.21	**β-glucosidase**	Hydrolysis of terminal, non-reducing β-glucosyl residues with release of β-glucose	1r, 3r, 9i, 30r, 116r	Cellulose, lichenan, cereal β-glucans, xyloglucan, glucomannan
3.2.1.4	**Endoglucanase**	1,4-β-glucosidic linkages	5r, 6i, 7r, 8i, 9i , 12r , 44r, 45i, 48i, 51r, 61nd, 74i, 124i	Cellulose, lichenan and cereal β-D-glucans
3.2.1.91	**Cellobiohydrolase**	1,4-β-glucosidic linkages in cellulose and cellotetraose, releasing cellobiose from the non-reducing ends of the chains	5r, 6i, 9i, 48i	Cellulose
3.2.1.73	**β-1,3-1,4-glucanase**	Hydrolysis of 1,4-β-glucosidic linkages in β-glucans containing 1,3- and 1,4-bonds	5r, 7r, 8i, 12r, 16r, 17r	Lichenan, cereal β-glucan
3.2.1.74	**Glucan β-glucosidase**	Hydrolysis of 1,4-linkages in 1,4-β--glucans, to remove successive glucose units	1r, 3r	Cereal β-glucan
3.2.1.39	**Endo 1,3-β-glucosidase**	Hydrolysis of 1,3-β-glucosidic linkages	16r, 17r, 55i, 64i, 81i	Cereal β-D-glucan, lichenan
3.2.1.8	**Endoxylanase**	1,4-β-xylosidic linkages in xylans	5r, 8i, 10r, 11r, 43i	Xylan, arabinoxylan
3.2.1.37	**β-xylosidase**	Removal of successive D-xylose residues from the non-reducing terminus	3r, 30r, 39r, 43i, 52r, 54r, 116r, 120nd	Xylan, arabinoxylan,
3.2.1.55	**α-*N*-arabinofuranosidase**	Hydrolysis of terminal non-reducing α-L-arabinofuranoside residues in α-L-arabinosides	3r, 43i, 51r, 54r, 62nd	Arabinans, xylan, pectin
3.2.1.131	**α-glucuronidase**	Hydrolysis of α-1,2-(4-*O*-methyl)glucuronosyl links in the main chain of xylans	67i, 115i	Xylan
3.2.1.151	**Endoglucanase**	1,4-β-glucosidic linkage	5r,12r, 16r, 44r, 74i	Xyloglucan
3.2.1.-	**α-xylosidase**	Removal of unsubstituted D-xylose residues attached to the glucose located at the non-reducing terminus in xyloglucan	31r	Xyloglucan
3.2.1.78	**Endomannosidase**	Random hydrolysis of 1,4-β-mannosidic linkages	5r, 26r, 113r	Mannan, glucomannan, galactomannan
3.2.1.25	**β-mannosidase**	Hydrolysis of terminal, non-reducing β-mannose	1r, 2r, 5r	Glucomannan, galactomannan
3.2.1.22	**α-galactosidase**	Hydrolysis of terminal non-reducing α-galactose residues	4r,27r, 36r, 57r, 97i/r, 110i	Galactomannan
3.2.1.23	**β-galactosidase**	Hydrolysis of terminal non-reducing β-galactose	1r, 2r, 35r, 42r	Pectin
3.2.1.15	**Polygalacturonase**	Random hydrolysis of 1,4-α-galactosiduronic linkages	28i	Homogalaturonan, rhamnogalcturonan II
3.2.1.40	**α-rhamnosidase**	Hydrolysis of terminal non-reducing α-L-rhamnose residues	28i, 78i, 106nd	Rhamnnogalacturonan
3.2.1.99	**Endo-1,5-α-L-arabinosidase**	Hydrolysis of 1,5-α-arabinofuranosidic linkages	43i	Rhamnogalacturonan
3.2.1.89	**Endo-1,4-β-galactanase**	Hydrolysis of 1,4-β-galactosidic linkages in arabinogalactans	53r	Rhamnogalacturonan
3.2.1.1	**α-amylase**	1,4-α-glucosidic linkages in polysaccharides containing three or more 1,4-α-linked D-glucose units	13r, 57r, 119r	Starch, glycogen
3.2.1.41	**Pullulanase**	1,6-α-glucosidic linkages	13r, 57r	Pullulan
3.2.1.20	**α-glucosidase**	Hydrolysis of terminal, non-reducing 1,4-linked α-glucose residues with release of D-glucose	4r, 13r, 31r, 63i, 97i/r, 122nd	Starch, glycogen
3.2.1.135	**Neopullulanase**	Hydrolysis of pullulan to panose (6-α-D-glucosylmaltose)	13r	Amylopectin

Table 1.1 Glycoside hydrolase enzymes involved in dietary polysaccharides degradation.

i = inverting mechanism of hydrolysis, r = retaining mechanism of hydrolysis, nd- not determined

1.7 Metagenomics

The combination of culture-dependent and molecular studies has been very successful determining the microbial diversity and metabolic activities that occurs in the largely inaccessible human GI tract (Qin *et al.,* 2010, Arumugam *et al.,* 2011, Tap *et al.,* 2009). These studies provide useful information which is needed to fully understand the interaction and impact of this highly complex ecosystem on the human health. The remaining challenge is to relate the sequence data to function and identify potential targets for disease therapy. Several large metagenomic surveys have been conducted on DNA recovered from the human gut microbiota in recent years. Metagenomics is a culture-independent approach to study the genetic pool present in an environmental sample. It is a multi-step approach which requires sampling, sample processing, DNA extraction and data analysis based on the sequence or function (Figure 1.7). Sampling and DNA extraction are first crucial steps in the metagenomic application as they affect the downstream procedures. The DNA extraction method depends on the insert size required and the analysis (sequence vs. function). If the aim of the study is high-throughput sequencing, PCR amplification or small-insert size clone library then DNA can be extracted using commercial kits. For high molecular weight (HMW) DNA, a specific protocol should be used and optimised. There are many protocols available to extract HMW DNA from different environmental samples (soil bacteria, marine bacteria and gut bacteria) which yield good-quality insert DNA (Liles *et al.,* 2009, Ouyang *et al.,* 2010, Reigstad *et al.,* 2011, Rosewarne *et al.,* 2011, Ruiz and Rubio 2009).

1.7.1. Sequence-driven analysis

Sequence-based screening of a metagenomic library depends on the identification of homologies between randomly sequenced clones and already characterized genes (Simon and Daniel 2011). This analysis can disclose genes of interest and catalogue the genetic potential, but will not detect fundamentally novel gene functions. A successful sequence-based approach depends on sequencing effort and good microbial coverage of the sample. The development of next-generation high-throughput sequencing (Wommack *et al.,* 2008, MacLean *et al.,* 2009) produces a great number of scattered pieces of sequences that can be assembled into longer contigs which can be then analysed. The analysis of these data has also improved

with a variety of bioinformatics tools for gene analysis e.g. MetaGene and MEGAN (Huson *et al.,* 2011). However, in a highly diverse community such as the human microbiome there is little or no chance of reconstructing complete bacterial genomes even with a deep-sequencing coverage (Qin *et al.,* 2010). The second approach for a sequence-based analysis involves designing DNA probes or primers which are derived from regions of already known genes or protein families (Handelsman 2004). The analysis of sequences from unrelated microbes is difficult, if no related organisms have been sequenced previously.

1.7.2 Function-driven analysis

The function-based approach to metagenomics relies on constructing a library and expressing genes in a heterologous host. The advantage of this method is the ability to access previously unknown genes and their phenotypic traits, which could find application in medicine, agriculture or industry (Taupp *et al.,* 2011). Depending on the insert size, metagenomic libraries have been constructed using different cloning vectors such as plasmids (up to 15 kb), fosmids and cosmids (both up to 40 kb) and bacterial artificial chromosomes (> 40 kb). Small insert libraries are usually produced using plasmids as a cloning vector (Simon and Daniel 2011). Small insert size libraries are employed to identify single genes (mostly enzymes) or small operons. They are usually constructed in *Escherichia coli* as a heterologous host, therefore the transformation efficiency is high ($>10^5$. µg DNA). Large insert size libraries are produced to recover biosynthetic pathways and large clusters of genes involved in the synthesis of complex enzymes and antimicrobial compounds (Kakirde *et al.,* 2010). Cosmids and fosmids are used commonly to produce large insert libraries; their copy number is low (typically single copy) to ensure a high stability of the recombinant gene and therefore they have been used in multiple studies (Rhee *et al.,* 2005, Hardeman and Sjoling 2007, Feng *et al.,* 2007).

Functional analysis is dependent on the expression of recombinant gene in a heterologous host. It requires successful transcription and translation of the gene or genes of interest. Therefore, the full-length gene or a gene cluster needs to be cloned. There are many obstacles that can limit successful gene expression. Codon bias, regulatory elements including promoters, post-translational modification and processing could be barriers to a functional metagenomic study. Conventionally, *E.*

coli is used as a background host for expression studies. However, the *in silico* results presented by Gabor *et al*. (2004) showed that only 40% of genes derived from diverse microbial origins are readily expressed in *E. coli,* with strong variations between different groups of microorganisms. The most readily independently expressed genes (73%) in *E. coli* were predicted to derive from Firmicutes. In contrast, only 7% of Actinobacteria genes were predicted to be independently expressed in *E. coli.* In order to increase the chance of expressing the metabolic potential hidden in complex ecosystems like the human gut microbiota, multiple expression systems can be used. The studies conducted by Martinez *et al*. (2004) and Craig *et al.* (2010) demonstrated that broad host screening is likely to increase the number and diversity of positive clones from functional metagenomic studies. The alternative hosts which have been used for the functional screening included *Agrobacterium tumefaciens*, *Burkholderia graminis*, *Caulobacter vibrioides*, *Pseudomonas putida* (Craig *et al.,* 2010), *Streptomyces lividans* (Meilleur *et al.,* 2009), *Thermus thermophilus* (Angelov *et al.,* 2009), *Xanthomonas campestris* and *Pseudomonas fluorescens* (Aakvik *et al.,* 2009).

Function-based screening needs reliable and high-throughput assay methods for detection of the activity of interest. This could for example be based on a phenotype that is readily visible. Plating colonies on a substrate-containing medium and observing visible changes in colony colour or production of clear zones around the colony is one of the potential screening methods for the function–based approach (Taupp *et al.,* 2011). It is an easy method that can be adapted for high-throughput screening for a large number of clones by applying colony picking robots, microplate readers and liquid handlers. Automation of the functional screening process increases the number of screened clones, reduces the required labour and improves reproducibility. Detection of an enzyme activity based on screening on substrate-containing plates was successfully used to recover a wide range of biocatalysts such as lipases (Berlemont *et al.,* 2011, Glogauer *et al.,* 2011), amylases (Yun *et al.,* 2004a, Sharma *et al.,* 2010) and cellulases (Duan *et al.,* 2009, Liu *et al.,* 2011, Kim *et al.,* 2011). Another function-based approach is the heterologous complementation of host strains or mutants. In this case, growth of clones is observed only if they encode the gene of interest and produce the active compound. This method is simple and fast and enables the screening of large libraries (Schallmey *et al.,* 2011).

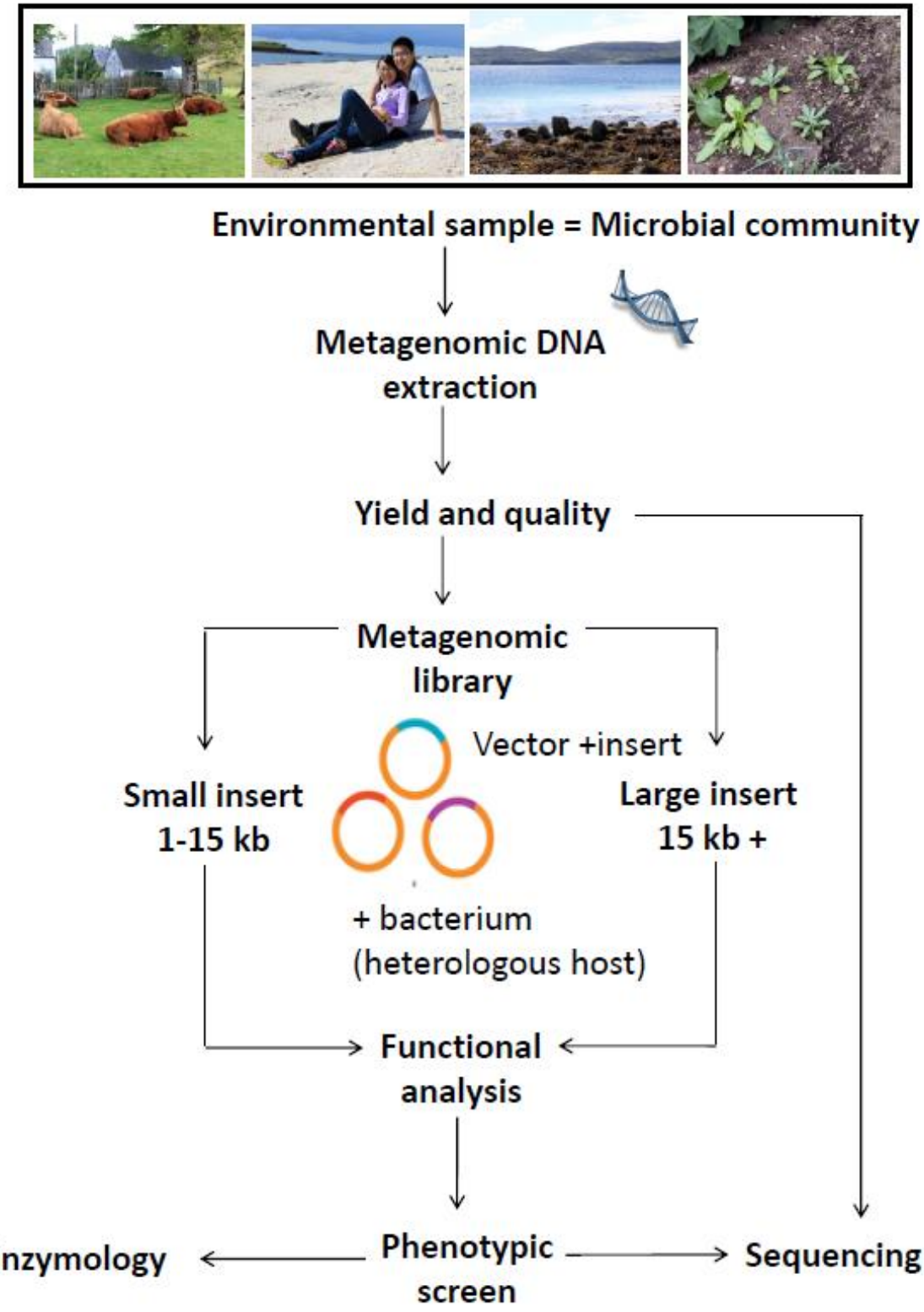


Figure 1.7 An overview of processes involved in production of a metagenomic library and functional screening.

The functional screening of metagenomic libraries which relies on substrate induced gene expression – SIGEX was used by Uchiyama and Watanabe (2008). This technique requires construction of an expression vector that carries a promotorless green fluorescence protein gene – *gfp* (p18GFP). The target gene, encoding the desired trait, is cloned from metagenomic DNA upstream of the *gfp* gene and is induced by the substrate resulting in the production of GFP protein allowing selection of positive and negative clones by utilizing fluorescence–activated cell sorting.

1.8 Assessment of microbial functionality from the human GI tract using metagenomics

The metagenomic approach has been used to detect novel GH enzymes (Venter *et al.*, 2004, Noonan *et al.*, 2005, Tyson *et al.*, 2004). A great deal of attention has been paid to soil metagenomics (Daniel 2005, Baveye 2009) since the likelihood of identifying uncultured and previously unknown microbes is high, and the environment is a rich source of new biocatalysts. Soil samples taken from different parts of the world became a starting point for many studies which led to the identification of novel amylases (Yun *et al.*, 2004), cellulases (Liu *et al.*, 2011), glucosidases (Jiang *et al.*, 2009) and other highly potent enzymes (Li *et al.*, 2009).

The gut microbiota is an obvious place to search for novel GH enzymes. The microbiota of grazing animals such as the bovine rumen community (Hess *et al.*, 2011, Chen *et al.*, 2010) and the rabbit GI tract (Feng *et al.*, 2007) were screened for enzymatic activities. Hess and colleagues (2011) reported a high-throughput sequencing metagenomic project of the gut microbiota from the cow rumen. They predicted around 27,000 candidate genes with significant similarities to GH enzymes. The active clones expressed in *E. coli* showed the ability to degrade lichenan, cellulose, xylan and potential biofuel feed-stocks – switchgrass and miscanthus. The wide range of GH enzymes was recovered from the metagenomic of bovine rumen by Ferrer *et al.* (2005). Positive clones of the small-insert expression library showed the activity to utilise different substrates of plant cell wall origin. The latter approach was used by Palackai *et al.*, (2007) and led to discovery of multi-functional GH enzymes from ruminant microbiota. The native herbivore of Australia,

the Tammar wallaby, was the subject of another study conducted by Pope *et al.* (2010). They reported more than 800 putative GH and associated modules including 41 dockerin modules. The sequence-based approach was also applied to the wood-feeding termite metagenome by Warnacke and colleagues (2007). It led to the identification of more than 700 glycoside hydrolase catalytic domains corresponding to 45 different families, including not surprisingly a great diversity of cellulases and hemicellulases.

The human gut microbiota has also been studied extensively in recent years by applying metagenomic approaches. These studies have focused mainly on better understanding the relationship of this highly complex ecosystem and human health. The common effort of the scientific community is to relate the sequence-derived data to function and clearly identify the potentially harmful and beneficial states of the microbial community in the human GI tract. Several studies have applied deep-sequencing analysis of the human microbiome and presented results on the diversity and function of the gut community. The sequence-based metagenome studies reported abundance of genes involved in glycan metabolism of the human gut bacteria. It estimated the presence of number of different GH families, which should allow sufficient fermentation of fibre material and which are not encoded by the human genome. Numerous key genes involved in SCFA production – the end-product of carbohydrate metabolism of gut bacteria were also reported (Gill *et al.*, 2006). A similar conclusion was drawn by Kurokawa *et al.* (2007), who also identified a great number of carbohydrate metabolism genes in the adult but also the infant microbiome, suggesting that the gut microbiota is programmed to utilise plant derived carbohydrates to some extent before weaning. As part of the international collaboration called MetaHIT (Metagenomics of the Human Intestinal Tract), scientists produced a catalogue of genes present in the human intestine (Qin *et al.*, 2010) that should improve the understanding of the underlying mechanisms of host/microbe interactions. This information could determine the differences between individuals, healthy versus sick (IBS, IBD etc) at the level of shared bacterial species and their functionality. Another scheme called Human Microbiome Project (HMP) aims to establish a reference catalogue of the microbiome from different sites of the human body including skin, oral cavity, GI tract and urogenital tract. Both programmes will sequence around 1,000 genomes of bacterial species isolated from

human volunteers (Dusko Ehrlich and MetaHIT consortium 2010). The first report on the human gut microbial gene catalogue was presented in 2010 (Qin *et al.*, 2010). The MetaHIT consortium analysed DNA extracted from 124 volunteers by applying next-generation sequencing methods and predicted a 'minimal gut metagenome' which is present in the gut samples of most or all individuals. The enzymes involved in biodegradation of dietary-derived carbohydrates (pectin, cellulose, mannose, and fructose), the synthesis of vitamins, amino acids and SCFAs and the ability of the gut microbiota to degrade numerous xenobiotics are part of the minimal gut metagenome (Qin *et al.*, 2010). The other study directed by MetaHIT consortium showed that people around the world can be classified based on their gut microbiota into three enterotypes (Arumugam *et al.*, 2011). The studies cited above produced a great number of sequencing data with millions of predicted genes encoded by the gut microbiota. However, they also reported that most of these data could not be assigned to a specific functionality group since they encoded uncharacterised orthologous groups and completely novel gene families (Qin *et al.*, 2010). Therefore the combination of sequence- and function based analysis should be combined to successfully predict the metabolic potential of human gut microbiota. A study conducted by Tasse *et al*. (2010) is an example of a combination of sequence and function based approach to determine the abundance of GH enzymes in the human gut. The scientists showed prevalence of proteins involved in carbohydrate transport and metabolism and the sequenced genes were assigned to 35 known CAZy families with activities such as glucanases, xylanases, amylases, pectinases and galactanases.

1.9 Lactococcus lactis *as a heterologous host*

Several alternative heterologous hosts have been used for metagenomic library screening. This project investigated the use of the Gram-positive bacterium *Lactococcus lactis* as an alternative heterologous host for a functional screening of a metagenomic library. *L. lactis* is widely recognized as an attractive alternative heterologous host to the *E. coli* expression system (Kunji *et al.*, 2003, Le Loir *et al.*, 2005). The positive correlation between codon usage of individual gene and surrogate host has been reported (Kurland 1991). It was hypothesized that the genes derived from low %G+C gut Firmicutes would be expressed in a host such as *L. lactis* which has a similar codon usage.

Lactococcus lactis is a Gram-positive low %G+C coccus belonging to the phylum Firmicutes. The ability of *L. lactis* to ferment lactose to lactic acid, hydrolyse casein and ferment citric acid makes it important for the food industry. It is used as a starter culture in dairy product manufacturing including cheese, yogurt and fermented milk processing. It performs bioconversions in fermented meats and vegetables. The lactic acid bacteria (LAB) include several genera with similar metabolic capabilities. Within the group there are industrially important genera, including *Pediococcus, Streptococcus, Leuconostoc, Lactococcus* and *Lactobacillus* species (Pfeiler and Klaenhammer 2007). *Lactococcus lactis* MG1363 was chosen in the present study to extend the expression host range for the functional screening of metagenomic libraries. *L. lactis* MG1363 is a plasmid-free strain (Wegmann *et al.,* 2007) obtained through the sequential protoplasting and regeneration of *L. lactis* NCDO712, which led to the creation of strains that retain none of the plasmids (MG1363) or only pLP712 (MG1299) (Gasson 1983). The genome sequence of this strain and a significant number of well established molecular techniques are available. Strain MG1363 does not produce any extracellular proteases, which is highly beneficial if the heterologous product is secreted. Therefore, *L. lactis* MG1363 has been employed as a cell factory for the production of macromolecules (bacteriocins), enzymes and metabolites (Morello *et al.,* 2008). It has also been used as an oral vaccine delivery system for different antigens and cytokines (Morello *et al.,* 2008). *L. lactis* MG1363 is a model microorganism used worldwide and alongside other lactic acid bacteria, is a generally regarded as a safe (GRAS) organism.

1.9.1 Bacterial gene expression in the metagenomic library

The expression of the recombinant gene in the metagenomic library has to be assured in order to detect clones producing the protein of interest. There are two approaches that enable expression of a cloned gene. Firstly, the expression and translation of cloned gene utilises a promoter and ribosomal binding site (RBS) specific to the cloning vector. This approach is commonly used if large quantities of desired product are needed. Often the strong promoter system may have a toxic effect to the bacterial host or can lead to a slower growth; therefore various inducible promoter systems are available to overcome this problem (Baneyx 1999). Secondly, the cloned gene can be under the control of its own promoter, meaning that both the promoter and RBS are

provided by the insert. The expression of the gene relies on the compatibility with the transcription/translation machinery of the background host bacteria. This method assures better stability of the recombinant genes and could minimise a potentially toxic effect on a surrogate host. However, if the expression signals are not recognised by the background bacterium the gene encoding potentially interesting trait can be lost.

Transcription of a bacterial gene starts with the binding of a sigma factor to the promoter which is a specific sequence in DNA. The main housekeeping sigma factor in *E. coli* is σ^{70} which is responsible for recognition of promoters of genes essential for bacterial survival (Browning and Busby 2004). In the Gram-positive bacterium *Bacillus subtilis,* σ^{43} (Voskuil and Chambliss 1998) and in *L. lactis* σ^{39} (Araya *et al.,* 1993) are the main sigma factors homologous to *E. coli* σ^{70}. In *E. coli,* σ^{70} binds to two conserved DNA regions, TATAAT (Pribnow box) and TTGACA, located at positions -10 and -35 upstream of the transcription initiation start point of the gene, respectively (Browning and Busby 2004, Gruber and Gross 2003). In *L. lactis* both regions are well conserved (Jeong *et al.,* 2006). In addition, *L. lactis* promoters contain a TG motif, located one base pair upstream of the -10 region. In other Gram-positive bacteria the TG motif is also well conserved (Voskuil and Chambliss 1998). The spacing between the -10 and -35 region is an important factor for gene expression. *E. coli* promoters have usually 17 ± 1 nucleotides between both regions. For *L. lactis* the spacing is usually longer than observed in *E. coli*. The start of transcription is located 7 ± 1 nucleotides downstream of the -10 region (Jensen and Hammer 1998).

Most bacteria contain additional sigma factors which recognise alternative promoter sequences, and are involved in expression of genes in response to particular changes in the surrounding environment e.g. an increase in temperature or a lack of nutrients. *E. coli* and *B. subtilis* contain seven and 18 alternative sigma factors, respectively (Dale and Park 2010). *L. lactis* does not have any stress-related alternative sigma factors (Bolotin *et al.,* 2001). Once the mRNA is synthesised, the next step in gene expression is to translate it into a functional protein. The process of translation (protein synthesis) starts at a specific sequence of the mRNA called the ribosomal binding site (RBS or Shine-Dalgarno) which is usually located up to 7 bases upstream to translational start codon (Shine and Dalgarno 1974). The RBS sequence

is complementary to the 3'-end of bacterial 16S rRNA. The Shine-Dalgarno sequences in *L. lactis* resemble those of *E. coli* based on their average free energy value (ΔG). However, the greater complementarily of RBS to the 16S rRNA sequence has been reported for Gram-positive bacteria (Van De Guchte *et al.*, 1992). The separation between RBS and the translational start codon determines the strength of gene expression (Baneyx 1999).

Another factor that can suppress the translational efficiency of the recombinant gene is codon usage. If the expressed gene contains rare codons, which are infrequently used by the heterologous host, then problems during translation can occur. The codon usage of *E. coli* and *L. lactis* differs due to the different %G+C content. *E. coli* infrequently uses ATA for isoleucine, CTA for leucine and AGG and particularly AGA for arginine (Appendix 3) (Chen and Inouye 1994). The rare codons in *L. lactis* are CTG and CTA for leucine and AGG and CGG for arginine. Overall *L. lactis* shows a preference for codons with A or T at the wobble position which reflects its low %G+C content. Codon optimisation is a recognised procedure to increase expression of heterologous genes. It requires either replacement of rare codons and usage of the optimum start and stop codon or supplementation of a bacterial host with tRNA genes encoding tRNAs that recognise these rare codons (Fuglsang 2003). The phylogenetic distance and codon usage between the gene donor and the host organism may have a significant impact on successful gene expression. This agrees with the hypothesis that a similar %G+C content of surrogate host and recombinant gene will lead to a similar codon usage and therefore higher gene expression (Kurland 1991, Kurland and Gallant 1996). Therefore, expression of genes derived from low %G+C gut Firmicutes should be more likely to be achieved in *L. lactis* since their codon usage is similar.

1.9.2 Protein secretion and targeting

Synthesized bacterial proteins can remain in the cytoplasm or be exported outside the cell (secreted). A secreted protein can be tethered to the cell surface or released into the extracellular environment. Bacteria have evolved several transport systems which allow successful secretion of proteins to the extracellular space. The two main secretory pathways are the Sec-dependent (Sec) and twin-arginine translocation (Tat) systems (Mergulhão *et al.*, 2005). Proteins for secretion are tagged with an N-

terminal signal peptide (SP) (Figure 1.8) which allows the bacterium to distinguish them from cytoplasmic proteins. The signal peptide for the Sec-pathway contains a positively charged N-terminus (N region), followed by a hydrophobic region (H) and a short cleavage region (C) with the consensus sequence A-X-A for the signal peptidase type I (which removes the tag from the protein) (Mergulhão *et al.,* 2005, Natale *et al.,* 2008). The *L. lactis* MG1363 genome encodes one signal peptidase type I (SipL), in contrast to five enzymes encoded by *B. subtilis* (Wegmann *et al.,* 2007). A second type of signal peptide is characteristic for lipoproteins and consists of shorter N- and H- regions followed by a consensus cleavage site called lipobox, which is recognised by signal peptidase type II encoded by *lspA* (Wegmann *et al.,* 2007, Hutchings *et al.,* 2009). The proteins secreted via Tat-system are tagged with a signal peptide which contains the twin arginine motif (SRRxFLK) at the junction between N- and H-regions. The Tat signal peptides are longer and less hydrophobic than the signal peptide of Sec-proteins (Natale *et al.,* 2008).

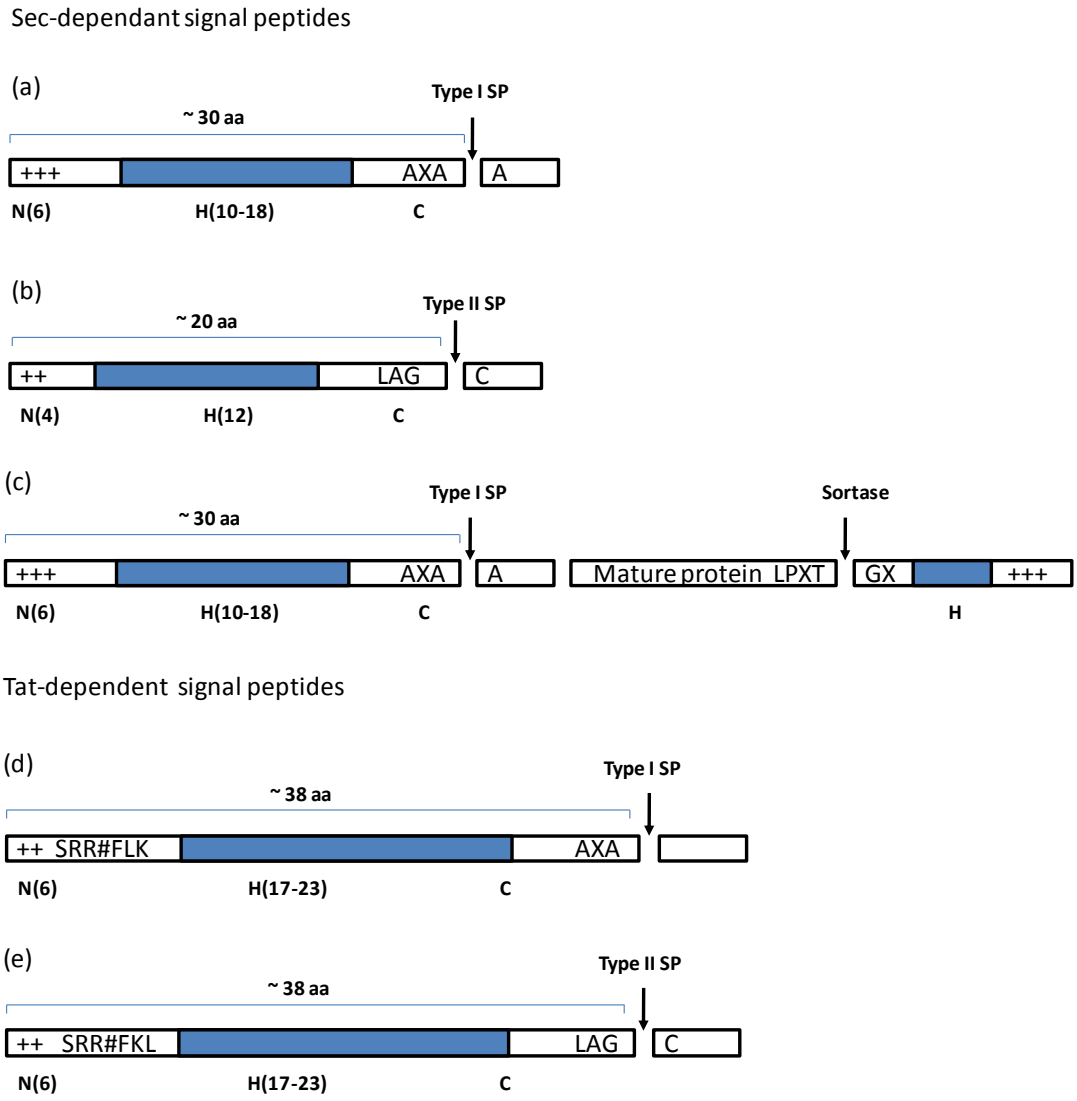


Figure 1.8 Schematic overview of the signal peptides of bacterial proteins.

The N-terminal (N), hydrophobic (H) and cleavage (C) regions are identified, the length in amino acids is presented in brackets. Cleavage sites are indicated by arrows. (a) – Sec-dependent signal peptide cleaved by type I signal peptidase at the AXA cleavage site. (b) – lipoprotein signal peptide cleaved by type II signal peptidase at LAG-C motif known as lipobox. (c) – signal peptide of cell wall anchored protein with C-terminal sortase motif LPXTGX cleaved by sortase, followed by hydrophobic region and positively charged tail. (d) – Tat-dependent signal peptide with twin-arginine motif (SRR#FKL), cleaved by signal peptidase type I. (e) – Tat-dependant signal peptide of lipoprotein, cleaved by signal peptidase type II. Adapted from Harwood and Cranenburgh (2008)

Sec-dependent (Figure 1.9) newly synthesised proteins in the cytoplasm must remain in an unfolded state (secretion-competent state) otherwise they will not be exported through the translocase (the multi-protein complex in the bacterial membrane which directs translocation). An intracellular chaperone system prevents folding of the protein by creating a protein-chaperone complex and directs the protein towards the translocase. In *E. coli*, the SecB chaperone binds to the polypeptide, preventing its folding and transfers it to the membrane protein SecA which is a part of the translocase. The second chaperone system present in *E. coli* and Gram-positive bacteria is the SPR- (signal recognition particle) pathway which is composed of the Ffh protein and a small cytoplasmic RNA molecule which binds to the signal peptide of secretory proteins as they emerge from the ribosome. The SRP-bound ribosome nascent chain complex (RNC) is positioned next to the membrane-bound receptor FtsY, followed by the transfer of protein complex to the translocase (Harwood and Cranenburgh 2008). The translocase is a membrane pore through which the nascent protein is exported. The translocase proteins SecY, SecE and SecG form a heterotrimeric core and interact with the SecA protein which drives the translocation by hydrolysis of ATP in *E. coli* and *B. subtilis*. Homologs of the Sec translocase components are encoded by the *L. lactis* MG1363 genome. A second heterotrimeric translocase complex identified in *E. coli* as SecDF-YajC (modulates the catalytic cycle of SecA) and in *B. subtilis* as SecDF-YrbF (contributes to efficient secretion) is not encoded in the *L. lactis* genomes (van Wely *et al.*, 2001). The chaperone-bound pre-protein arrives at SecA and is transported through the complex SecYEG. ATP is hydrolysed by SecA and the pre-protein is released from the chaperone complex. The signal peptide is cleaved off by signal peptidase I and the mature protein is pushed through the membrane to the periplasm (*E. coli*) or extracellular environment (Gram-positive bacteria) (Natale *et al.*, 2008). The lipoproteins are firstly lipid-modified and subsequently the signal peptide is cleaved off by signal peptidase II.

The Tat-dependant pathway (Figure 1.9) has been recognised in *E. coli* and *B. subtilis*, however homologous genes are not encoded in the genome of *L. lactis* MG1363 (Wegmann *et al.*, 2007, Bolotin *et al.*, 2001). The Tat-pathway in contrast to the Sec-system is capable of secretion of folded proteins across the inner membrane to the periplasm (*E. coli*) or extracellular space (*B. subtilis*). The Tat translocase is a multi-protein complex and consists of TatA, TatB, TatC, TatD and

TatE proteins. Recently it was shown that a putative lipoprotein DmsA from *Shewenella oneidensis* was translocated by Tat-dependant system. In the Gram-positive bacterium *B. subtilis* very few of proteins are exported via Tat and there are no reports on Tat-dependent lipoproteins from other low %G+C Firmicutes (Hutchings *et al.,* 2009).

In Gram-positive bacteria, the protein can be anchored to the cytoplasmic membrane, cell wall matrix or be part of a surface layer (S-layer) if present (Harwood and Cranenburgh 2008). Lipoproteins are synthesized as pre-lipoproteins and have to be modified by the diacylglyceryl transferase (Lgt) before the lipoprotein signal peptide is cleaved off by signal peptidase II (lspA). The diacylglyceryl group, attached to the cysteine residue of the mature lipoprotein, inserts into the lipid bilayer of the cytoplasmic membrane, preventing release of the protein into the environment (Natale *et al.,* 2008, Hutchings *et al.,* 2009). In Gram-negative bacteria, lipoproteins are either retained in the cytoplasmic membrane after lipid modification or they are transported to the outer membrane by the Lol (lipoprotein localisation) pathway (Hutchings *et al.,* 2009). A special group of proteins remain covalently anchored to the cell wall via the C-terminus of the protein. These proteins, apart from the N-terminal signal peptide, carry a C-terminal cell wall anchoring motif LPXTG, followed by a hydrophobic region and positively charged tail (Figure 1.9). A specific transpeptidase, the sortase A (SrtA), recognises the cell wall sorting signal and cuts it between threonine and glycine residues. The mature protein is then covalently attached to the cell wall by the carboxyl group of the threonine residue (Natale *et al.,* 2008).

A limiting factor in heterologous protein secretion is degradation by housekeeping proteases. *E. coli* produces several proteases located in the cytoplasm and in the cell envelope (Baneyx and Mujacic 2004). *B. subtilis* encodes three main serine proteases WprA, HtrA and HtrB, which degrade misfolded and non-native secretory proteins. It was reported that a *wprA*-null mutant of *B. subtilis* enhances the production of the heterologous protein (Stephenson and Harwood 1998). However this 'quality control' system remains an issue in heterologous expression studies using *B. subtilis* as a host. In contrast, *L. lactis* MG1363 was reported to encode only one housekeeping protease HtrA which is involved in proteolysis of aberrant proteins,

maturation of native proteins and processing of pro-peptides. Unlike the *B. subtilis wprA*- mutant, which improved the production of heterologous proteins, the inactivation of *L. lactis htrA* had the opposite effect (Sriraman and Jayaraman 2008). . It was instead reported that the HtrA protease is essential for efficient secretion of recombinant proteins in *L. lactis*. It was proposed that bacterial cell aggregation which was observed in HtrA-depleted cells could potentially reduce the secretion efficiency (Sriraman and Jayaraman 2008).

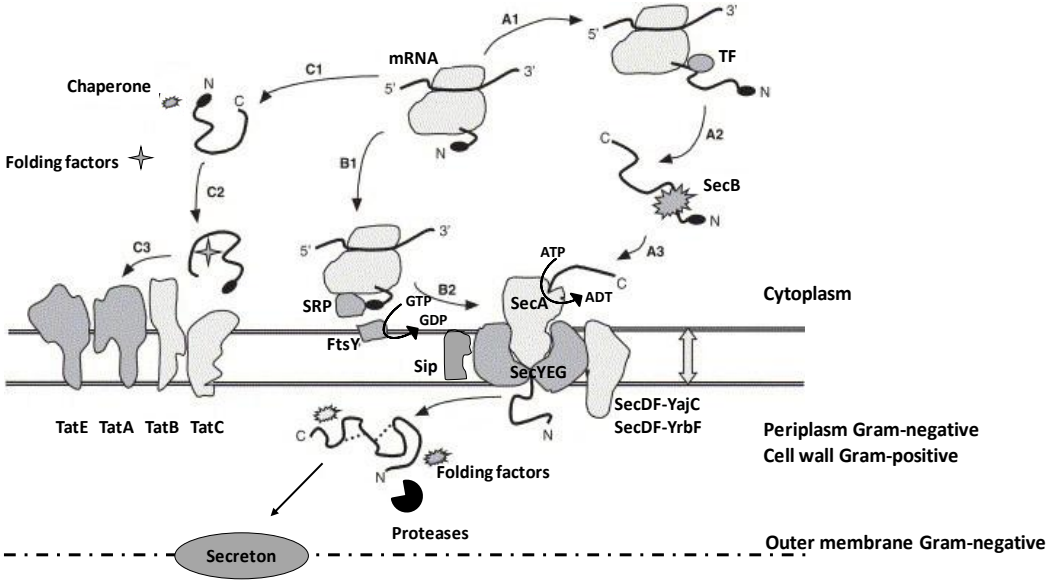


Figure 1.9 Schematic overview of bacterial protein translocation systems.

E. coli: A1 - the protein emerges from the ribosome and binds to trigger factor (TF), step A2 – protein is recognised by SecB, step A3 – targeting of the complex to SecA. *E. coli, B. subtilis, L. lactis* step B1 - the protein emerges from the ribosome and is recognised by the SRP complex, step B2 – interaction with FtsY and release to the translocation site. The translocase is a membrane protein complex consisting of SecA, SecYEG (*E. coli, B. subtilis, L. lactis*) and SecDF-YajC (*E. coli*) or SecDF-YrbF (*B. subtilis*). The signal peptide is cleaved by the Sip protein. In the periplasm (*E. coli*) or outside the cell wall the protein is folded (Gram-positive bacteria). The Tat-dependent system transports folded proteins (*E. coli* and *B. subtilis*). The lipoproteins are secreted by either Sec- or Tat-dependent system. Adapted from Mergulhão *et al.*, 2005.

In *E. coli* the presence of the outer membrane prevents direct secretion into the culture medium. Most of the heterologous proteins remain in the periplasm where they are folded by several folding modulators (Baneyx and Mujacic 2004). The signal peptides (SP) originating from Gram-positive bacteria such as *Ruminococcus albus* or *Bacillus subtilis* were reported to be recognised by the translocation machinery of *E. coli*. Two signal peptides from two ruminococcal cellulases were used to track the translation pathway by fusion to the GFP protein. It was observed that SP from cellulase Cel48A was a substrate for the Sec-pathway, in contrast to SP from Cel9B which served as substrate for Tat-pathway. Both proteins were secreted to the periplasmic space of *E. coli* (Esbelin *et al.,* 2009). The secretion of heterologous proteins in *E. coli* is often mediated by replacing the native signal peptide with the SP of *E. coli* OmpA protein. This system has been used to produce a number of heterologous proteins from Gram-positive bacteria such as α-amylase from *B. subtilis* or endomannosidase from *B. licheniformis*. The secreted protein was detected to the periplasm and culture medium of *E. coli* used as a heterologous host (Yamabhai *et al.,* 2008, Songsiriritthigul *et al.,* 2010). In Gram-negative bacteria, the proteins can be exported through the outer membrane by non-specific periplasmic leakage, or via the main terminal branch of the Sec-dependant pathway. The extracellular secretion of endoglucanase from *B. subtilis* was reported in *E. coli* by Lo *et al.* (1988). In Gram-positive bacteria such as *L. lactis* and *B. subtilis*, the folding of the native protein must occur rapidly when they reach the external environment and it is meditated by several folding factors (Harwood and Cranenburgh 2008).

1.10 Aims of this work

The overall aim of this project was to better understand the metabolic potential of the human gut microbiota in dietary polysaccharides breakdown.

1. At the technical level, a culture-independent approach was chosen as the method of study. *Lactococcus lactis* MG1363 would be use as an alternative expression host for the functional screening of the created libraries in parallel to *Escherichia coli* to investigate carbohydrate degrading enzymes from human gut microbiota. The first objective was to develop a shuttle vector based on plasmid pLP712 from

Lactococcus lactis NCDO712 and test its cloning abilities by constructing a genomic library from the novel human gut isolate *Ruminococcus* sp. 80/3 using *L. lactis* MG1363 and *E. coli* as heterologous hosts. Subsequently, a metagenomic library from the human gut microbiota would be created, in both hosts. The functional screening of the libraries would be performed to detect novel microbial activities. The two hosts would be compared for their suitability to generate metagenomic libraries. It was hypothesized that the expression of genes derived from low %G+C gut Firmicutes would be enhanced in *L. lactis* leading to recovery of a higher number of positive clones. The metagenomic libraries generated here would not be expression libraries with cloned genes expressed from native promoters.

2. At the biological level, the second objective was to analyse novel glycoside hydrolase encoding genes derived from the human gut bacteria that were identified during objective one, and to determine their substrate specificity, structure and origin using bioinformatics and functional approaches. This information would provide better understanding on the carbohydrate-degrading capacity of the human gut microbiota. This project also aimed to identify potentially novel genes and their products involved in carbohydrate metabolism in the human GI tract that are relevant to microbiota dynamics.

This study was developed through the collaboration between the Institute of Food Research (Norwich, UK) and the Rowett Institute of Nutrition and Health (University of Aberdeen, Aberdeen, UK)

Chapter 2

Materials and Methods

2.1 Buffers and solutions

All the components were dissolved in distilled water and sterilised by autoclaving at 121°C for 20 minutes and stored at room temperature, unless stated otherwise.

2.1.1 Solutions for bacterial growth media

10% Glucose

Distilled water	to 100 ml
Glucose	10 g

1 M $MgCl_2$/ 1 M $MgSO_4$ (sterilised by filtration through 0.22 µm membrane)

Distilled water	to 100 ml
$MgCl_2$	9.52 g
$MgSO_4$	12.04 g

Mineral solution 1

Distilled water	to 1 litre
K_2HPO_4	3 g

Mineral solution 2

Distilled water	to 1 litre
KH_2PO_4	3 g
$(NH_4)_2SO_4$	6 g
NaCl	6 g
$MgSO_4$	0.6 g
$CaCl_2$	0.6 g

Vitamin I solution (stored at -20°C)

Distilled water	to 100 ml
Biotin	1 mg
Cobalamin	1 mg
ρ-aminobenzoic acid	3 mg
Folic acid	5 mg
Pyridoxamine	15 mg

Vitamin II solution (sterilised by filtration through 0.22 µm membrane, stored at -20°C)

Distilled water	to 100 ml
Thiamine	5 mg
Riboflavin	5 mg

Haemin solution

Distilled water	to 100 ml
KOH	0.28 g
95% Ethanol	25 ml
Haemin	100mg

Does not require sterilisation

VFA mix

Acetic acid	17 ml
Propionic acid	6 ml
n-valeric acid	1 ml
Isovaleric acid	1 ml
Isobutyric acid	1 ml

Does not require sterilisation

Antibiotic	**Stock solution**	**Final concentration**		**Solvent**
		E. coli	***L. lactis***	
Erythromycin	30 $mg.ml^{-1}$	150 $\mu g.ml^{-1}$	5 $\mu g.ml^{-1}$	Ethanol
Chloramphenicol	10 $mg.ml^{-1}$	-	5 $\mu g.ml^{-1}$	Ethanol
Ampicillin[a]	100 $mg.ml^{-1}$	100 $\mu g.ml^{-1}$	-	H_2O

Table 2.1 Antibiotic solutions used in this study.

[a]- sterilised by filtration through 0.22 µm membrane. Antibiotic solutions were stored at -20°C.

2.1.2 Buffers for DNA manipulation techniques

0.5 M Tris-HCl, pH 8.0

Distilled water	to 100 ml
Tris base	12.1 g

Adjust the pH to 8.0 with concentrated HCl

0.5 M EDTA, pH 8.0

Distilled water	to 100 ml
EDTA	14.6 g

Adjust the pH to 8.0 with 5 M NaOH

STE buffer (Sambrook and Russell 2001)

Distilled water	to 1 litre
NaCl	5.8 g
0.5 M Tris-HCl pH 8.0	20 ml
0.5 M EDTA pH 8.0	2 ml

TE pH 8.0 (Sambrook and Russell 2001)

Distilled water	to 100 ml
0.5 M Tris-HCl pH 8.0	2 ml
0.5 M EDTA pH 8.0	0.2 ml

THMS buffer

Distilled water	to 100 ml
0.5 M Tris HCl pH 8.0	10 ml
1 M $MgCl_2$	0.2 ml
Sucrose	6.7 g

Buffer TL

Distilled water	to 15 ml
0.5 M Tris HCl pH 8.0	0.75 ml
Lysozyme from chicken egg (Sigma, UK)	150 mg

Does not require sterilisation, freshly prepared before use.

Buffer A

Distilled water	to 100 ml
0.5 M Tris HCl pH 8.0	10 ml
0.5 M EDTA pH 8.0	50 ml

Buffer B

Distilled water	to 100 ml
SDS	20 g
0.5 M Tris HCl pH 8.0	10 ml
0.5 M EDTA pH 8.0	4 ml

2.5 M Potassium acetate pH 5.2

Distilled water	to 100 ml
Potassium acetate	24.54 g

Adjust the pH to 5.2 with acetic acid

10x TBE

Distilled water	to 1 litre
Tris Base	108 g
Boric acid	55 g
0.5 M EDTA pH 8.0	40 ml

Does not require sterilisation

6x loading buffer (Sambrook and Russell 2001)

Distilled water	to 100 ml
Bromophenol blue	0.25 g
Xylene cyanol FF	0.25 g
Ficoll	15 g

Does not require sterilisation.

2.1.3 Solutions for transformation techniques

Electroporation buffer 1 (EP1)

Distilled water	to 100 ml
Sucrose	17.11 g
Glycerol	10 ml

Electroporation buffer 2 (EP2)

Distilled water	to 100 ml
Sucrose	17.11 g
Glycerol	10 ml
0.5 M EDTA pH 8.0	10 ml (after autoclaving)

1 M $MgCl_2$

Distilled water	to 100 ml
$MgCl_2.6H_2O$	20.33 g

1 M $CaCl_2$

Distilled water	to 100 ml
$CaCl_2.2H_2O$	14.7 g

2.1.4 Solutions for genomics techniques

1 M $MgSO_4$

Distilled water	to 100 ml
$MgSO_4.7H_2O$	24.65 g

20x freezing mix

Distilled water	to 100 ml
K_2HPO_4	12.54 g
KH_2PO_4	3.59 g
Na Citrate	1.00 g
$(NH_4)_2SO_4$	1.80 g
1 M $MgSO_4$	0.8 ml (after autoclaving)

2.1.5 Solutions for enzyme assays

Congo red staining solution

Distilled water	to 1 litre
Congo Red	1 g
NaOH	0.2 g

Destaining solution

Distilled water	to 1 litre
NaCl	58.33 g
NaOH	0.2 g

Ruthenium red

Distilled water	to 100 ml
Ruthenium red	0.05 g

100 mM sodium phosphate buffer, pH 6.5

100 mM Na_2HPO_4 solution

Distilled water	to 0.5 litre
Na_2HPO_4	7.098 g

100 mM NaH_2PO_4 solution

Distilled water	to 0.5 litre
NaH_2PO_4	7.800 g

Mix two solutions to obtain pH 6.5. Store at 4°C.

Lowry reagents

Solution 1

Distilled water	to 500 ml
Na_2CO_3	0.25 g

Solution 2

Distilled water	to 100 ml
Sodium potassium tartrate	1 g
$CuSO_4.5H_2O$	0.5 g

Adjust pH to 7.0 with 1M NaOH

Mix 50 ml of solution 1 and 2 ml of solution 2 before use.

Lever reagents

Bismuth reagent – store at 4°C up to 1 week

Distilled water	to 10 ml
Bismuth nitrate	4.84 g
Sodium potassium tartrate	2.82 g
NaOH	1.2 g

PAHBAH reagent – freshly prepared before use

0.5 M NaOH	to 100 ml
p-hydroxy benzoic acid hydrazide	0.761 g
Bismuth reagent	1 ml

SDS-PAGE gel

Separating gel (8%)	10 ml (2 gels 0.75 mm)
Distilled water	9.352 ml
1.5 M Tris-HCl pH 8.8	5 ml
10% SDS	0.2 ml
Acrylamide/Bis 37.5:1 (30% solution)	5.333 ml
10% ammonium persulfate[a]	0.1 ml
TEMED	0.015 ml

Stacking gel (4%)	10 ml (2 gels 0.75 mm)
Distilled water	6.1 ml
0.5 M Tris-HCl pH 6.8	2.5 ml
10% SDS	0.1 ml
Acrylamide/Bis 37.5:1 (30% solution)	1.3 ml
10% ammonium persulfate	0.060 ml
TEMED	0.020 ml

Electrophoresis buffer (5x conc.)

Distilled water	to 1 litre
Tris base	15 g
Glycine	72 g
SDS	5 g

Coomassie stain

Distilled water	to 100 ml
Coomassie blue R250	0.1 g
Methanol	8 ml
Acetic acid	7 ml

Destaining solution

Distilled water	to 100 ml
Methanol	8 ml
Acetic acid	7 ml

Solution 1

Distilled water	to 500 ml
Tris base	0.6055 g
DTT	0.1542

Isopropanol	100 ml
Adjust pH to 7.5 with HCl	

Solution 2

Distilled water	to 500 ml
Tris base	3.0275 g
DTT	0.1542 g
0.5 M EDTA	1 ml
Adjust pH to 6.8 with HCl	

For zymogram gel replace 2 ml of water with 2% CMC.
[a]- Freshly prepared before use.

2.2 Bacterial strains and plasmids used in this work

Strain	Characteristics	Reference
Lactococcus lactis* subsp. *cremoris		
MG1363	Plasmid-free strain, Lac^-	(Gasson 1983)
MG1629	Derivative of NCDO712, Lac^+, Prt^+	Mike Gasson, personal communication
FI10792	MG1363 containing pFI2672	This study
FI10793	MG1363 containing pFI2673	This study
FI10794	MG1363 containing pFI2674	This study
FI10795	MG1363 containing pFI2675	This study
FI10796	MG1363 containing pFI2676	This study
FI10858	MG1363 containing pFI2709	This study
FI10859	MG1363 containing pFI2710	This study
-	MG1363 containing pTRKL2	This study
Escherichia coli		
pir 116	F^- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80d*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ^- *rps*L *nup*G *pir*-116(DHFR)	Epicentre
pir	F^- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80d*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ^- *rps*L *nup*G *pir* (DHFR)	Epicentre
XL1 Blue	*endA1 supE44 thi-1 hsdR17 recA1 gyrA96 relA1 lac* [F´ *proAB lacI*q*ZΔM15* Tn*10* (Tet^r)	Stratagene
FI10862	*pir* 116 containing pFI2676	This study
FI10863	*pir* 116 containing pFI2710	This study
-	XL1 Blue containing pTRKL2	This study
Ruminococcus sp. 80/3 (EU266551)		(Dabek *et al.*, 2008)
Coprococcus eutactus ART55/1 (AY350746)		(Louis *et al.*, 2004)
Coprococcus sp. L2-50 (AJ270491)		(Duncan *et al.*, 2002)

Plasmid	Characteritics	Reference
pLP712	55-kb, Lac^+, Prt^+	(Gasson 1983)
pFI2672	10-kb *Ban*II fragment of pLP712 ligated to *Xho*I/*Eci*I ery^R gene of pG^+ host9	This study
pFI2673	8.8-kb *Msp*I fragment of pFI2672 ligated to PCR amplified cm^r gene of pUK200	This study
pFI2674	pFI2673 cut with *Psh*AI/*Drd*I and recircularised	This study
pFI2675	8.5-kb *Bam*HI fragment of pFI2674 ligated to *Bam*HI/*Bgl*II MCS of pMTL23p	This study
pFI2676	8.5-kb *Pci*I fragment of pFI2675 ligated to PCR amplified R6Kγori from pMOD™-3	This study
pFI2709	8.8 kb pFI2674 cut with *Bst*UI and ligated to *Sma*I linker	This study
pFI2710	8.8 kb *Pci*I fragment of pFI2709 ligated to PCR amplified R6Kγori from pMOD™-3	This study
pG^+ host9	Ery^R, thermosensitive replicative plasmid in *L. lactis*, 3.7 kb	(Maguin *et al.,* 1996)
pUK200	Cm^R, P_{nisA}, pSH71 replicon with terminator of *brnQ*, 3.2 kb	(Wegmann *et al.,* 1999)
pMTL23p	Amp^R, cloning vector, 2.8 kb	(Chambers *et al.,* 1988)
pMOD™-3	Amp^R, R6Kγ*ori* located within the ME seqeunces, 2.8 kb	Epicentre
pTRKL2	Ery^R, *lac*Z, 6.4 kb	(O'Sullivan and Klaenhammer 1993)
pGEM®-T	Amp^R, *lac*Z, T7 RNA polymerase promoter, 3 kb	Promega
pIVEX 2.3d	C-terminal 6xHis-Tag, Amp^R, RBS, T7Prom and T7Term, MCS,	Roche
pIVEX 2.4 d	Cleavable N-terminal 6xHis-Tag, Amp^R, RBS, T7Prom and T7Term, MCS, Xa	Roche

2.3 Bacterial growth media and growth conditions

2.3.1 Aerobic growth media

Luria-Bertani (LB) medium (Sambrook and Russell 2001)

Distilled water	to 1 litre
Bacto tryptone	10 g
Bacto yeast extract	5 g
NaCl	10 g
For LB agar add	
Agar	15 g

Brain Heart Infusion (BHI) medium

Distilled water	to 1 litre
BHI dehydrated broth	37 g
For BHI agar add	
Agar	15 g

SOC medium

Distilled water	to 100 ml
Bacto tryptone	2 g
Bacto yeast extract	0.5 g
NaCl	0.05 g
1M $MgCl_2$/1M $MgSO_4$ solution	1 ml (after autoclaving, before use)
10 % Glucose	2 ml (after autoclaving, before use)

M17 medium

Distilled water	to 0.95 litre
M17 dehydrated broth	37.25 g
For GM17 add	
10 % Glucose	50 ml (after autoclaving)
For M17 agar add	
Agar	15 g

Modified M17 medium (mGM17)

Distilled water	to 95 ml
M17 dehydrated broth	3.725 g
Glycine	2.5 g
Sucrose	17.1 g
10 % Glucose	5 ml (after autoclaving)

The media components were dissolved in distilled water and sterilised by autoclaving at 121°C for 20 minutes. The dehydrated media were obtained from Oxoid (UK). Media were supplemented with appropriate antibiotics as required, according to table 2.1.

2.3.2 Anaerobic growth media

M2GSC medium

Distilled water	to 100 ml
Bacto casitone	1 g
Yeast extract	0.25 g
$NaHCO_3$	0.4 g
Glucose	0.2 g
Starch	0.2 g
Cellobiose	0.2 g
Clarified rumen fluid	30 ml
Mineral solution 1	15 ml
Mineral solution 2	15 ml
0.1 % Resazurin	0.1 ml
Cysteine (after boiling)	0.1 g

YCFA (C)

Distilled water	to 100 ml
Bacto casitone	1 g
Yeast extract	0.25 g
$NaHCO_3$	0.4 g
Cellobiose	0.2 g
Mineral solution 1	15 ml
Mineral solution 2	15 ml
Vitamin I solution	100 µl
Vitamin II solution	100 µl (after autoclaving)
Haemin solution	1 ml
VFA mix	0.31 ml
0.1 % Resazurin	0.1 ml
Cysteine (after boiling)	0.1 g

For M2GSC and YCFA(C) media, the various components were dissolved in distilled water and then boiled for 2 minutes. Cysteine was added, and the medium was re-boiled and allowed to cool down under O_2-free CO_2 (100%) atmosphere using metal gassing hooks. The medium was dispensed into Belco tubes flushed with CO_2

in order to maintain anaerobic conditions. The tubes were tightly closed with a rubber stopper and a plastic cap. All media were autoclaved at 121°C for 20 minutes.

2.3.3 Growth conditions

Escherichia coli strains were grown at 37°C in LB or BHI medium with shaking at 220 rpm, unless stated otherwise. *Lactococcus lactis* strains were grown at 30°C in M17 medium supplemented with 0.5% (w/v) glucose, under static conditions. Anaerobic strains were grown at 37ºC in YCFA(C) or M2GSC medium, under static conditions.

2.3.4 Storage of bacterial strains

All strains were stored in -20°C and -80°C freezers. One volume of the bacterial overnight culture was added to one volume of 40% (v/v) sterile glycerol. For anaerobic bacteria, glycerol was prepared under O_2-free CO_2 (100%) atmosphere. Bacterial cultures growing on agar plates were stored at 4°C up to one month.

2.4 DNA manipulation techniques

2.4.1 Plasmid DNA purification – small scale preparation

Plasmid DNA from overnight *E. coli* and *L. lactis* cultures was isolated using a QIAprep Spin Mini Kit (Qiagen, UK) according to the manufacturer's instructions or with the following modification for *L. lactis*. The cell pellet was re-suspended in 250 µl of P1 buffer supplemented with lysozyme at the concentration 10 $mg.ml^{-1}$. For efficient cell lysis, tubes were incubated at 37°C for 30 minutes.

2.4.2 Plasmid DNA purification – maxi scale preparation

Plasmid DNA from 500 ml of overnight *E. coli* culture was harvested by centrifugation at 5000 x g for 15 minutes at 4°C in a Sorvall STE-28 rotor. The bacterial pellet was washed twice with 100 ml of ice-cold STE buffer, followed by harvesting and alkaline lysis according to Sambrook and Russell (2001). Plasmid DNA was purified by equilibrium centrifugation in a CsCl - ethidium bromide gradient in Beckman Quick Seal tubes (13.5 ml) at 99,300 x g for 60 hours (rotor Ti 90.1) at 20°C. The plasmid DNA band was collected by puncturing the Quick Seal tubes and collecting the fraction with a 21G x 1" UTW hypodermic needle attached to a 10 ml syringe (Terumo, Japan). Ethidium bromide was removed from DNA by

extraction with an equal volume of n-butanol saturated with water until the pink colour disappeared from both the aqueous and organic phase. DNA was precipitated with isopropanol according to section 2.4.13.

2.4.3 Plasmid stability test

Segregational stability of created constructs in *L. lactis* MG1363 was tested in the following way. Overnight cultures of each relevant plasmid-carrying *L. lactis* were grown in GM17 broth with the appropriate antibiotic. These cultures were subjected to six serial ten-fold dilutions and 20 µl of each dilution was spotted on GM17 agar plates with and without antibiotic. The plates were incubated at 30°C overnight. Then 0.5 ml of the undiluted overnight cultures was used to inoculate 4.5 ml of the GM17 broth without antibiotic. Six further serial ten-fold dilutions were made. The cells were grown for 280 minutes, which is equal to approximately 8 generations (based on previously assessed growth curve) at 30°C. The penultimate sample that grew (as judged by eye) was then used to make the serial dilutions for spotting and sub-culturing fresh GM17 broth for the next subculture. Sub-culturing was carried out consecutively for up to approximately 100 generations. The percentage of cells that had retained the plasmid was calculated by dividing the average number of colonies grown on selective agar plates of GM17 by the average number of colonies grown on non-selective agar plates.

2.4.4 Isolation of genomic DNA

Isolation of genomic DNA from *Ruminococcus* sp. 80/3 for library construction was performed as follows. An overnight culture (500 ml) was harvested by centrifugation at 5,000 x g for 15 minutes at 4°C in a Sorvall STE-28 rotor. The bacterial pellet was washed twice with 100 ml STE buffer and centrifuged as before. The bacterial cells were resuspended in 57 ml of THMS buffer, with RNase (final concentration 100 $\mu g.ml^{-1}$) and 13.5 ml of TL buffer added. The bottle was incubated at 37ºC for 0.5 hour. When the lysis was complete, 7.5 ml of buffer A and 4.2 ml of buffer B were added and supplemented with proteinase K (30 $U.mg^{-1}$). The mix was incubated at 55°C for 0.5 hour. Genomic DNA was purified twice by equilibrium centrifugation in a CsCl - ethidium bromide gradient in Beckman Quick Seal tubes (39 ml) at 151,000 x g for 40 hours (rotor Ti 70.1) at 20°C. The genomic DNA band was collected by puncturing the Quick Seal tubes and collecting the fraction with a

21G x 1”UTW hypodermic needle attached to 10 ml syringe (Terumo, Japan). Ethidium bromide was removed from DNA by extraction with an equal volume of isopropanol saturated with TE. An equal volume of TE was added followed by phenol:chloroform:isoamyl alcohol (25:24:1) and chloroform: isoamyl alcohol (24:1) extraction. The DNA was precipitated with ethanol according to method 2.4.13.

2.4.5 Metagenomic DNA isolation

A faecal sample was collected from a healthy volunteer by using the commode specimen collection system (Fisher Scientific, USA) and was processed immediately. The sample was placed in double plastic bags and all the air was removed before sealing with sellotape. The sample was placed between the paddles of a stomacher (Lab Blender 80, Seward Medical, UK) and it was blended three times for 1 minute. The thoroughly mixed sample was divided into 400 mg aliquots. The aliquots were snap frozen in liquid nitrogen and stored at -80°C until freeze drying in a freeze dry system (Labconco, USA). DNA from lyophilised samples was extracted using the FastDNA® Spin kit for soil (MP Biomedicals, UK) using the mechanical cell disruption with provided glass beads and manufacturer’s protocol with the modifications as follows. The samples were processed in the FastPrep® instrument for 15 sec at speed 6.5, unless stated otherwise. The DNA was eluted with 120 µl of DNase/Pyrogen Free water.

2.4.6 Mechanical shearing of genomic DNA using HydroShear DNA device (GeneMachines®)

Genomic DNA from a pure culture of *Ruminococcus* sp. 80/3 was extracted according to section 2.4.4. High molecular weight DNA was mechanically sheared with the HydroShear DNA device (Gene Machine, USA) in order to produce 5-10 kb size inserts. The syringe and tubes were flushed with 0.2 M HCl (3 cycles), 0.2 M NaOH (3 cycles) and TE pH 8.0 (3 cycles) before and after use. Genomic DNA (10 µg) in 100 µl volume was sheared at speed 8 for 20 cycles. The sheared DNA was analysed by agarose gel electrophoresis (see section 2.4.7).

2.4.7 Agarose gel electrophoresis

DNA samples were analysed using a horizontal electrophoresis apparatus (Bio-Rad Laboratories, UK). Agarose solution (1% (w/v)) was prepared in 1x TBE buffer. The

slurry was heated in a microwave oven until the agarose dissolved. Samples were mixed with 0.2 volumes of 6x loading buffer and were loaded on the gel alongside 5 µl of a 10 kb marker (Hyperladder I, Bioline, UK) or λ*Hind*III DNA marker (Promega, UK). Gels were run in 1x TBE buffer at 120 V for 40-80 minutes. The DNA was stained by immersing the agarose gel in an ethidium bromide solution (0.5 $\mu g.ml^{-1}$) or in 3x concentrated Gel Red (Biotium, USA) stain in H_2O with 0.1 M NaCl. In each case, the gel was agitated gently at room temperature for ~ 30 minutes.

2.4.8 DNA recovery from agarose gels

DNA was recovered from agarose gels using electroelution into dialysis tubing in 1x TBE buffer. Dialysis tubing (Medicell International, UK, size 2, 18/32", MWCO – 12 – 14 kDa) was prepared by boiling in a solution of 2% sodium bicarbonate/ 1 mM EDTA for 10 minutes, washing in distilled H_2O and re-boiling in a solution of 1 mM EDTA for 10 minutes. The tubing was allowed to cool down and was stored at 4°C in a 2% ethanol solution. The tubing was rinsed with 1x TBE prior to use.

In order to define the band of interest, the part of the stained gel with the ladder was exposed to UV light on a UV transilluminator. The DNA band was cut from the gel according to the marker size using a sterile, sharp blade and the piece of gel was transferred into dialysis tubing. The tubing was filled with 1x TBE, securely closed with clips and placed in the electrophoresis tank. The electroelution was run at 120 V for 40 minutes and at the reverse current for 30 sec. The buffer with eluted DNA was collected and the gel slice was examined under UV light to check the efficiency of the electroelution.

2.4.9 Restriction endonuclease digestion of plasmid DNA

Restriction enzymes were purchased from New England Biolabs (UK) or Promega (UK). The reaction was prepared in a volume of 10 µl for a routine digestion or 50 µl for digesting a cloning vector for genomic and metagenomic library construction. The amount of enzyme required per 1 µg of DNA was calculated according to the equation:

$$\frac{\text{size of } \lambda \text{ DNA (kb)}}{\text{size of plasmid (kb)}} \times \frac{\text{number of sites in plasmid}}{\text{number of sites in } \lambda \text{ DNA}} \times 10 = \text{enzyme U.}\mu g^{-1} \text{ DNA}$$

The cloning vector for genomic and metagenomic library construction was digested with *Sma*I enzyme. The reaction was incubated at 25ºC for 3 hours. The *Sma*I enzyme was inactivated by heating at 65°C for 20 minutes and the plasmid DNA was used for dephosphorylation with Antarctic phosphatase according the method 2.4.10. For routine digestion the reaction was incubated for 1 hour at the recommended optimal temperature. Complete digestion was verified by running an aliquot of digested vector next to undigested vector by agarose gel electrophoresis (see section 2.4.7). If required the restriction digested DNA sample was cleaned to remove enzyme/ buffer residues as described in method 2.4.13.

2.4.10 Dephosphorylation of 5'- phosphorylated ends of vector DNA

After restriction digestion, plasmid DNA was dephosphorylated by using Antarctic phosphatase (New England Biolabs, UK). Digested cloning vectors were incubated with 1 U of enzyme per 1 µg DNA at 37°C for 30 minutes, followed by heat inactivation at 65°C for 5 minutes and DNA clean up (method 2.4.13).

2.4.11 End repairing of fragmented DNA

The inserts DNA for genomic and metagenomic library construction were end repaired by End-It™ DNA End-Repair Kit (EPICENTRE Biotechnologies, USA) according to the manufacturer's instructions. The samples were incubated at room temperature for 45 minutes, followed by heat inactivation at 70°C for 10 minutes. DNA was purified prior to further manipulation according to method 2.4.13.

2.4.12 Ligation of DNA into cloning vector

T4 DNA ligase (Promega, UK, M1794) was used for the development of cloning plasmids and for genomic library construction. The LigaFast™ Rapid DNA Ligation System (Promega, UK) was used for metagenomic library construction. *Sma*I digested plasmid and insert were ligated in a molar ratio of 1:5. The ligation mix was incubated overnight at room temperature when T4 DNA ligase was applied or for 2 hours at room temperature followed by overnight incubation at 4°C when the LigaFast system was used. The ligation mix was purified prior to electrotransformation by ethanol precipitation (method 2.4.13) or by MF membrane filter disc dialysis (0.025 µm VSWP, Millipore, USA). The ligation mix was applied

on a filter disc and placed gently on the surface of sterile water and left for 30 minutes for efficient purification.

2.4.13 DNA clean up

Several clean up methods were applied during this work:

1. Wizard SV Gel and PCR Clean-Up system (Promega, UK) was used according to the manufacturer's instructions following electroelution and enzyme manipulation (dephosphorylation, end repairing).
2. Ethanol precipitation was applied following genomic DNA extraction (section 2.4.4) and ligation (section 2.4.12). Cold ethanol (100%, 2.5 x sample volume) and 2.5 M potassium acetate pH 5.2 (0.1 x sample volume) was added to the DNA and incubated on ice for 0.5 hour. DNA was recovered by centrifugation at 16000 x g for 10 minutes at 4°C, followed by washing with 70% ethanol. The DNA pellet was dried and resuspended in sterile water.
3. Isopropanol precipitation was applied following plasmid purification (method 2.4.2) and electroelution (method 2.4.8, when volumes of collected buffer were bigger than 5 ml). Isopropanol (0.6 x sample volume) was mixed with the DNA sample and stored at room temperature for 20 minutes. DNA was recovered by centrifugation at 16,000 x g for 10 minutes at 20°C, followed by washing with 70% ethanol and centrifugation as previously. The DNA pellet was dissolved in sterile water.
4. Sure Clean (Bioline, UK) system was used according to the manufacturer's instruction for the end repaired genomic insert DNA.

If required, the cleaned up DNA was concentrated with a DNA 30-11 vacuum concentrator (1,200 x g, 37°C).

2.5 DNA amplification

2.5.1 Polymerase Chain Reaction (PCR)

All PCR reactions were carried out using a PCR thermo cycler (Bio-Rad Laboratories, UK, iCycler). The PCR reactions were performed using Phusion High-Fidelity DNA Polymerase (Finnzymes, New England Biolabs UK) or BioTaq DNA Polymerase (Promega, UK). Primers P194/ P195 and R6κγori F/R were used during

plasmid construction (Table 2.3). For routine colony screening of a genomic library of *Ruminococcus* sp. 80/3, primers C5_F and C5_R were used. For colony screening of human gut metagenomic library primers M13_F and M13_R were used (Table 2.3). The amplification of 16S rRNA genes from metagenomic DNA extracted from a human faecal sample used a mixture of forward primers 7F, 27F Chlo, 27F Bor, 27F Bif, 27F Ato (4:1:1:1:1) and reverse primer 1510R. Primers used for the amplification of genes encoding glycoside hydrolase enzymes from *Coprococcus* strains are shown in Appendix 3. Primers used for *in vitro* over-expression study with pIVEX plasmid (Table 2.3) were used to amplify genes from the genomic library of *Ruminococcus* sp. 80/3, followed by cloning into pIVEX plasmid (Roche) according to manufactures instructions.

Sterile H_2O	**up to 50 µl**	**Final concentration**	
5x buffer	10 µl	1x	
2 mM dNTP	5 µl	0.2 mM	
10 µM Primer forward	2.5 µl	0.5 µM	
10 µM primer reverse	2.5 µl	0.5 µM	
template	X µl	5-10 ng or 1 µl of bacterial suspension	
polymerase	Y µl	BioTaq – 0.025 $U.\mu l^{-1}$, Phusion – 0.02 $U.\mu l^{-1}$	
Step	**Temp**	**Time**	**Cycles**
Initial denaturation	95- 98ºC	2- 5 min	x1
Denaturation	95- 98ºC	10- 30 sec	
Annealing[a]	XºC	10- 30 sec	x25-30
Extension	72ºC	30- 60 sec/ 1kb	
Final extension	72ºC	5 min	x1

Table 2.2 The components and conditions of typical PCR reaction

[a] – annealing temperature for primers used in this work is provided in Table 2.3.

Label	Sequence (5'-3')	[°C]	Reference	Target DNA
Gene amplification primers				
P194_F	GCACCCATTAGTTCAACAAACG	52	This study	pUK200
P195_R	ACTAACGGGGCAGGTTAGTGAC		This study	pUK200
R6κγori_F	CAAGCTTTAAAAGCCTTATATATT	54	This study	pMOD™-3
R6κγori_R	GTTGGCTAGTGCGTAGTCGTTGGC		This study	pMOD™-3
7F	AGAGTTTGATYMTGGCTCAG	52	(Satokari *et al.*, 2001)	16S rRNA
27F Chlo	AGAATTTGATCTTGGTTCAG		(Frank *et al.*, 2008)	16S rRNA
27F Bor	AGAGTTTGATCCTGGCTTAG		(Frank *et al.*, 2008)	16S rRNA
27F Bif	AGGGTTCGATTCTGGCTCAG		(Frank *et al.*, 2008)	16S rRNA
27F Ato	AGAGTTCGATCCTGGCTCAG		(Frank *et al.*, 2008)	16S rRNA
1510R	ACGGYTACCTTGTTACGACTT		(Satokari *et al.*, 2001)	16S rRNA
1_GL_F	CGCACGCCATGGACAAAAAGGAAATAAATGAAATAAAGAAAAC	55	This study	pFI2170_1GL
1_GL_R	CGCACGCCCGGGTTTTATTGTAGTTTCAAGTCCGTTG		This study	pFI2170_1GL
1_GL_RS	CGCACGCCCGGGTTATTTTATTGTAGTTTCAAGTCCGTTG		This study	pFI2170_1GL
4_GL_F	CGCACGTCATGAACATAGACAAGATTTTGAAGGAAC		This study	pFI2170_4GL
4_GL_R	CGCAGCAGTACTTTTTTCTGCAAGCATTTCGTTGAG		This study	pFI2170_4GL
4_GL_RS	CGCAGCAGTACTTTATTTTTCTGCAAGCATTTCGTTGAG		This study	pFI2170_4GL
Screening for insert in genomic library				
C5_F	GTACCGTTACTTATGAGCAAG	52	This study	pFI2710
C5_R	CTCTTTTCTCTTCCAATTGTC		This study	pFI2710
Screening for insert in metagenomic library				
M13_F	GTTTTCCCAGTCACGAC	50	Sigma, UK	pTRKL2
M13_R	CAGGAAACAGCTATGAC		Sigma, UK	pTRKL2

Table 2.3 Primers used for gene amplification in this study.

Restriction sites for pIVEX cloning are underlined. [°C] = annealing temperature. Y = C or T, M = A or C

2.6 Preparation and transformation of competent cells

2.6.1 Preparation of electrocompetent cells of *Lactococcus lactis*

L. lactis MG1363 electrocompetent cells were prepared according to the modified protocol by Gerber and Solioz (2007). Cells were grown in GM17 medium supplemented with 2.5% (w/v) glycine and 0.5 M sucrose (mGM17). An aliquot (100 µl culture from a glycerol stock) was inoculated into 5 ml of mGM17, and grown overnight at 30°C. Then 1 ml of this culture was inoculated into 10 ml of mGM17 and grown overnight under the same conditions. An aliquot (10 ml) from the overnight culture was then inoculated into 100 ml of mGM17, until the OD_{600} reached 0.2-0.3. Cells were harvested by centrifugation (4,000 x g/ 4°C/10 min, Sorvall, rotor SS-34) in 50 ml cold sterile Falcon tubes. Subsequent steps were performed on ice with ice-cold buffers. Cells were washed firstly with 50 ml EP1 buffer, followed by washing with 25 ml EP2 and 50 ml EP1 buffer. Cells were gently re-suspended in 1 ml EP1 buffer and 40 µl was aliquoted into 0.2 ml Eppendorf tubes which were snap-frozen in liquid nitrogen and stored at -80°C.

2.6.2 Transformation of *L. lactis* by high voltage electroporation

Transformation was performed using a BioRad Gene Pulser apparatus, set to 2.0 kV, 25 µF and 200 Ω. An aliquot of 40 µl frozen cells was thawed on ice. Purified ligation mix (10 ng of vector) was added to the cell suspension, mixed gently and transferred into pre-chilled 2 mm electroporation cuvettes (Cell Projects, UK). The pulse was applied and immediately 960 µl of modified pre-chilled GM17 medium (supplemented with 20 mM $MgCl_2$ and 2 mM $CaCl_2$) was added. Cuvettes were placed on ice for 5 minutes, and then the cell suspension was transferred into a 1.5 ml Eppendorf tube which was incubated at 30°C for 2 hours, followed by plating on selective GM17 agar plates.

2.6.3 Preparation of electrocompetent cells of *Escherichia coli*

E. coli electrocompetent cells were prepared according to Sambrook and Russell (2001). An overnight starter culture was subcultured (1/100) into 200 ml of LB medium (in a 2 litre flask) and grown at 37°C with vigorous shaking at 220 rpm until an OD_{600} of 0.4-0.5 was reached. Cells were harvested by centrifugation in 50 ml cold, sterile Falcon tubes (4,000 x g /4°C/10min, Sorvall, rotor SS-34). Subsequent

steps were performed on ice with ice-cold buffers. Cells were washed twice with 50 ml sterile water and once with 50 ml sterile 10% glycerol. They were gently re-suspended in 0.5 ml 10% glycerol, aliquoted to 0.2 ml Eppendorf tubes, snap frozen in liquid nitrogen and stored at -80°C. For genomic and metagenomic library construction commercially available electrocompetent cells were purchased as stated in section 2.2.

2.6.4 Transformation of *E. coli* by high voltage electroporation

Transformation was performed using a BioRad Gene Pulser apparatus, set to 1.7 kV, 25 µF and 200 Ω. An aliquot of 40 µl frozen cells was thawed on ice. Purified ligation mix (10 ng of vector) was added to the cell suspension, mixed gently and transferred into pre-chilled 1 mm electroporation cuvettes (Cell Projects, UK). The pulse was applied and immediately 960 µl of pre-warmed (37°C) SOC medium was added. The cells suspension was transferred into 15 ml Falcon tubes and was incubated at 37°C for 1 hour at 220 rpm, followed by plating on selective BHI agar plates.

2.7 High throughput genomics techniques

2.7.1 Storage of the genomic library of *Ruminococcus sp*. 80/3

Plates with transformants were flooded with sterile BHI medium; all colonies were removed and pooled together in 15 ml Falcon tube. The plasmid DNA from pooled clones of the genomic library of *Ruminococcus* sp. 80/3 was extracted using Qiagen Midi Kit (UK). The genomic library DNA was stored at -20°C.

2.7.2 Storage of the metagenomic library in *E. coli* and *L. lactis*

The colonies from transformation into *E. coli* were archived in triplicate in 384-well format plates (Nunc, UK) by using an automated colony picking robot (BioRobotics, Isogen Life Science, the Netherlands) or manual picking. *E. coli* clones were grown overnight at 37°C at 200 rpm in BHI medium supplemented with 1x freezing mix, 4.35% glycerol and erythromycin (150 $\mu g.ml^{-1}$). OD_{600} was measured at that point for future reference. One set of plates was used for *E. coli* clone pooling. Overnight grown cultures (20 µl) from each well were pooled into 50 ml Falcon tubes using a multichannel pipette. DNA from all clones was extracted using the Midi Prep Pure Yield kit according to the manufacture's instruction (Promega, UK). Metagenomic

library DNA was transformed into *L. lactis* MG1363 by electroporation following the method 2.6.2. The transformants were selected on GM17 plates with erythromycin after overnight incubation at 30°C and were archived in duplicated 384- well format plates by using manual picking. *L. lactis* MG1363 clones were picked into GM17 medium supplemented with 1x freezing mix, 4.37% glycerol and erythromycin (5 $\mu g.ml^{-1}$) and were grown at 30°C overnight. OD_{600} was measured before storing the plates at -80°C.

2.7.3 Functional screening of metagenomic libraries

The metagenomic library stored in 384 well plates was arrayed on agar plates with a substrate (table 2.4) using OmniTray plates (Nunc, UK) and the MicrogridII replicator from BioRobotics (Isogen Life Science, the Netherlands). The plates were incubated overnight at 37°C or at 30°C for *E. coli* or *L. lactis* library, respectively. These were used the following day for plate enzyme assay according to method 2.8.1.

2.7.4 DNA Sequencing

All DNA sequencing was carried out using ABI PRISM™ Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer, UK). M13 forward and reverse primers were used for initial sequence runs for the clones selected from the functional screening of the metagenomic inserts. Internal primers were subsequently designed and produced by Sigma, UK. Sequencing reactions were set up according to manufacturer's recommendations and were carried out by the Genomics Service at the Rowett Institute of Nutrition and Health.

2.7.5 Bioinformatics analysis of clones

All sequence data were assembled using the Seqman programme, Lasergene version 6 (DNASTAR 1989-2004). The open reading frames were detected by using the ORF search tool provided by NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Homology searches were run against the GenBank database using the BLASTX and BLASTP algorithms. The InterPro (http://www.ebi.ac.uk/Tools/pfa/iprscan/), SMART (http://smart.embl-heidelberg.de/smart/set_mode.cgi), Prosite (http://expasy.org/tools/scanprosite/) and PFAM (http://pfam.sanger.ac.uk/) databases were utilised for protein analysis (conserved domains and internal repeats

prediction). SignalP (http://www.cbs.dtu.dk/services/SignalP/) was used to predict the presence and location of signal peptide cleavage sites in amino acid sequences from different clones. Multiple amino acid sequences alignments were prepared with ClustalW in BioEdit. The alignment was edited in GeneDoc (http://www.psc.edu/biomed/genedoc) and the numbers of amino acids correspond to the whole protein. The most common residue in each column (the residue with the highest count) is assigned to the first shading level (black) and the second most common residue in each column is assigned to the next shading level (grey). The hypothetical promoter regions were predicted based on visual inspection of the sequence.

2.7.6 Phylogenetic analysis of sequences derived from 16S rRNA clone library

Genes of 16S rRNA from the metagenomic DNA extracted from a human faecal sample were amplified according to method 2.5.1, purified according to the manufacture's instruction using Wizard PCR clean up system (Promega, UK) and ligated to pGEM®-T vector according to method 2.4.12 (Promega, UK). The sequencing was done according to method 2.7.4. The 16S rRNA PCR product from library clones was amplified with M13 primers and sequenced using internal primer 797R (Nadkarni *et al.*, 2002) modified GGACTACHVGGGTATCTAATCC). The presence of chimeras was tested by Mallard version 1.02 (Ashelford *et al.*, 2006) and manually inspecting BLAST results. Sequences were aligned using ClustalW (Chenna *et al.*, 2003), a distance matrix was generated with Dnadist (Phylip package, distributed by J Felsenstein, University of Washington, Seattle) and Operational taxonomic units (OTUs) at 98% sequence identity were obtained with Dotur (Schloss and Handelsman 2005) on the in-house RINH/BIOSS Beowulf cluster. Single OTUs were subject to manual inspection and correction. The phylogenetic tree was constructed with Mega 5.03, using neighbour-joining method and distance method Kimura.

2.8 Enzyme related assays

2.8.1 Qualitative plate assay

Enzyme activity was examined on agar plates supplemented with substrate (table 2.4). For chromogenic substrates, positive clones turned blue. Plates containing non-chromogenic substrate plates were flooded with staining solution. After removing the

detection solution by aspiration, clear halos should be observed around positive clones or colonies.

Substrate (Supplier) final Concentration	Detection method
Starch from potato (Sigma, S2004) – 1%	Gram's iodine solution (Sigma, UK)
Carboxymethyl cellulose (Sigma, C4888) – 0.5%	Congo Red – 20 min Destaining solution – 15 min
Xylan from oat spelts (Sigma, X0627) – 0.5%	Destaining solution – 20 min Congo Red – 20 min Destaining solution – 15 min
Lichenan (Sigma, L6133) – 0.05%	Congo Red – 20 min Destaining solution – 15 min
Polygalacturonic acid (Sigma, P0853) – 0.5%	Ruthenium red solution – 30 min
5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside X-Gal (Sigma B4252) - 80 $\mu g.ml^{-1}$	Chromogenic
5-Bromo-4-chloro-3-indoyl-B-d glucopyranoside X-Glu (Fisher Scientific BPE4000-100) - 50 $\mu g.ml^{-1}$	Chromogenic
4-Methylumbelliferyl α-L-arabinofuranoside (Sigma M9519) – 50 $\mu g.ml^{-1}$	Fluorescent

Table 2.4 Substrates and detection methods used in this study during functional screening of genomic and metagenomic libraries

2.8.2 Preparation of enzyme for assay with p-nitrophenyl substrates

E. coli cultures were grown according to method 2.3.3. Overnight cultures (10 ml) were harvested by centrifugation at 5000 x g/ 4°C/ 10 min (Sorvall, SS-34). The pellet was washed twice with 3 ml of pre-chilled 100 mM sodium phosphate buffer (pH 6.5) and re-suspended in 1 ml of buffer. The re-suspended cells were transferred into a Sarstedt tube containing 250 µl of 106 µm glass beads in sterile water. The sample was bead- beaten twice for 30 sec with 30 sec incubation on ice using a Mini Bead Beater (Stratech, UK). The supernatant was collected after centrifugation at 16,000 x g/ 4°C/ 10 min; it was stored on ice and was used on the day of extraction to perform enzyme assay according to section 2.8.3.

2.8.3 Enzyme activity assay with p-nitrophenyl substrates

Pre-warmed to 37°C cell-free extract, was mixed in equal volume with 10 mM p-nitrophenyl substrate (table 2.5). The absorbance (A_{405}) was measured every 2 min

for 30 minutes at 37°C with shaking for 2 sec and settling for 5 sec using a Tecan Safire^2 plate reader (Tecan Trading Switzerland).

Substrate - SIGMA	MW	10 mM [mg.ml^{-1}]
p-nitrophenyl-β-D-glucopyranoside – N7006	301.3	3.013
p-nitrophenyl-β-D-glucuronide – N1627	315.2	3.152
p-nitrophenyl-β-D-xylopyranoside – N2132	271.2	2.712
p-nitrophenyl-α-D-galactopyranoside – N0877	301.3	3.013
p-nitrophenyl-β-D-galactopyranoside – N1252	301.3	3.013
p-nitrophenyl-α-D-glucopyranoside – N1377	301.3	3.013
p-nitrophenyl-α-D-arabinofuranoside – N3641	271.2	2.712

Table 2.5 Para nitrophenyl derivatives used in enzyme activity assays.

2.8.4 Preparation of enzyme fractions for Lever assay

E. coli and *L. lactis* cultures were grown according to method 2.3.3. Freshly grown overnight culture was harvested by centrifugation at 4000 x g at 4°C for 10 minutes (Sorvall, SS-34). The supernatant was kept on ice and concentrated by using Amicon Ultra-15 centrifugal filter devices (10,000 NMWL, Millipore, USA) to 0.04 of initial culture volume. The concentrated supernatant was aliquoted and stored at -80°C. The pellet was washed twice with ice-cold 50 mM sodium phosphate buffer (pH 6.5) and spun as before. The pellet was re-suspended in 50 mM sodium phosphate buffer (0.04 of initial culture volume, pH 6.5). The cells were broken by sonication using Soniprep 150 (MSE, UK) with three strokes of 30 sec and 2 minute intervals on ice. The sonicate was then centrifuged for 10 min at 16,000 x g at 4°C and the cell free extract was aliquoted and stored at -80°C.

2.8.5 Determination of enzyme activity with Lever assay

The supernatant and cell free extract was thawed on ice followed by 10-fold dilution in 50 mM sodium phosphate buffer, pH 6.5. The diluted enzyme fraction was incubated with 1% (w/v) of the appropriate polysaccharide substrate (in 50 mM sodium phosphate buffer, pH 6.5) at 37°C for different time intervals. The reaction was stopped by adding p-hydrobenzoic acid (PAHBAH) reagent followed by heating at 70°C for 10 min. The release of reducing sugars was determined in a microtitre plate Tecan Safire^2 plate reader at 415 nm. Dilutions of a stock solution of 1 mg.ml^{-1} of glucose or xylose were assayed to obtain a standard curve.

2.8.6 Gel electrophoresis of proteins

Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) was according to method in Sambrook and Russell (2001). Polyacrylamide stacking gel (4%) and 8% polyacrylamide separating gel were used. Gels were cast and run in Mini-PROTEAN®3 system (BioRad, UK). Protein samples were denatured by heating at 60°C for 20 minutes in 5x SDS loading buffer, before loading onto the gel. The gel was run in standard electrophoresis buffer at constant 80 V until the samples entered the stacking gel followed by separation at 200 V. The gel was stained afterwards for 2 hours in Coomassie solution at room temperature. The bands became clear by destaining in methanol (8% v/v)/acetic acid (7% v/v) solution.

2.8.7 Zymogram analysis

SDS-PAGE gel was prepared and run according to the method 2.8.5 with modifications. Carboxymethyl cellulose (CMC, at 0.2% w/v) was incorporated to the separating gel. After electrophoresis, the gel was washed twice in 200 ml of solution 1 for 50 minutes at room temperature. The gel was renatured in 200 ml of solution 2 overnight at 4°C with gentle shaking, followed by washing in 200 ml of 50 mM sodium phosphate buffer (pH 6.8) for 1 hour at 4°C. The gel was transferred onto a glass plate, covered and sealed with cling film and incubated at 37°C overnight. Bands were visualised by staining with Coomassie solution at 60°C for 2 hours and destaining with methanol/acetic acid solution overnight at room temperature. The gel was then neutralized by extensive washing with 0.1M Tris-HCl (pH 8.0) for 8 hours. CMC hydrolysis was detected according to method 2.8.1.

2.8.8 *In vitro* protein overexpression

Overexpression of the gene 1_GL and 4_GL from clone pFI2710_1GL and pFI2710_4GL, respectively, was performed in vitro using Rapid Translation System kit RTS100 (50 µl). The *E. coli* lysate, reaction mix, mix of amino acids and methionine provided with RTS100 kit were reconstituted and reaction solutions were prepared according to the manufacturer's instructions (Roche, UK). The plasmid with cloned DNA was extracted according to section 2.4.1. 0.5 µg of plasmid was added to the reaction solution in a total volume of 10 µl, mixed by gentle shaking and incubated at 30°C for 6 hours. The reaction was stored at 4°C. The overexpressed protein was analysed by SDS-PAGE gel electrophoresis according to section 2.8.6.

Chapter 3

Functional analysis of a genomic library from the human gut bacterium *Ruminococcus* sp. 80/3

3.1 Introduction

Function-based analysis of a metagenomic library relies on heterologous gene expression; however there are many obstacles that can limit successful gene expression. Conventionally *Escherichia coli* is used as a background host for expression studies. However, significant differences in expression pattern were reported between different taxonomic groups of bacteria (Gabor *et al.,* 2004). Several studies demonstrated that a broad host screening is likely to increase the number and diversity of positive clones from a functional metagenomic library (Martinez *et al.,* 2004, Craig *et al.,* 2010). A previous study showed that there should be some degree of evolutionary conservation between donor and a surrogate host in order to successfully express heterologous genes (Holt *et al.,* 2007). The similar %G+C content of recombinant gene and heterologous host was shown to facilitate gene expression (Warren *et al.,* 2008). Having this in mind, *Lactococcus lactis* MG1363 was proposed in this study to extend the alternative expression host range for studying glycoside hydrolase activity in the human gut microbiota by applying metagenomic approach. The development of a novel shuttle cloning vector is reported, followed by the construction and functional screening of a genomic library from the human gut isolate *Ruminococcus* sp. 80/3.

3.2 Development of a cloning vector based on the pLP712 replicon

Lactococcal plasmids have been studied widely due to their importance in the dairy industry and milk fermentation. Many industrially important traits are plasmid-encoded including lactose and protein breakdown, citrate permease activity, phage resistance and bacteriocin production (Mills *et al.,* 2006).

An extensive effort has been made in the development of shuttle vectors based on lactococcal replicons that can be applied in a broad host screening. There are a number of shuttle vectors such as pNZ123, pMIG, pSA3, pTRKH2 and pTRKL2, created for different bacterial genera of lactic acid bacteria (LAB) that can also be used in a *E. coli* background (O'Sullivan and Klaenhammer 1993, Dao and Ferretti 1985, De Vos 1987, Wells *et al.,* 1993).

The initial aim of this project was to create a plasmid based on a lactococcal replicon that would accommodate large inserts of genomic/ metagenomic DNA. Plasmid

pLP712 was chosen as the backbone to construct a novel cloning vector (Figure 3.1). Plasmid pLP712 is a 55 kb plasmid isolated from the dairy starter strain *L. lactis* NCDO712. This plasmid has been shown to encode genes for lactose and protein utilisation and is the only plasmid of this strain required for growth in milk and lactic acid production. The remaining four plasmids of *L. lactis* NCDO712 appeared to be cryptic (pSH71, pSH72, pSH73, pSH74). Previous study showed that pLP712 can accommodate large (up to 59 kb) genetic elements known as sex factor through the re-location from the bacterial chromosome into the plasmid (Gasson *et al.,* 1992). That makes pLP712 a suitable vector for cloning large size inserts. Plasmid pLP712 replicates via theta mode replication. The segregational stability of theta-mode replicating plasmids was shown to be superior to plasmids of the RCR-type replication (Kiewiet *et al.,* 1993). Plasmid pLP712 also encodes a partition system which is a common feature of large lactococcal plasmids like pGdh442, pSK08 and pCI2000 (Tanous *et al.,* 2007). The proteins encoded by the *parA* and *parB* genes are essential for the genetic stability of the plasmid during cell division (Schumacher 2007). Moreover, it has been extensively characterized by restriction endonuclease mapping and it has been sequenced recently (Wegmann *et al*., unpublished data). These characteristics of pLP712 make it useful for genetic manipulation and cloning application.

3.2.1 Construction of vectors carrying the pLP712 origin of replication

The 55 kb low-copy-number plasmid pLP712 was isolated from *Lactococcus lactis* MG1629 and used to construct a series of vectors (Figure 3.1). Firstly the pLP712 plasmid DNA was cut with *Ban*II to produce a 10 kb fragment, which carries the replication region. The fragment was blunt-ended and ligated to the 1.0-kb erythromycin resistance gene from plasmid pG^{+}host9 (Maguin *et al.,* 1996), resulting in pFI2672. Subsequently, pFI2672 was cut with *Msp*I, resulting in two fragments of 8.8 kb and 2.3 kb respectively. The 8.8 kb fragment was blunt ended and ligated to the chloramphenicol resistance gene amplified by PCR from plasmid pUK200 (Wegmann *et al.,* 1999). The resulting construct pFI2673 was further modified by deleting the nonessential 1.2 kb *Psh*AI/ *Drd*I region, resulting in the smaller derivative pFI2674 (8.5 kb). Next a *Bam*HI/ *Bgl*II fragment carrying the multiple cloning site of pMTL23p was cloned into pFI2674 linearized with *Bam*HI (construct

pFI2675). Finally, the *Escherichia coli* replicon R6κγori amplified from commercially available plasmid pMOD™-3 (Epicentre, UK) was cloned into *Pci*I linearized, blunt ended pFI2675, giving pFI2676. The replication of this well-characterised R6κγori replicon is mediated by the direct interaction with initiator π protein produced by the plasmid itself. It also requires additional protein produces by the host bacterial cells (Dellis and Filutowicz, 1991). The plasmids based on R6κγori were used in previous studies for the rescue of transposons therefore they can support the replication of large insert constructs and thus this replicon was chosen for construction pFI2676.

Structural rearrangements were discovered within the multiple cloning site (MCS) of plasmid pFI2676 (see section 3.2.3) rendering it unsuitable as a cloning vector. Therefore plasmid pFI2674 was cut with *Bst*UI and a *Sma*I linker was inserted resulting in pFI2709. The latter plasmid was digested with *Pci*I, blunt ended with T4 polymerase and the *E. coli* replicon R6κγori was inserted. The resulting construct was named pFI2710 and this shuttle vector was used for the construction of a *Ruminococcus* sp. 80/3 genomic library. All the constructs were propagated in *L. lactis* MG1363 according to method 2.6.2, apart from pFI2676 and pFI2710 which were selected in *E. coli* EC100D *pir*$^{+}$ after electrotransformation according to section 2.6.4.

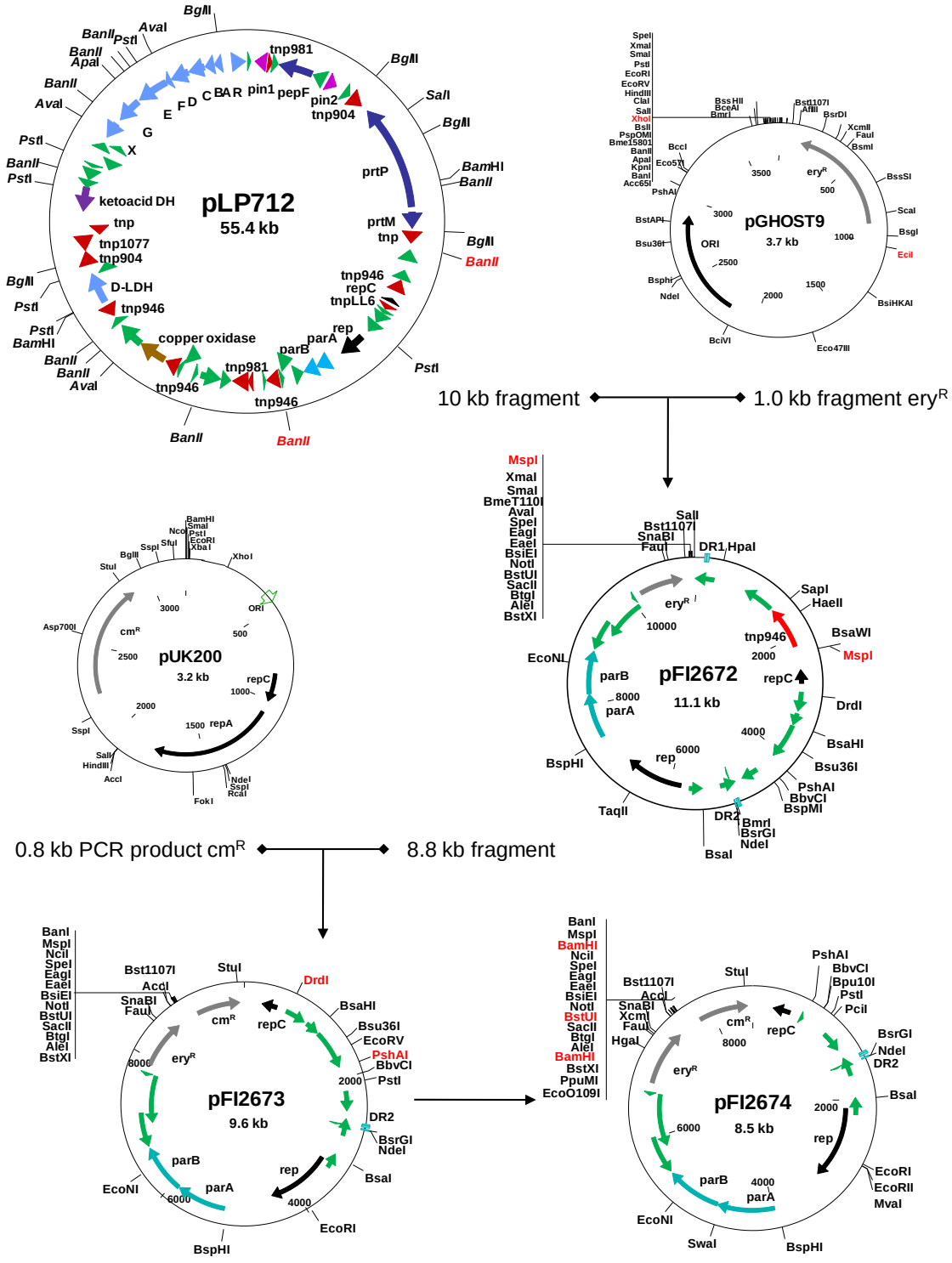


Figure 3.1 Construction of the pFI series vectors.

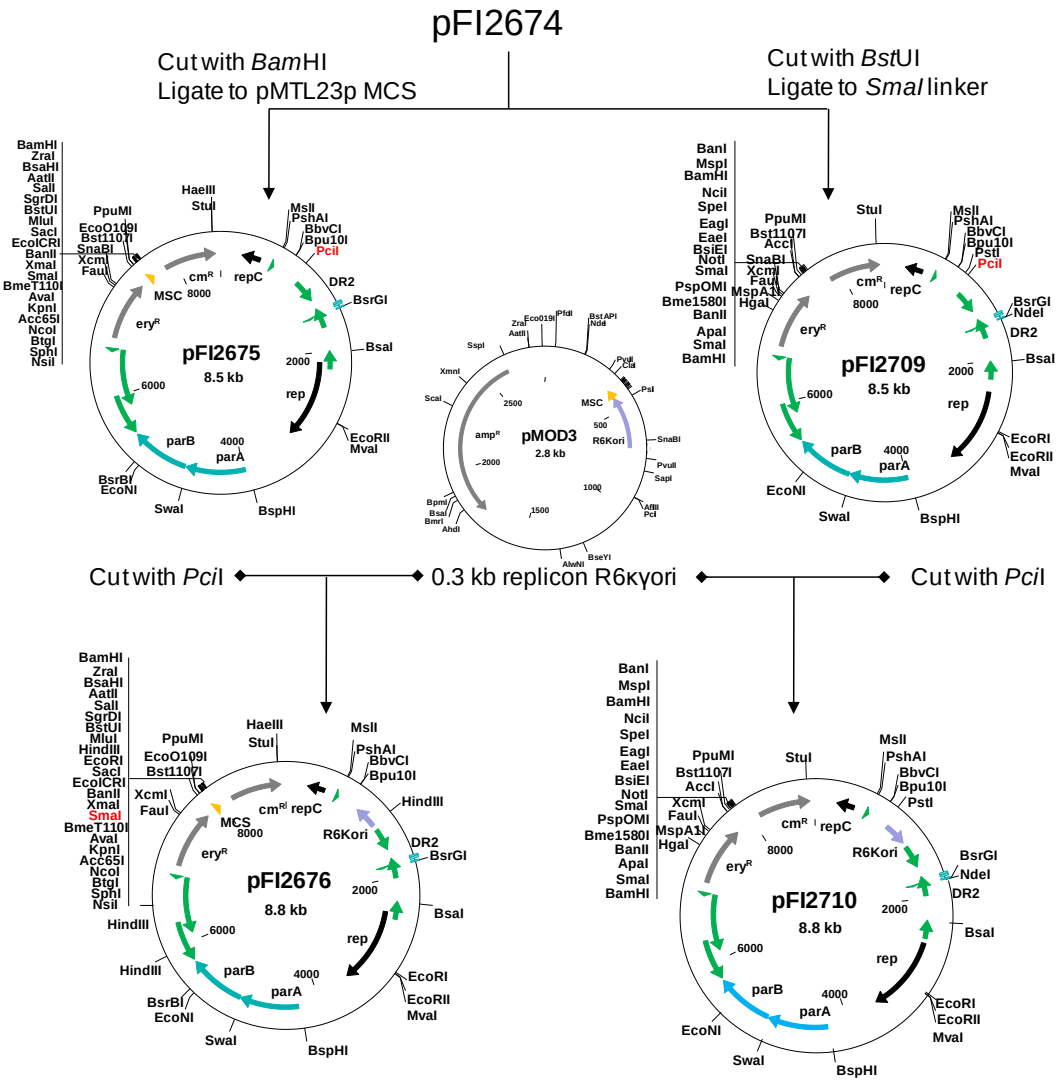


Figure 3.1 Construction of the pFI series of vectors.

The restriction enzyme sites used for genetic modification are marked in red. The arrows represent genes. The following genes are presented in this figure:

Lactose metabolism (*lacR* – transcriptional regulator, *lacA*, *lacB* – galactose-6-phosphate isomerase, *lacC* – tagose-6-phosphate kinase, *lacD* – tagose-1,6-phosphate aldolase, *lacF*, *lacE* – PTS translocation system, *lacG* – phospho-β-galactosidase, *lacX* – protein of unknown function, D-LDH – lactate dehydrogenase)

Transposition (*tnp* genes)

Proteolysis - peptide and amino acid utilisation (*prtP* – cell wall bound serine proteinase, *prtM* - maturase, *pepF* - oligopeptidase)

Plasmid replication (*rep*, *repC*)

Segregational stability (*parA*, *parB* – partitioning genes)

Ketoacid dehydrogenase

Copper oxidase

Hypothetical proteins

Antibiotic resistance genes

E. coli origin of replication

3.2.2 Stability of new plasmid derivatives based on pLP712

The maintenance of the plasmids harboring the pLP712 replicon was assessed in the absence of selective antibiotics using the method described in 2.4.3. The results of all tested constructs are summarized in Figure 3.2, in which the percentage of plasmid-containing cells is plotted against time. The results show that all derivatives were stably maintained during the entire testing period (approximately 100 generations, growth time approximately 60 hours).

Although the segregational stability of the tested constructs in *L. lactis* MG1363 was very high, the structure of the created plasmids proved to be unstable during cloning.

The first noticeable structural instability was observed during construction of a genomic library of *Ruminococcus* sp. 80/3, which tested the cloning abilities of plasmid pFI2676 and a new protocol was devised for future metagenomic library construction. Three independent libraries were prepared using plasmid pFI2676 and the numbers of clones in *L. lactis* MG1363 were 329, 1139 and 236 respectively. Amongst all these clones, 35 were screened by restriction enzyme digest and none showed the presence of an insert. This also led to the observation that the selected clones were lacking a *Bam*HI site. The multiple cloning site (MCS) region of pFI2676 was sequenced and showed structural rearrangement (Figure 3.3) which affected successful cloning. The instability was mediated by two short repeat sequences which participate in homologous recombination, causing *Bam*HI deletion. The product of pairing led to the formation of derivatives with a repeat sequence of 19 nucleotides.

A new derivative was constructed based on plasmid pFI2674, featuring a *Sma*I site and *E. coli* EC100D *pir*$^{+}$ replicon. Sequencing confirmed the absence of repeats and the segregational stability of the resulting plasmid pFI2710 proved to be high in *L. lactis* MG1363 (Figure 3.2). Plasmid pFI2710 was used as a cloning vector for genomic library construction; clones were screened (by restriction digestion and PCR) and stably maintained in *E. coli* EC100D *pir*$^{+}$ and *L. lactis* MG1363 (see section 3.3.1).

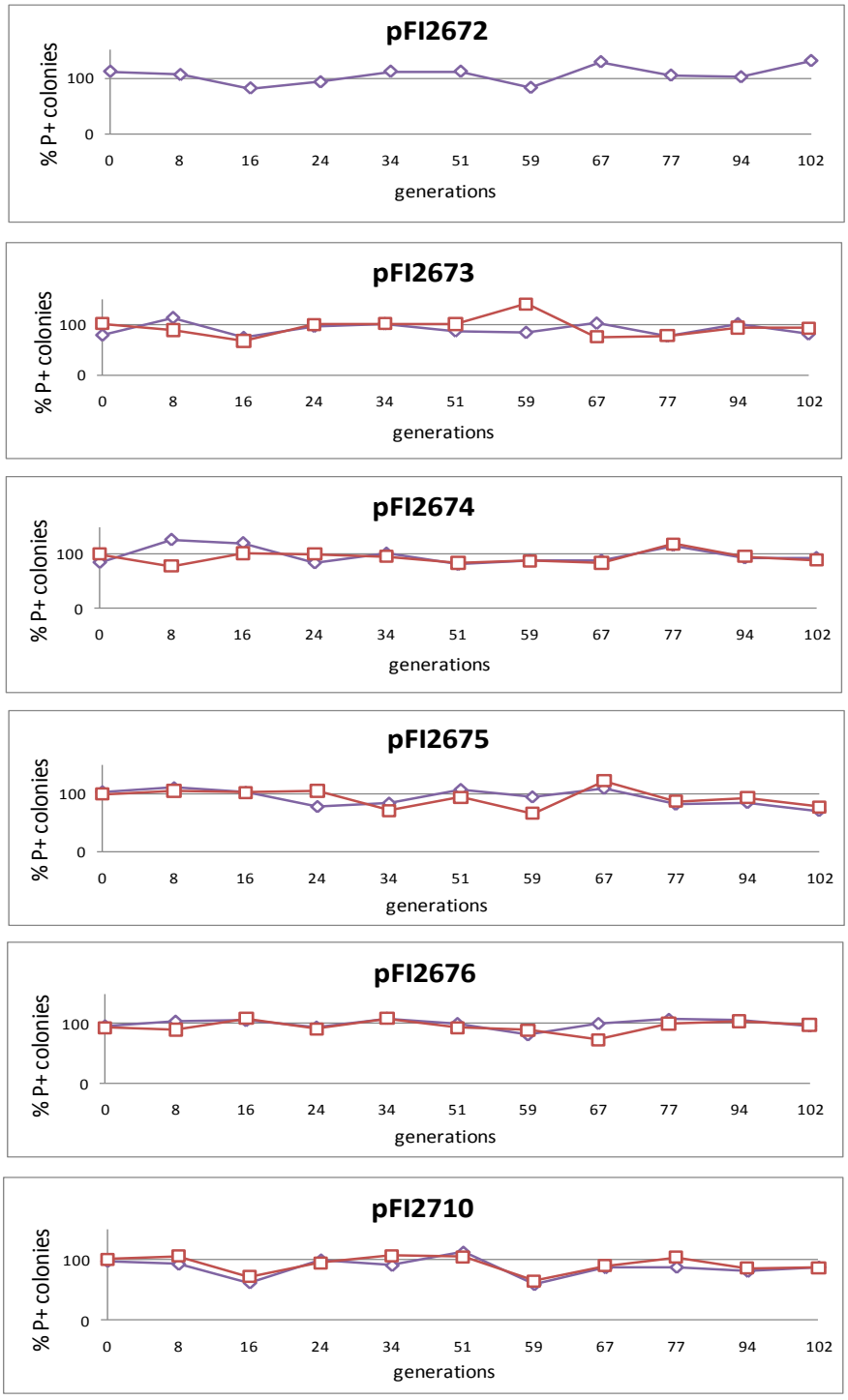
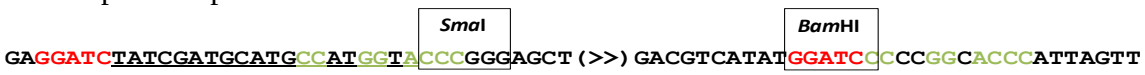


Figure 3.2 Presence of plasmid-carrying colonies of pLP712 derivatives in *L. lactis* MG1363.

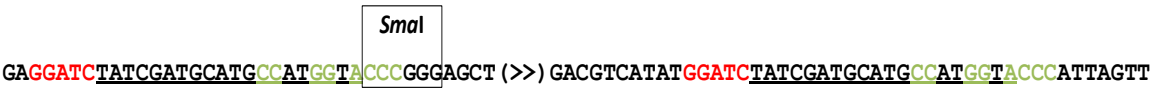
Cultures of plasmid carrying *L. lactis* were diluted in antibiotic free medium and subcultured for approximately 100 generations. Every 8 generations samples were plated onto selective and non-selective agar plates. The ratio of the average number of colonies from selective versus nonselective plates was multiplied x100. Erythromycin -resistant colonies - purple, chloramphenicol-resistant colonies - red.

A

Parental plasmid pFI2676



Recombination product



B

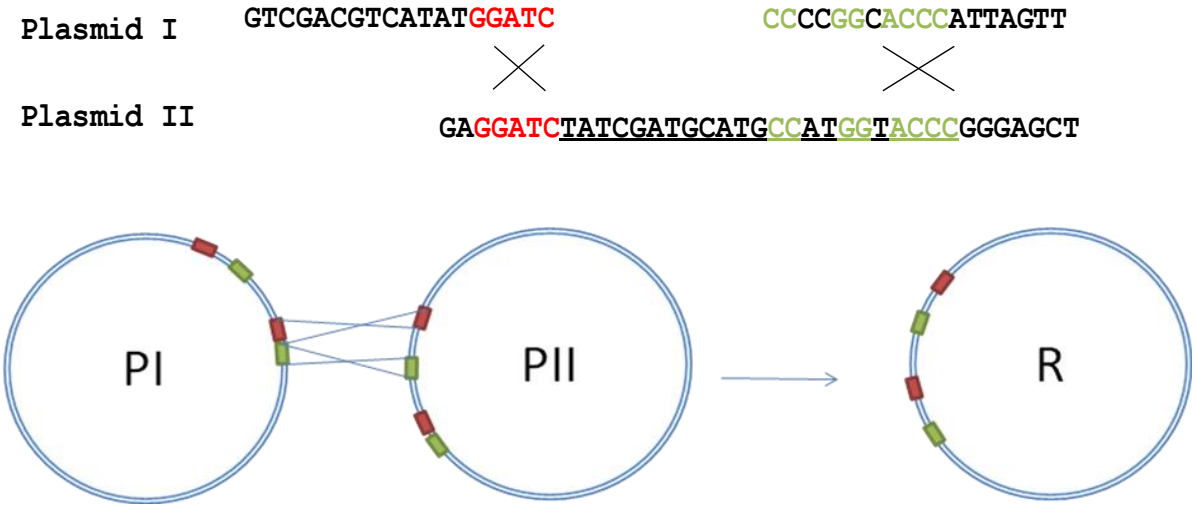


Figure 3.3 Schematic overview of homologous recombination between plasmids pFI2676.

Panel A – sequences of parental (original) plasmid pFI2676 and recombination product. Short repeat sequences which mediate homologous recombination are presented in red and green. *Sma*I and *Bam*HI sites are shown in boxes, they are both present in the parental plasmid but only the *Sma*I site is present in the recombinant plasmid. The underlined sequence is present only once in parental plasmid and twice in the recombinant. Panel B – exchange of the genetic material between two parental plasmids pFI2676.

3.3 Construction of **Ruminococcus** *sp. 80/3 genomic library*

Following construction of pFI plasmids, the next aim of this research was the creation of a large insert *Ruminococcus* sp. 80/3 genomic library using *L. lactis* MG1363 as the primary host and the newly created lactococcal plasmids pFI2710.
Several techniques were employed in order to prepare good quality, high molecular weight (~ 20 kb) insert DNA. Initially, fragments of genomic DNA were prepared as follows: shearing of genomic DNA was performed by passing through a 200 µl small bore pipette tip followed by electrophoresis analysis. If the DNA migrated with 23 kb λ DNA marker, end-repairing was performed, the enzyme was heat inactivated and the DNA was purified. Several independent libraries were prepared using plasmid pFI2710 and *L. lactis* MG1363 as a cloning host. The results demonstrated that direct transformation of *L. lactis* MG1363 with ligation mix yielded no positive clones. For *L. lactis* MG1363, highly reproducible transformation efficiencies were observed during plasmid construction (10^6 CFU. μg^{-1} DNA). In contrast, only low number of *L. lactis* MG1363 transformants was achieved during genomic library construction (average efficiency of 10^2 CFU. μg^{-1} DNA). A previous study showed that the insert size can decrease the transformation efficiency and is strongly dependent on the host (Sheng et al., 1995). Therefore, it was decided to decrease the insert size to 5-10 kb and use a second approach for the insert preparation which relies on mechanical fractionation of genomic DNA using a HydroShear machine. Two independent libraries were prepared applying Sure Clean or phenol extraction for DNA insert purification after shearing and end-repairing steps. The purification procedure affected the transformation rate which was three fold lower for phenol purified insert DNA. The libraries consisted of 435 and 135 colonies respectively. PCR confirmed the presence of an insert in 8 out of 24 screened clones with an average insert size of 2 kb. The results showed low efficiency of the cloning procedure that restricted the preparation of *Ruminococcus* sp. 80/3 genomic library directly in *L. lactis* MG1363. A report by Papagianni *et al.* (2007) reported high efficiency electrotransformation of *L. lactis* cells treated with lithium acetate and dithiothreitol. Several attempts to reproduce their work failed to increase the transformation rate of *L. lactis* MG1363. The decision was made to use *E. coli* EC100D pir^+ as a primary host for the library construction. The shearing procedure was optimized and produced inserts of 5-10 kb which were end-repaired and purified

with Sure Clean. Transformation into *E. coli* EC100D *pir*$^{+}$ produced 11,700 colonies with a transformation rate of 10^5 CFU. µg^{-1} DNA. The insert size of 12 randomly selected clones was estimated at 5 kb following restriction digestion analysis (Figure 3.4) which indicates cloning of 58.5 Mb genomic DNA.

The results showed that the observed transformation efficiency in *L. lactis* MG1363 (10^2 CFU.µg^{-1} DNA) was 1000-fold lower than in *E. coli* EC100D *pir*$^{+}$ (10^5 CFU. µg^{-1} DNA) using the same ligation mixture. The number of recombinants in *L. lactis* MG1363 was not sufficient to produce a representative genomic library. The higher transformation efficiency of commercially available electrocompetent *E. coli* EC100D *pir*$^{+}$ cells enabled the production of a genomic library from *Ruminococcus* sp. 80/3. The genome of *Ruminococcus* sp. 80/3 was covered 20-times based on 2.9 Mb size from draft genome information.

3.3.1 Transfer of *E. coli* library into *L. lactis*

In order to compare the insert distribution of *Ruminococcus* sp. 80/3 genomic DNA in both hosts, the *E. coli* EC100D *pir*$^{+}$ clones were pooled (method 2.7.1), plasmid DNA was extracted and re-transformed into *E. coli* EC100D *pir*$^{+}$ and *L. lactis* MG1363 electrocompetent cells. The results showed that the transformation efficiency was still lower in *L. lactis* ($2.4x10^5$ CFU. µg^{-1} DNA) compared to *E. coli* EC100D *pir*$^{+}$ ($2.5x10^8$ CFU.µg^{-1} DNA). However, it was 1000-fold more efficient than using the ligation mix. Next, 24 randomly selected transformants from *E. coli* EC100D *pir*$^{+}$ and *L. lactis* MG1363 were analyzed by restriction enzyme digest for the presence of insert DNA and 87% of the selected clones carried a DNA insert with an average size of 5.0 kb (~1.8 kb to 12 kb) for *E. coli* EC100D *pir*$^{+}$ and 4.5 kb (~3.5 kb to 7 kb) for *L. lactis* MG1363 clones (Figure 3.5). The presence of different fragments of *Ruminococcus* sp. 80/3 genomic DNA was confirmed by end sequencing insert DNA of random clones isolated from both hosts. The sequencing data were compared to draft genome information (Wegmann *et al*., unpublished data) and presented an equal distribution of random inserts from selected and sequenced clones within the bacterial genome (Figure 3.5). Each *E. coli* EC100D *pir*$^{+}$ clone restriction pattern was unique, indicating that a variety of different genomic DNA had been cloned. *L. lactis* MG1363 clones showed some degree of insert redundancy due to the amplification step in *E. coli* EC100D *pir*$^{+}$ followed by the library conversion.

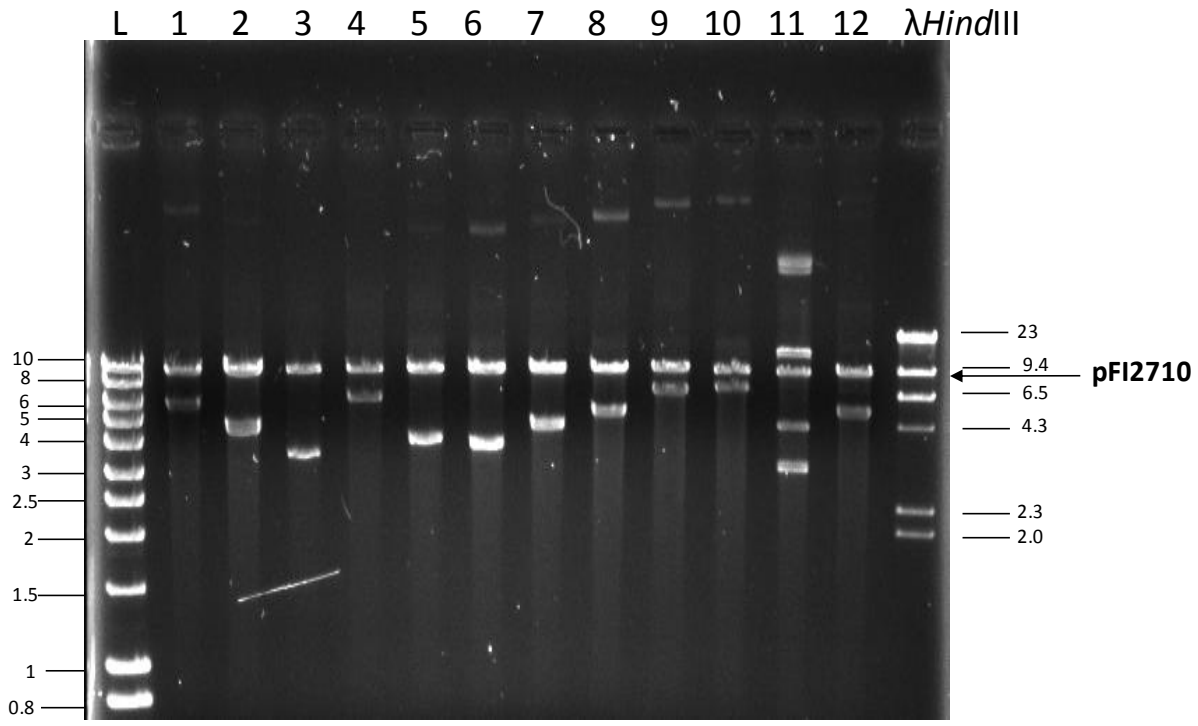


Figure 3.4 Randomly selected *E. coli* EC100D *pir*$^{+}$ clones which carry genomic insert DNA of *Ruminococcus* sp. 80/3.

Clones were analyzed by restriction digestion with *Bam*HI. The arrow represents plasmid pFI2710 (8.8 kb). 1-12 random clones, L: Hyperladder I, *Hin*dIII: *Hin*dIII λ DNA marker.

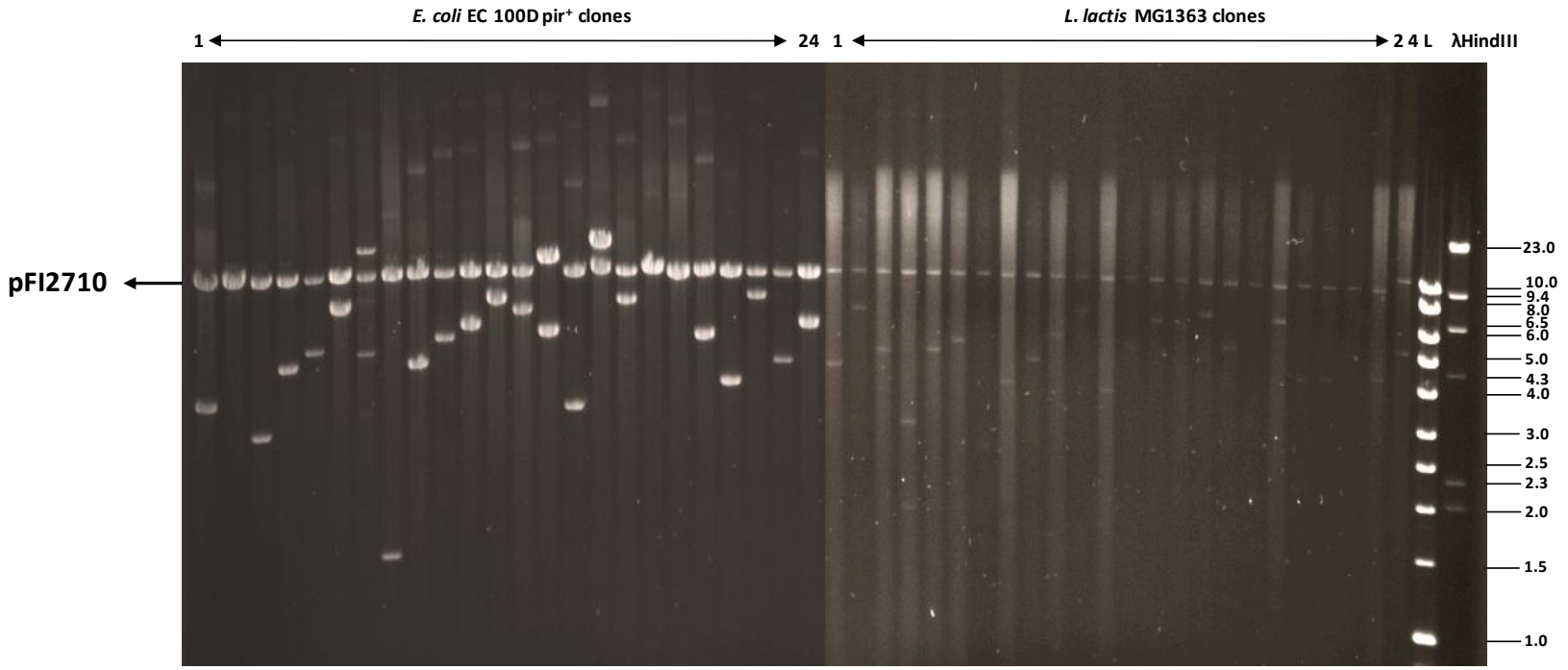


Figure 3.5 Restriction digestion profile of *E. coli* EC100D *pir*$^{+}$ and *L. lactis* MG1363 clones selected for insert end sequencing.

L- ladder [kb], λ*Hind*III ladder [kb], plasmid pFI2710 is shown with the arrow (8.8 kb). The random insert distribution suggests a variety of different fragments of genomic DNA from *Ruminococcus* sp. 80/3 that have been cloned in both hosts.

3.4 Functional screening of the genomic library of **Ruminococcus** *sp. 80/3*

An earlier study showed that *Ruminococcus* sp. 80/3 belongs to the cluster IV ruminococci group which is firmly associated with fiber particles in the human faecal samples (Walker *et al.,* 2008). The genomic library of *Ruminococcus* sp. 80/3 was used for function-based analysis in order to identify genes involved in dietary fiber breakdown. Briefly, the plasmid from the genomic library was electrotransformed into *E. coli* EC100D *pir*$^{+}$ or *L. lactis* MG1363 and a plate assay was done on substrate-containing plates according to section 2.8.1. Genes encoding β-galactosidase, β-glucosidase and cellulase were targeted in this work. These enzymes enable degradation of plant-derived components such as cellulose, β-glucan, xyloglucan, glucomannan, galactomannan and pectins (see Table 1.2).

In *E. coli* EC100D *pir*$^{+}$, 4500 clones were screened for β-galactosidase activity and 16 colonies were selected for further analysis (Table 3.1). Enzymatic activity for β-glucosidase was observed in 34 clones amongst 6200 screened clones. The numbers of clones screened for carboxymethyl cellulase (CMCase) activity was 1200 colonies amongst which four showed detectable enzyme activity. All the clones from the primary screening were streaked on substrate containing plates in order to confirm the detected enzyme activity. The positive re-screened clones were used for DNA extraction, followed by restriction enzyme analysis, the profiles of which are shown in Figure 3.6. The final number of positive clones excluding false positives picked during primary screening was 12 clones with β-galactosidase activity (0.26%, 1 per 2 Mb of cloned DNA), four clones with CMCase activity (0.33%, 1 per 1.5 Mb of cloned DNA) and 12 clones with β-glucosidase activity (0.19%, 1 per 2.6 Mb of cloned DNA). Clones showing a distinctive restriction profile were selected for sequencing and further analysis (five clones with β-galactosidase - pFI2710_3GA, _4GA, _5GA, 11_GA and _12GA, five clones with β-glucosidase – pFI2710_1GL, _3GL, _4GL, _7GL, and _12GL and two clone with CMCase activity- pFI2710_1CMC and pFI2710_2CMC).

Function-based screening of the genomic library of *Ruminococcus* sp. 80/3 was also performed in *L. lactis* MG1363. Transformants were screened for β-galactosidase and CMCase activity. Screening for β-glucosidase was impossible due to high background activity. No positive clones were detected during the screening in *L. lactis* MG1363 (374 colonies screened for β-galactosidase and 324 colonies

screened for CMCase). Therefore positive sequenced constructs from *E. coli* EC100D *pir*$^{+}$ were re-transformed into *L. lactis*, and enzyme activity was determined on substrate-containing plates. The results showed that *L. lactis* containing pFI2710_4GA, pFI2710_11GA or pFI2710_1CMC expressed β-galactosidase and CMCase activity, respectively, when screened on selective plates. The transformation with pFI26710_3GA, pFI2710_5GA and pFI2710_12GA plasmids showed no transformants (three independent experiments).

Enzyme Activity	β- glucosidase	β - galactosidase	CMCase
Clone assayed	6200	4500	1200
Number of positive clones	(34) 12	(16) 12	(4) 4
Frequency of positive clones	0.19 %	0.26 %	0.33 %
Clones sequenced	5	5	2

Table 3.1 The number of clones assayed functional screening of *Ruminococcus* sp. 80/3 genomic library in *E. coli* EC100D *pir*$^{+}$.

Number of positive clones selected during initial (number in brackets) and secondary screening is presented. The frequency of positive clones is shown as the percentage of total screened clones for specific enzyme activity. The selection of clones for sequencing was based as being different according to the restriction digestion profile (Figure 3.6).

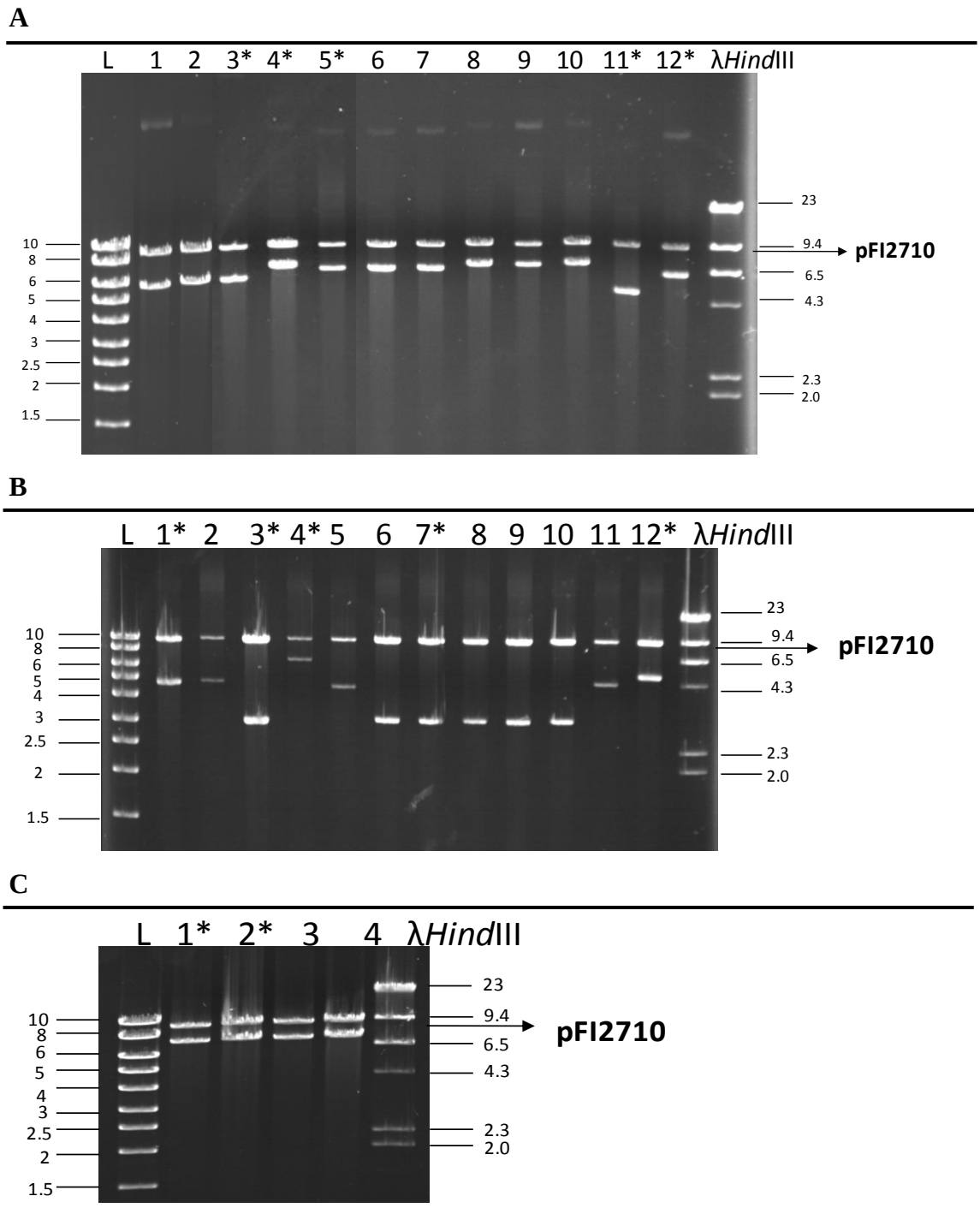


Figure 3 6 Agarose gel electrophoresis of *Bam*HI digested positive clones recovered during functional screening of *Ruminococcus* sp. 80/3 genomic library

β-galactosidase (panel A), β-glucosidase (panel B) and CMCase (panel C). Clones with (*) were selected for end-sequencing and further analysis, based on their restriction profile showing a different insert size. Plasmid pFI2710 is marked with an arrow (8.8kb). L – Hyperladder I, λ *Hind*III marker.

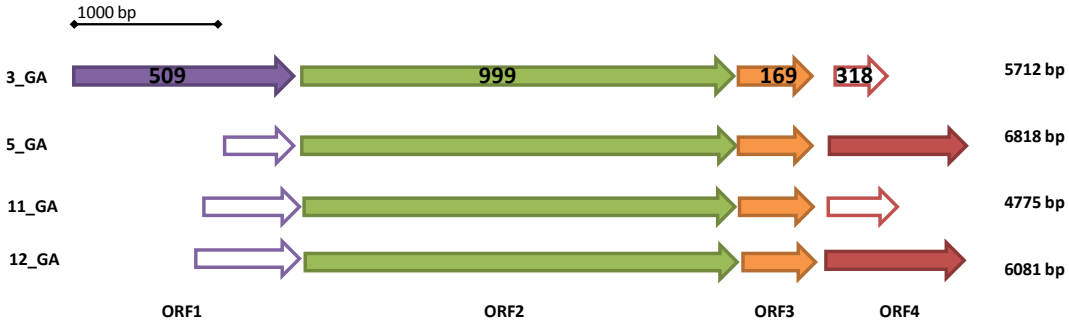
3.5 Sequence and bioinformatics analysis of clones selected from the **Ruminococcus** *sp. 80/3 library*

The cloned inserts were sequenced from both ends and the genomic information that they contained was obtained based on the draft genome sequence of *Ruminococcus* sp. 80/3 (Wegmann *et al*., unpublished data). The sequence representation of ORFs encoding glycoside hydrolases and their structure can be found in Appendix 1.

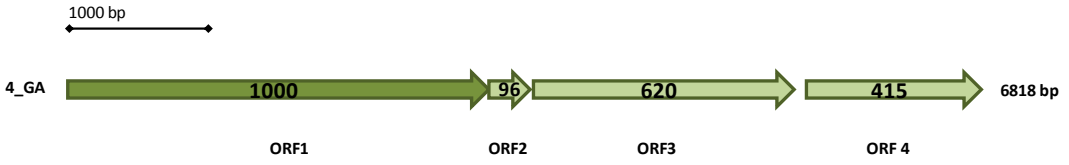
3.5.1 Analysis of β-galactosidase expressing clones

Five clones obtained for β-galactosidase activity carried inserts of different length due to random shearing of the genomic DNA. The schematic arrangement of the genes in each insert is presented in Figure 3.7. The insert from clones pFI2710_3GA, pFI2710_5GA, pFI2710_11GA and pFI2710_12GA contained several ORFs, one of which is encoding a protein with high similarity to a predicted β-galactosidase (ORF2). The inserts of these clones carry overlapping fragments of chromosomal DNA. The insert from clone pFI2710_ 4GA is distinctive and contained four ORFs, one of which encodes a protein with high similarity to putative β-galactosidases from *Ruminococcus albus* (ORF1). Downstream to the latter ORF, the three remaining ORFs (ORF2, ORF3 and ORF4) showed similarity to the gene from *Ruminococcus albus* encoding putative GH2 sugar binding protein.

The predicted number of putative β-galactosidases based on a draft genome information of *Ruminococcus* sp. 80/3 (Wegmann *et al.*, unpublished data) showed the presence of six genes encoding enzymes with similarity to proteins from the GH2 family. Five of these genes were defined during functional screening of the genomic library in this work, including ORF2 from clones pFI2710_3GA, pFI2710_5GA, pFI2710_11GA and pFI2710_12GA and ORF1 from clone pFI2710_ 4GA. ORF2, ORF3 and ORF4 from the latter clone showed also similarity to GH2 proteins. This indicates that one remaining gene predicted as β-galactosidase was not detected during functional screening of *Ruminococcus* sp. 80/3 library.



Closest BlastP match:				
ORF	ORF1	ORF2	ORF3	ORF4
Accession No	CBK96945.1	CBK96946.1	CBK96947.1	CBL16904.1
Source organism	*Eubacterium siraeum* 70/3	*Eubacterium siraeum* 70/3	*Eubacterium siraeum* 70/3	*Ruminococcus* sp. 18P13
Predicted function	sugar transporter	β- galactosidase	galactokinase	galactose-1-phosphate uridyl transferase
% identity	99	96	96	64
% similarity	99	98	97	78



Closest BlastP match:				
ORF	ORF1	ORF2	ORF3	ORF4
Accession No	YP_004103399.1	YP_004105882.1	YP_004105882.1	YP_004105882.1
Source organism	*Ruminococcus albus* 7	*Ruminococcus albus* 7	*Ruminococcus albus* 7	*Ruminococcus albus* 7
Predicted function	β-galactosidase	GH2 family sugar binding protein	GH2 family sugar binding protein	GH2 family sugar binding protein
% identity	59	57	61	42
% similarity	72	73	74	60

Figure 3.7 Schematic overview of inserts from *Ruminococcus* sp. 80/3 clones encoding β-galactosidases.

The arrows represent the ORFs detected in each insert. The full length ORFs are filled arrows, the open arrows represent truncated ORFs. The number shows the length of deduced protein in amino acids. The closest BlastP matches are given.

The predicted gene product of ORF2 from clones pFI2710_3GA, _5GA, _11GA and _12GA showed extensive amino acid sequence identity to a putative β-galactosidase/ β-glucuronidase from *Eubacterium siraeum* 70/3 (CBK96946.1). It also showed a high level of identity to well characterized β-galactosidase enzymes from *Clostridium acetobutylicum* (P24131.2), *Streptococcus salivarius* (AF389474.1) and *Streptococcus thermophilus* (AF389475.1) (Hancock *et al.*, 1991, Vaillancourt *et al.*, 2002).

A potential ribosomal binding site with the sequence GGAGG was observed at positions -12 to -6 with respect to the ATG start codon. Analysis of the upstream sequence to a translational start codon identified -10 region with the sequence TGGAAT. The extended promoter with -16 region characteristic for Gram-positive bacteria with the highly conserved motif TG was found for this β-galactosidase (see Appendix 1.1) (Voskuil and Chambliss 1998). The deduced amino acid sequence of ORF2 contained 999 amino acids residues. The protein sequence analysis showed the presence of domains characteristic for glycoside hydrolase family 2, which comprises several known enzymes exhibiting different activities including β-galactosidase (EC:3.2.1.23), β-mannosidase (EC:3.2.1.25) and β-glucuronidase (EC:3.2.1.31). The gene from *Ruminococcus* sp. 80/3 contains a conserved region (432-446 aa) with a glutamic acid residue. Glutamate acts as a general acid/ base catalyst in the active site which had been demonstrated in β-galactosidase (LacZ) and β-glucuronidase (UidA) enzymes from *E. coli* (Gebler *et al.*, 1992). The β-galactosidase from *Ruminococcus* sp. 80/3 has a TIM-barrel-like core (a conserved protein fold) and immunoglobulin-like beta-sandwich domain similar to the *E. coli* enzyme (Matthews 2005). The second conserved region (362-387 aa) is located around 60 residues upstream of the active site and is a signature pattern found in all enzymes from GH2 family.

The upstream and downstream region of ORF2 encodes genes found in well characterized gal-lac regulons (Figure 3.8) (Vaillancourt *et al.*, 2002) which consist of the genes involved in lactose and galactose metabolism. The sugar transporter (ORF1), galactokinase (ORF3) and galactose-1-phosphate uridyl transferase (ORF4) encoded by *Ruminococcus* sp. 80/3 are involved in substrate binding, transport across the membrane and galactose phosphorylation (Frey 1996) required for entering the LeLoir pathway. The cluster of these genes is well conserved across different

bacterial genera. However the gene organization in *Ruminococcus* sp. 80/3 is considerably shorter than *Streptococcus* gal-lac operons (Vaillancourt *et al.*, 2002).

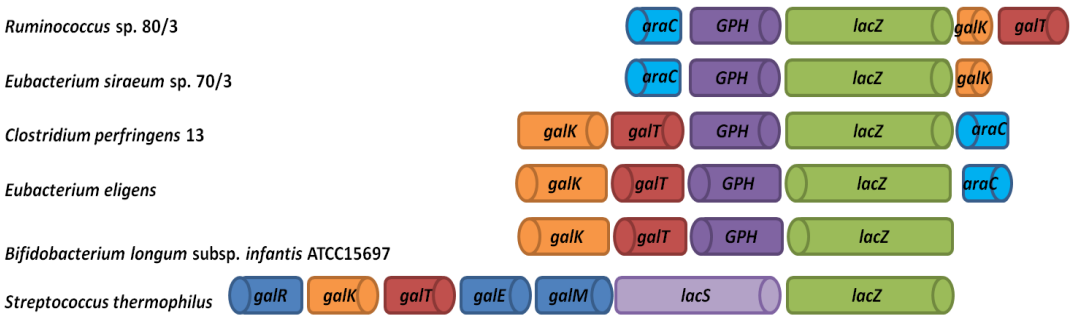


Figure 3.8 Organization of the galactose and lactose catabolism genes in various Gram-positive bacteria.

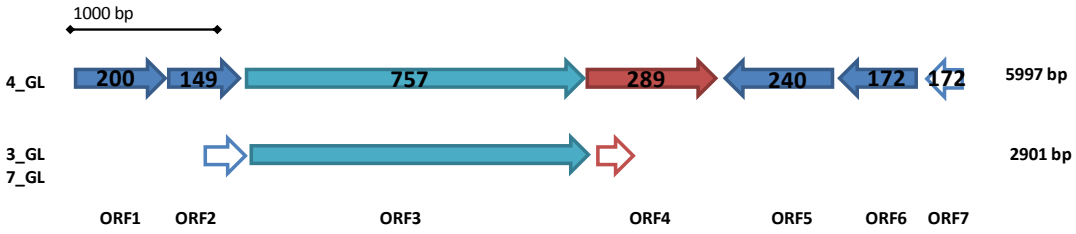
Cylinders represent gene and its orientation. The genes encodes for the following proteins: AraC family transcriptional regulator (pale blue), GPH - sugar transport protein (purple), *lacZ* - β-galactosidase (green), *galK* - galactokinase (orange), *galT* galactose-1-phosphate uridyl transferase (red), *galR* – transcriptional regulator, *galE* – UDP-glucose 4-epimerase, *galM* – galactose mutarotase (dark blue), lacS – lactose transporter (pale purple).

Clone pFI2710_4GA was shown to encode a second putative β-galactosidase (ORF1) from *Ruminococcus* sp. 80/3, with high similarity to enzymes from *Ruminococcus albus* strains (Figure 3.7). A putative ribosomal binding site at position -11 to -6 was identified in front of the start codon (see Appendix 1.2). Consensus promoter sequences were not identified in the cloned insert. The analysis of the genomic region upstream of *orf1* based on draft genome information showed the presence of extended promoter characteristic for Gram-positive bacteria. This enzyme encoded by *orf1* is predicted to be a multi-domain protein which consists of a N-terminal sugar binding domain, catalytic domain with a highly conserved glutamate residue and a C-terminal TIM-barrel-like domain similarly to previously described β-galactosidase encoded by ORF2 in clones pFI2710_3GA, pFI2710_5GA, pFI2710_11GA and pFI2710_12GA. The Blast analysis showed that these β-galactosidases share only 31% similarity (coverage 96%) and belong to two groups based on their amino acid sequence.

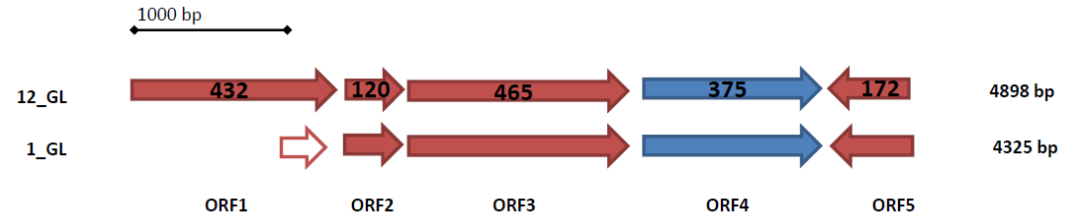
3.5.2 Analysis of β-glucosidase expressing clones

Screening for β-glucosidase activity allowed the selection and subsequent characterization of five positive clones that showed the enzyme activity on selective plates. Clones pFI2710_3GL, pFI2710_4GL and pFI2710_7GL harboured an insert that contained a complete ORF homologous to a known β-glucosidase (ORF3). Sequence analysis of clones pFI2710_1GL and pFI2710_12GL showed no homology to previously characterized glycoside hydrolase enzyme (Figure 3.9).

The gene product of ORF3 from clones pFI2710_3GL, pFI2710_4GL and pFI2710_7GL showed a high level of identity with a β-glucosidase from *Ruminococcus albus* strains. A putative ribosome binding site (RBS) AGAGG was present upstream of the translational initiation codon (see Appendix 1.3). Possible promoter sites were also identified. The -35 region sequence TAGAAT and -10 region sequence TACAAT were separated by a conserved TG motif previously observed in galactosidase genes of *Ruminococcus* sp. 80/3. The protein encoded by ORF3 belongs to glycoside hydrolase family 3 which are two-domain globular proteins. The N-terminal domain located between 27-248 residues contains the active site with a conserved aspartic acid residue (Harvey *et al.,* 2000). The C-terminal domain (312-522 aa) contains a catalytic acid residue and is likely to be involved in binding substrates such as β-glucans (Varghese *et al.,* 1999). A conserved COOH-terminal antiparallel loop sequence was also identified in the protein product of ORF3.



Closest BlastP match:							
ORF	ORF1	ORF2	ORF3	ORF4	ORF5	ORF6	ORF7
Accession No	YP_004104936	YP_004104935	YP_004104933	YP_004104932	ZP_05347445	ZP_02026321	ZP_02026323
Source organism	*Ruminococcus albus* 7	*Ruminococcus albus* 7	*Ruminococcus albus* 7	*Ruminococcus albus* 7	*Bryantella formatexigens* DSM 14469	*Eubacterium ventriosum* ATCC 27560	*Eubacterium ventriosum* ATCC 27560
Predicted function	hypothetical protein	hypothetical protein	β-glucosidase	Helix-turn-helix motif protein	hypothetical protein	hypothetical protein	hypothetical protein
% identity	79	59	68	41	26	50	38
% similarity	87	75	80	60	43	66	54



Closest BlastP match:					
ORF	ORF1	ORF2	ORF3	ORF4	ORF5
Accession No	YP_004103807.1	ZP_08158872.1	ZP_08158933.1	YP_004104720.1	ZP_08158467.1
Source organism	*Ruminococcus albus* 7	*Ruminococcus albus* 8	*Ruminococcus albus* 8	*Ruminococcus albus* 7	*Ruminococcus albus* 8
Predicted function	carboxyl transferase	glutaconyl-CoA decarboxylase	oxaloacetate decarboxylase	hypothetical protein	phosphodiesterase
% identity	69	79	93	75	86
% similarity	82	84	97	86	94

Figure 3.9 Schematic representation of inserts from *Ruminococcus* sp. 80/3 clones selected for β-glucosidase activity.

The arrows represent detected ORFs and their length in amino acids. The truncated ORFs are open arrows. The predicted function is described below. All the hypothetical proteins are shown as blue arrows. The genes of strain *Ruminococcus* sp. 80/3, the closest relatives and their identity and similarity (shown as percentage) according to BlastP search are shown in tables.

The clones pFI2710_1GL and pFI2710_12GL had been selected for β-glucosidase activity. However the gene products of the ORFs detected in the inserts did not show any similarity to previously characterized glycoside hydrolase enzymes. These two clones carry an overlapping fragment of *Ruminococcus* sp. 80/3 genomic DNA. ORF1, ORF2 and ORF3 showed high level identity to an oxaloaceate decarboxylate (OAD) gene cluster present in various anaerobic bacteria. The OAD genes are required for bacterial growth on different substrates and conversion of the energy through decarboxylation (Dimroth *et al.*, 2001). This free energy is used by bacteria to pump Na^+ ions out of the cell into the environment. The enzymes contain biotin as the prosthetic group and three subunits with α-, β- and γ- chains. ORF2 and ORF3 showed a high level of similarity to the well characterized α- and γ- subunit OAD from *Klebsiella pneumoniae* (Dimroth *et al.,* 2001).The deduced product of *orf4* gene showed homology to hypothetical proteins from *Ruminococcus albus* strains. A search in the proteins, domains and functional sites databases (e.g. ProSite, Interpro) showed no hits to previously characterized proteins (Appendix 1.4). The sequence between 200-318 amino acids classified this protein as a member of the glycoside hydrolase family 76, but the motif match was not significant (SMART and PFAM database, E- value 0.23). Family GH 76 is not very well characterized and the only known member is α-1,6 mannosidase. Several conserved amino acids including tryptophan, tyrosine and cysteine residues are found in known bacterial GH76 enzymes. Alignment of ORF4 with bacterial mannosidases suggested that this protein is not a member of the bacterial GH76 enzyme family.

The draft genome of *Ruminococcus* sp. 80/3 predicted three genes to encode β-glucosidases. In this study, functional screening of the genomic library recovered one gene predicted as β-glucosidase (ORF3 from clone pFI2710_3GL, pFI2710_4GL and pFI2710_7GL). The two remaining genes, predicted as putative β-glucosidases, were not detected during functional screening of genomic library. However, ORF4 from clone pFI2710_1GL and pFI2710_12GL, predicted to be a hypothetical protein was detected during screening on selective plates and could possibly encode a novel β-glucosidase.

3.5.2.1 Determination of enzyme activity of a novel β-glucosidase from *Ruminococcus* sp. 80/3

In order to confirm the enzyme activity of potentially novel β-glucosidases, encoded by pFI2710_1GL and pFI2710_12, a series of assays were conducted. The clone pFI2710_4GL with *orf3* encoding a predicted β-glucosidase was used as a positive control. The enzymic activity of clones pFI2710_1GL and pFI2710_4GL were assayed using p-nitrophenyl (pNP)–β-D-glucopyranoside according to method 2.8.3. Enzymes were prepared according to method 2.8.2. Briefly, a cell free extract from freshly grown culture of *E. coli* EC100D *pir*$^{+}$ was prepared and β-glucosidase activity was measured using a spectrophotometer. The concentration of released p-nitrophenyl was determined from a standard curve. The enzyme activity was also measured in supernatant from an overnight culture. β-glucosidase activity was detected in cell free extract (9.33 µmol. min^{-1}. ml^{-1}) and the supernatant fraction (0.34 µmol.min^{-1}.ml^{-1}) of clone pFI2710_4GL. The vast majority of β-glucosidase activity in recombinant *E. coli* was cell-associated. An increase of enzyme activity in the supernatant was observed when stationary phase was reached (data not shown). This may be due to bacterial cell lysis which led to release of the enzyme to the extracellular environment. No detectable enzyme activity was found in cells or supernatant fraction for clone pFI2710_1GL in samples prepared from an overnight culture or from samples taken in exponential or stationary growth phase. It was hypothesized that proteolytic inactivation of the protein might have been a reason for not obtaining assayable β-glucosidase activity of *E. coli* clone pFI2710_1GL from broth culture. Interestingly, however, using a plate assay (Figure 3.10), enzymatic activity was detected for both clones when using a chromogenic substrate for β-glucosidase.

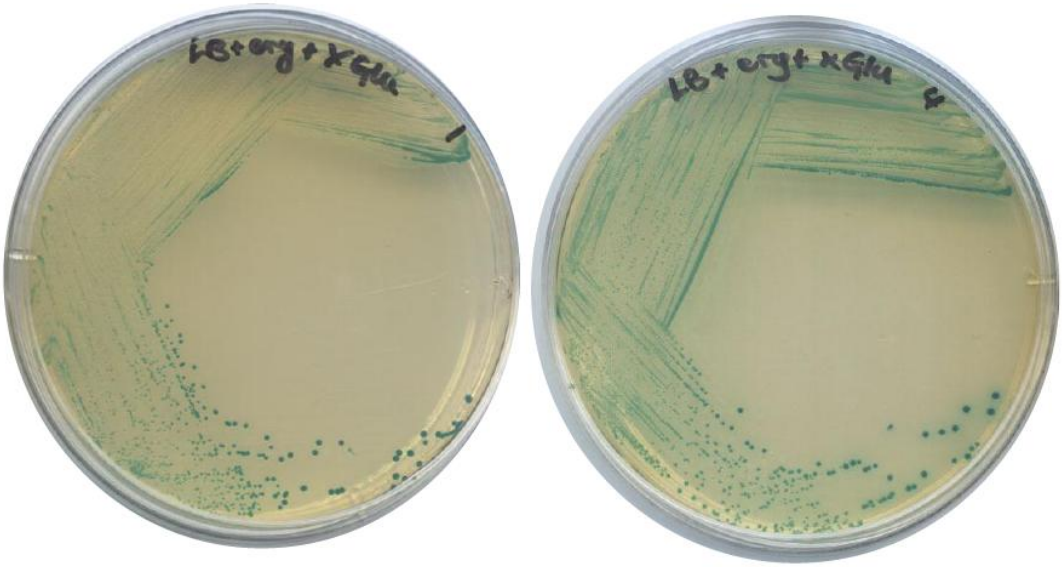


Figure 3.10 Plate assay for β-glucosidase activity in two clones derived from genomic library of *Ruminococcus* sp. 80/3.

A chromogenic substrate plate (LB medium) assay was performed to assess enzyme activity in clone pFI2710_1GL (left) and pFI2710_4GL (right). The streaking of the plates was performed and *E. coli* EC100D *pir*$^{+}$ clones were grown overnight at 37°C. The colonies on both plates are blue/green in colour showing the β-glucosidase activity. The bacterial host for the cloning *E. coli* EC100D *pir*$^{+}$ was tested on X-Glu plates and was negative.

ORF4 was predicted to be responsible for the enzyme activity in clone pFI2710_1GL. In order to inactivate *orf4* gene, the plasmid was cut with the single cutter *Bts*I and was treated with T4 polymerase in order to remove two base pairs which led to a frameshift mutation. The religation product (ΔpFI2710_1GL) showed no β-glucosidase enzyme activity on selective plates (not shown), confirming that this gene encoded β-glucosidase and might represents a new family of glycoside hydrolase.

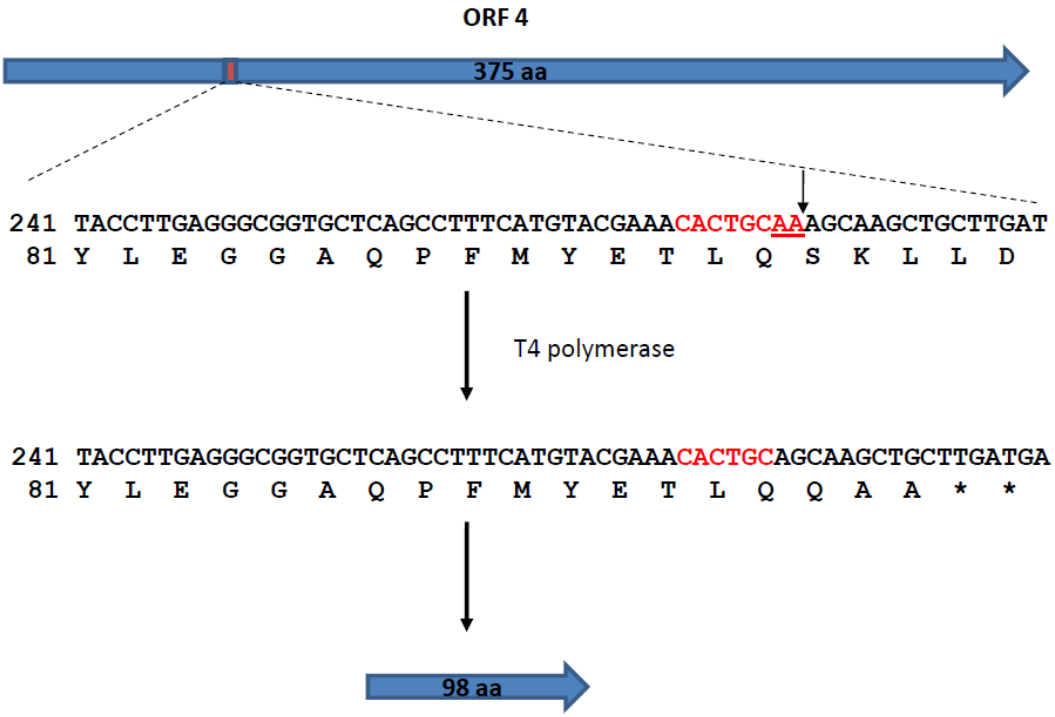


Figure 3.11 Inactivating mutation of ORF4 from clone pFI2710_1GL.
*Bts*I enzyme specific sequence is presented in red. Two nucleotides removed by T4 polymerase treatment are underlined.

Orf3 gene encoding putative β-glucosidase from clone pFI2710_4GL and ORF4 encoding the potentially novel glycoside hydrolase from clone pFI2710_1GL, were used for *in vitro* overexpression using RTS100 kit (see section 2.8.7). Two sets of primers were designed for cloning *orf3* and *orf4* into overexpression vectors pIVEX2.3d and pIVEX2.4d with a C-terminal His-tag and an N-terminal His-tag, respectively (see Table 2.3). Following *in vitro* overexpression β-glucosidase activity was determined using p-nitrophenyl linked substrate. Enzyme activity was successfully detected in N- (0.18 µmol.min^{-1}.ml^{-1}) and C- terminal (0.34 µmol.min^{-}

^{1}.ml^{-1}) His-tagged recombinant β-glucosidase ORF3 from clone pFI2710_4GL. The overexpression of a potentially novel β-glucosidase (ORF4) from clone pFI2710_1GL failed to display a detectable enzyme activity. The qualitative plate assay using chromogenic substrate also confirmed lack of activity of overexpressed ORF4.

The cell free overexpression system used in this study failed to obtain assayable enzyme activity of ORF4 from pFI2710_1GL. The results showed that clone pFI2710_1GL was only active when tested on agar plates using chromogenic substrate 5-bromo-4-chloro-3-indoxyl- β-D-glucoside (X-Glu). Previous studies reported differences between enzyme activities on different chromogenic substrates (Perry *et al.,* 2007). Edberg *et al*., (1986) demonstrated that some substrates may induce β-glucosidase activity in *E. coli* which could lead to misinterpretation of the results. The *E. coli* EC100D *pir*$^{+}$ strain used in this study for the functional screening was negative when tested on X-Glu-containing plates. Moreover this explanation was excluded by the inactivation experiment of ORF4 from clone pFI2710_1GL which produced white colonies. Future experiments are necessary to resolve this problem and confirm the discovery of a novel glycoside hydrolase enzyme.

3.5.3 Analysis of cellulase positive clones

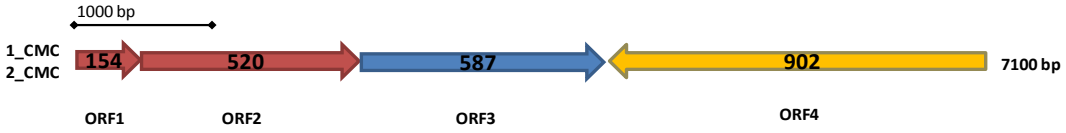
Two positive clones (pFI2710_1CMC and pFI2710_2CMC) that harboured the same insert of genomic DNA were selected during functional screening of the *Ruminococcus* sp. 80/3 genomic library on CMC-containing plates according to method 2.8.1 (Appendix 1.5). The insert contained several ORFs and a schematic overview is presented in Figure 3.12. The genome of *Ruminococcus* sp. 80/3 encodes 15 different genes which were predicted as cellulase/endoglucanase enzymes and were classified to the GH5 family.

The gene product of *orf4* showed high similarity to an endoglucanase from *Eubacterium siraeum*. The translation product of the full gene was a protein of 904 amino acids which display a modular architecture (Figure 3.13, Appendix 1.5). A putative ribosomal binding site (RBS) AAAGG was located upstream of the ATG translational initiation codon. The promoter site with -10 region TATAAT was identified as well as a TG -16 region. The N-terminal part of the *Ruminococcus* sp. 80/3 protein had a typical predicted signal peptide pattern (1-35 aa) with a cleavage

site between alanine and leucine. The N-terminal signal peptide of cellulase encoded by *orf4* could be involved in transport across the bacterial membrane since most of these enzymes are located extracellularly (Natale *et al.,* 2008). The region between 61 – 372 amino acids contains a cellulase domain found in glycoside hydrolase family 5 (Yoda *et al.,* 2005). Glycoside hydrolase family 5 comprises enzymes with several known activities; endoglucanase (EC 3.2.1.4), β-mannanase (EC 3.2.1.78), exo-1,3-glucanase (EC 3.2.1.58), endo-1,6-glucanase (EC 3.2.1.75), xylanase (EC 3.2.1.8) and endoglycoceramidase (EC 3.2.1.123). The signature pattern (187-196 aa) found in the *Ruminococcus* sp. 80/3 gene contains a glutamic acid residue that is potentially involved in the catalytic mechanism of the enzyme (Tull *et al.,* 1991). The second conserved glutamate residue (E329) acts as a nucleophile (Ducros *et al.,* 1995). Two catalytic residues (E194 and E329) in the *Ruminococcus* sp. 80/3 cellulase were found to be highly conserved amongst different GH5 enzymes (Appendix 1.6) (Posta *et al.,* 2004, Pereira *et al.,* 2010). Downstream of the cellulase region, a Ig-like domain was detected (407-484 aa). The Ig-like domain (Big-2) is usually found in a variety of bacterial and phage surface proteins such as intimins, as well as in some uncharacterized eukaryotic proteins. Intimin is a bacterial cell-adhesion molecule that mediates intimate bacterial host-cell interaction (Nougayrede *et al.,* 2003). At the C-terminus, two domains known as fibronectin type III domains were identified. These two regions (640-733 aa and 810-902 aa) share 62% homology when analyzed by BlastP. Fibronectin type III domains are characteristic for extracellular glycohydrolases; however they are randomly distributed across different GH families. These domains can play several functions: they could facilitate cellulose and other polysaccharide binding. They may coordinate interaction between enzymes involved in cellulosome assembly or they may be involved in the anchoring of the enzyme to the bacterial cell surface (Kataeva *et al.,* 2002, Xu *et al.,* 2004).

The remaining ORFs located downstream of the endoglucanase gene showed similarity to predicted proteins involved in synthesis of cell surface polysaccharides (ORF1 and ORF3) and amine oxidation (ORF2). ORF1 showed similarity to GtrA-like protein (PF04138) which is an integral membrane protein of bacteriophage origin and may be involved in glucose transport across the cell membrane (Adams *et al.,* 2001). The putative amine oxidase encoded by ORF2 is a predicted flavin-containing amine oxidoreductase (EC 1.4.3.4) which is a cell bound enzyme

produced by various bacteria, involved in amine oxidation (Murooka *et al.*, 1979). ORF3 showed a similarity to putative glycosyl transferase enzyme from family 39 with transmembrane domains. These enzymes transfer mannose units to serine or threonine residues of target proteins. Interestingly, all the genes detected in clone pFI2710_1/2CMC showed a high level of identity to the putative proteins (ADD61970-ADD61973) detected in a β-glucanase positive clone derived from the functional metagenomic study conducted by Tasse *et al.* (2010).



Closest BlastP match:				
ORF	**ORF1**	**ORF2**	**ORF3**	**ORF4**
Accession No	ZP_06142830.1	ZP_05793125.1	ZP_06142864.1	CBL34359.1
Source organism	*Ruminococcus flavefaciens* FD-1	*Butyrivibrio crossotus* DSM 2876	*Ruminococcus flavefaciens* FD-1	*Eubacterium siraeum* V10Sc8a
Predicted function	GtrA-like protein	Amine oxidase	Hypothetical protein	Endoglucanase
% identity	47	69	35	54
% similarity	63	81	53	68

Figure 3.12 Schematic representation of the insert from clones selected for cellulase activity.

The arrows represent the detected ORFs and their length in amino acids. The genes of strain *Ruminococcus* sp. 80/3, the closest relatives and their identity and similarity (shown as percentage) according to BlastP search are shown in the table.

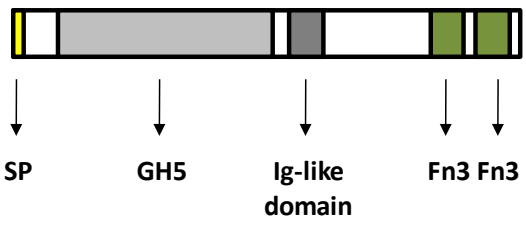


Figure 3.13 Schematic overview of domain structure of cellulase encoded by ORF4 in clones pFI2710_1CMC and _2CMC.

SP – signal peptide, GH5- catalytic domain, glycoside hydrolase family 5, Ig-like domain, Fn3 – fibronectin III domain.

3.6 Discussion

Several studies have now employed alternative hosts for genomic and metagenomic library construction in order to increase the number and diversity of positive clones (Craig *et al.*, 2010, Angelov *et al.*, 2009). *Lactococcus lactis* MG1363 had been used as a heterologous expression host in a number of studies (Morello *et al.*, 2008). Carbohydrate active enzymes derived from different microorganisms had been cloned and expressed in this background. The expression of rumen-derived β-glucanase was examined in *L. lactis* by Ekinci *et al*. (1997) and showed detectable enzyme activity. Other carbohydrate active enzymes including xylanases and cellulases have also been studied using lactic acid bacteria as a heterologous host (Liu *et al.*, 2007, Ozkose *et al.*, 2009). The present study assessed the suitability of *L. lactis* for the construction of a metagenomic library. Novel lactococcal based plasmid vectors were created and used to construct a genomic library of the gut bacterium *Ruminococcus* sp. 80/3. The library was screened for a number of enzyme activities involved in carbohydrate breakdown since the bacterium was shown to utilize complex β-glucan polymers such as cellulose, lichenan and xylan (weak activity).

Several attempts were made to produce the genomic library directly in *L. lactis* but a limited number of transformants was obtained. Initially the limiting factor was structural instability of plasmid pFI2676 which was shown to be a major issue preventing successful cloning. A new derivative pFI2710 was created and used for the further cloning study. The latter shuttle vector incorporates several features which make it useful for cloning in *L. lactis* and *E. coli*. It possesses erythromycin (Ery^R) and chloramphenicol (Cm^R) resistance genes. Genes conferring Cm^R and Ery^R have been used to genetically mark many lactococcal and nonlactococcal plasmids including derivatives of pWV01, pAMβ1 and pIP501 (Mills *et al.*, 2006). These two genes were expressed in *L. lactis* MG1363 and *E. coli* pir^+ which made them easily selectable. Although the erythromycin encoded by pFI2710 was expressed in *E. coli* pir^+, the selection had to be done on BHI medium, similarly to work done by O'Sullivan and Klaenhammer (1993). Plasmid pFI2710 has also got a multiple cloning site (MCS) inherited from plasmid pLP712 with several commonly used restriction sites (*Bam*HI, *Sma*I etc. see Figure 3.1). Although the structure of plasmid pFI2710 improved, only a limited number of transformants were produced restricting the preparation of a genomic library in *L. lactis.* Geertsma *et al.* (2007) experienced

similar difficulties in their study, which led to library construction in *E. coli* followed by a transfer into *L. lactis.* Accordingly, this approach was used for production of the *Ruminococcus* sp. 80/3 library in the present work. Other studies also reported using an intermediate *E. coli* step with subsequent transfer into the desired host (Angelov *et al.,* 2009). A major limitation for direct library construction in many hosts other than *E. coli* is still a lack of convenient cloning vectors and a low efficiency of transformation.

Functional screening in *E. coli* yielded several clones encoding predicted glycoside hydrolase enzymes such as β-galactosidase, β-glucosidase and cellulase activity that were closely related to those previously reported in species of clostridial cluster IV. β-galactosidase activity was shown to be high in faecal samples from volunteers fed on different diets (P. Louis, personal communication). Moreover, a metagenomic study has demonstrated a relatively large number of sequences being assigned to GH2 family (Tasse *et al.,* 2010). As shown by Dabek *et al*. (2008), β-glucosidase activity is also common among the colonic microbiota. Out of 40 different bacterial strains, which belong to the predominant groups of human gut microbiota, detectable β-glucosidase activity was found in 23, with the highest activity found in cell free extracts of *Bifidobacterium* species. *Ruminococcus* sp. 80/3 was also tested during this study and showed the highest activity of all strains examined from clostridial cluster IV. The present study reported a potentially novel glycoside hydrolase, with β-glucosidase activity. A cellulase (CMCase) from *Ruminococcus* sp. 80/3 showed a modular architecture and shared strong similarity with GH5 enzymes (see Appendix 1). The similar domain structure was found in well characterized endoglucanase D from *Cellulomonas fimi* (Meinke *et al.,* 1993) and cellulase Cel5G from *Ruminococcus albus* 8 (Xu *et al.,* 2004).

The number of clones screened in *L. lactis* MG1363 was not sufficient to detect positive clones due to lower transformation efficiency in this host. The study showed however that plasmids pFI2710_1CMC, pFI2710_4GA and pFI2710_11GA were successfully transformed and expressed in *L. lactis* MG1363. In contrast, plasmid pFI2710_3GA, pFI2710_5GA and pFI2710_12GA did not produce any transformants in *L. lactis* MG1363 which could indicate the presence of genes in the insert (ORF1 or ORF4) that had a toxic effect on the host. The toxicity of the recombinant genes derived from *Ruminococcus* sp. 80/3 could have been a major

limitation in this research. A previous study showed that foreign genes propagated in *E. coli* can have a lethal effect on the host (Holt *et al.*, 2007).

Recently genome of *Ruminococcus* sp. 80/3 have been sequenced and manually annotated by Wegmann *et al*. (unpublished data). This information offered the possibility of identifying glycoside hydrolase genes sequences and comparing it to results derived from functional screening of *Ruminococcus* sp. 80/3 genomic library produced in this study. Analysis of the 2.9 Mbp genome of *Ruminococcus* sp. 80/3 had shown the presence of 47 putative glycoside hydrolase enzymes belonging to 11 different GH families. The most abundant GH families in the *Ruminococcus* sp. 80/3 genome are GH5, GH2, GH3 and GH13.

The functional screening of the genomic library validate the *in silico* prediction of some of the ORFs encoding enzymes from GH2, GH3 and GH5 families. It was shown that less than the half (30%) of predicted ORFs from those families were detected during functional screening in *E. coli*. Unfortunately, the screening of the library in *L. lactis* did not detect positive clones. However, it has been shown that ORF4 in clone pFI2710_1GL and pFI2710_12GL which was annotated as a hypothetical protein and was not assigned as GH enzyme could be potentially novel GH enzyme since it showed the β-glucosidase activity on selective plates. Further analysis of this gene should verify its enzyme specificity.

This study confirmed the metabolic potential of *Ruminococcus* sp. 80/3 involved in carbohydrate utilization, previously established by *in vitro* study (S. Duncan and M. Pudenz, unpublished data). It also showed that functional screening is a powerful tool in discovery of novel enzymes. The function based approach and the protocol established for *L. lactis* heterologous expression will be applied to study a more complex ecosystem, the human gut microbiota.

Chapter 4

Exploring carbohydrate active enzymes from the human gut microbiota by applying a functional metagenomics approach

4.1 Introduction

Following work presented in chapter 3 which established the protocol for using *Lactococcus lactis* as an alternative heterologous host, the investigation was now focused on a complex microbial community from human gut. Other reports showed that the genomes of bacteria within the human gut microbiota contain many genes encoding glycoside hydrolases, providing enzymes for dietary fibre fermentation, which the human genome lacks coding capacity for (Gill *et al.*, 2006, Kurokawa *et al.*, 2007, Tasse *et al.*, 2010). As a result of this microbial fermentation, the gut microbiota has a profound effect on human nutrition and health by metabolizing dietary compounds which escaped digestion in the upper GI tract and which cannot be degraded by the host. The metagenomic approach was shown to successfully extract information on glycoside hydrolase enzymes from different environmental samples (Li *et al.*, 2009). It was also presented that the commonly used *E. coli* expression system may fail due to differences in codon usage, promoter recognition signals or protein-processing while the use of alternative hosts can improve the efficiency of function-based study (Martinez *et al.*, 2004, Craig *et al.*, 2010, Angelov *et al.*, 2009).

At this stage of the research, a metagenomic library of the human gut microbiota was constructed and used for high-throughput functional screening for polysaccharide and plant cell wall biomass degrading enzymes using *E. coli* and *L. lactis* as heterologous hosts.

4.2 Metagenomic library preparation

4.2.1 Choice of plasmid and insert size

The previously created shuttle vector pFI2710 was used for metagenomic library construction during initial trials. However, it was observed that plasmid pFI2710 increased in size after transformation and purification from *E. coli* EC100D *pir*$^{+}$ and was not able to replicate in *L. lactis* MG1363. Sequencing results demonstrated the presence of an *E. coli* IS element that caused disruption of the lactococcal replication gene. In the context of a previous cloning study with *Ruminococcus* sp. 80/3, plasmid pFI2710 was clearly a functional shuttle vector. However, the structural instability of construct pFI2676 due to homologous recombination and pFI2710 due to transposition of IS element, showed that further testing and possibly improvement in

their design was required. Due to these structural rearrangements occurring within the plasmid it was decided to use another shuttle vector. Plasmid pTRKL2 was chosen since it can replicate in *E. coli* and *L. lactis* (O'Sullivan and Klaenhammer 1993) and had previously been used for a number of cloning studies. The pTRKL2 plasmid contains a well-characterized theta-replicating pAMβ-1 replicon from *Enterococcus faecalis* and the *E. coli* P15A replicon (Bruand *et al.,* 1993). The plasmid encodes erythromycin resistance, which is expressed in Gram-positive bacteria and *E. coli.* It contains a *lacZ* cassette with a multiple cloning site which is amenable to blue/white screening. It is a low-copy number plasmid in *L. lactis* (6-9 copies per cell) and moderate copy number in *E. coli* (30-40 copies per cell). Unfortunately, there were no reports of large inserts cloned into pTRKL2. The largest insert sizes cloned into pTRKL2 came from a *L. lactis* LM0230 genomic library and were on average 7.9 kb (Dudley and Steele 2001). Therefore it was decided to create a small insert metagenomic library.

4.2.2 Sampling and DNA extraction

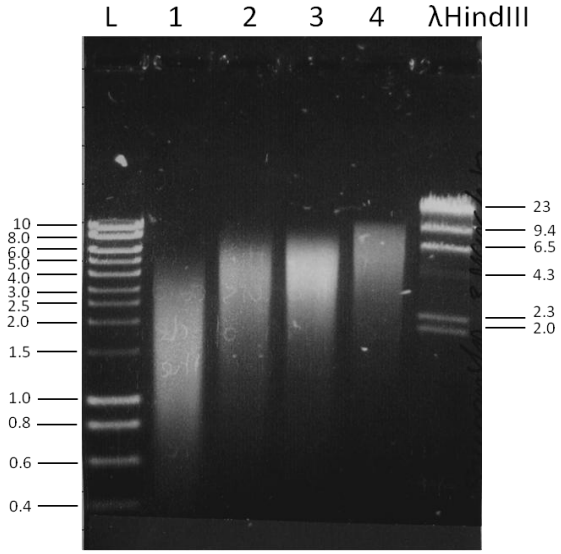
A healthy, 27 year old, female volunteer, consuming a Western diet, provided a fresh faecal sample for this study. The volunteer did not take any antibiotics or other drugs known to influence the faecal microbiota for six months prior to the study.

The desired average insert size was 5-10 kb, therefore commercially available kits applying chemical or mechanical cell disruption were used (QIAamp DNA stool kit QIAGEN, Extract Master Fecal DNA, Epicentre and Fast®DNA Spin kit for soil, Table 4.1). Mechanical lysis of cells is a harsh but very efficient method for DNA recovery from a wide range of bacteria. The Fast®DNA Spin kit system (method 2.4.5) resulted in the best yield of metagenomic DNA, for freshly processed and freeze-dried samples. The extraction was more efficient than using the QIAamp DNA stool kit from QIAGEN, which was tested only with freshly prepared samples (Table 4.1). The Epicentre kit gave the lowest yield of metagenomic DNA (data not shown). For mechanical cell disruption using the FastPrep instrument (FP120 machine, BIO 101, Thermo Savant), three different times and speeds were tested in order to establish the optimal DNA profile. The results showed that sample processed or 15 sec at speed setting 6.5 gave the least sheared DNA profile and the highest DNA yield compared with the other tested settings (Figure 4.1).

Extraction method	Sample 1			Sample 2		
	Yield µg.DNA g^{-1} faeces	A260/280	A260/230	Yield µg.DNA g^{-1} faeces	A260/280	A260/230
QIAamp DNA stool kit (fresh sample)	23	2.1	0.72	nd	nd	nd
Fast®DNA Spin kit (fresh sample)	59	1.79	0.33	136	1.80	0.90
Fast®DNA Spin kit (freeze-dried sample)	nd	nd	nd	146	1.80	1.0

Table 4.1 DNA yield and absorbance ratios of metagenomic DNA from a human faecal sample purified with different commercial kits.

Absorbance (A) was measured for purified DNA using the NanoDrop apparatus. Two samples collected at different time points were used for the DNA extraction.
nd = not determined



Aliquot	Time [sec]	Setting[a]	Yield µg.DNA g^{-1} faeces
1	30	6.5	162
2	45	4.5	101
3	15	6.5	198
4	30	5.5	112

Figure 4.1 Metagenomic DNA purified by modified mechanical cell disruption using Fast®DNA Spin kit for soil.

Equal aliquots of the same faecal sample were used for metagenomic DNA purification using different time and speed as presented in the table. Equal volumes of purified DNA were loaded into each lane. L – Hyperladder I [kb], λHindIII marker [kb]. Setting = speed on FastPrep FP120 machine, BIO 101, Thermo Savant.

4.2.3 Phylogenetic analysis of faecal sample DNA

The diversity of the microbial population in the extracted faecal DNA was assessed by analysis of the 16S rRNA gene to establish whether the used sample contained a typical microbial profile. The sequence analysis and tree construction were done according to section 2.7.6. Briefly, the 16S rRNA gene was amplified from the metagenomic DNA extracted from the human faecal sample and cloned into pGEM®-T vector, followed by transformation into *E. coli* XL1 Blue competent cell (see method 2.6.4). The V4 region of 16S rRNA gene was sequenced from 288 clones using internal primer 797R. The final dataset included 263 chimera-checked sequences and contained estimated 74 phylotypes using 98% identity cut-off which could be mapped to four bacterial phyla (Figure 4.2). At the phylum level the proportions of Bacteroidetes (37.6%), Firmicutes (61.2%), Actinobacteria (0.38%) and Proteobacteria (0.76%) were consistent with previous reports (Tap *et al.*, 2009, Duncan *et al.*, 2007). Within Firmicutes phylum, the vast majority of sequences grouped into two families Lachnospiraceae (35.7%) and Ruminococcaceae (44.2%) which comprise clostridial cluster XIVa and IV, respectively. The Bacteroidetes sequences were predominantly from the Bacteroidaceae family (81.8%) but also included the Prevotellaceae (9.1%) and the Porphyromonadaceae (6.1%) families. The most abundant phylotypes fell into different phyla (Table 4.2). A recent study by Walker *et al*. (2011) showed that abundant phylotypes contained more sequences of cultured strains than uncultured strains. It was reported that the sequenced phylotypes with a frequency higher than 2%, represent bacterial strains that were cultured. The present study showed that most abundant phylotypes were cultured except for clone 107H1 (3% of total clones) which showed 87% identity to *Clostridium xylanolyticum*, therefore it could represent a novel phylotype. During the analysis of 16S rRNA data in the present project, 24 phylotypes showed less than a 95% similarity to previously cultured bacterial strains (Fig. 4.2). Their abundance was usually less than 0.5% and in accordance with Walker *et al*. (2011) those phylotypes primarily derived from uncultured bacteria. Therefore, these clones were classed as potentially novel phylotypes.

Phylotype, species	**Phylum**	**% total clones**
Bacteroides plebeius	B	21.7
Bacteroides dorei	B	6.1
Paraprevotella clara	B	3.4
Faecalibacterium prausnitzii M21/2	F (R)	8.0
Anaerostipes coli SSC/2	F (L)	6.1
Ruminococcus bromii	F (R)	4.6
Eubacterium rectale	F (L)	4.2
Clone 107H1	F (u)	3.0
Coprococcus eutactus ART55/1	F (L)	2.3
Faecalibacterium prausnitzii SL3/3	F (R)	2.3

Table 4.2 The most abundant 16S rRNA phylotypes detected in the faecal sample used for the metagenomic study.

B – Bacteroidetes, F – Firmicutes, L – Lachnospiraceae, R – Ruminococcaceae, u - unclassified. Cut off ≥ 2%.

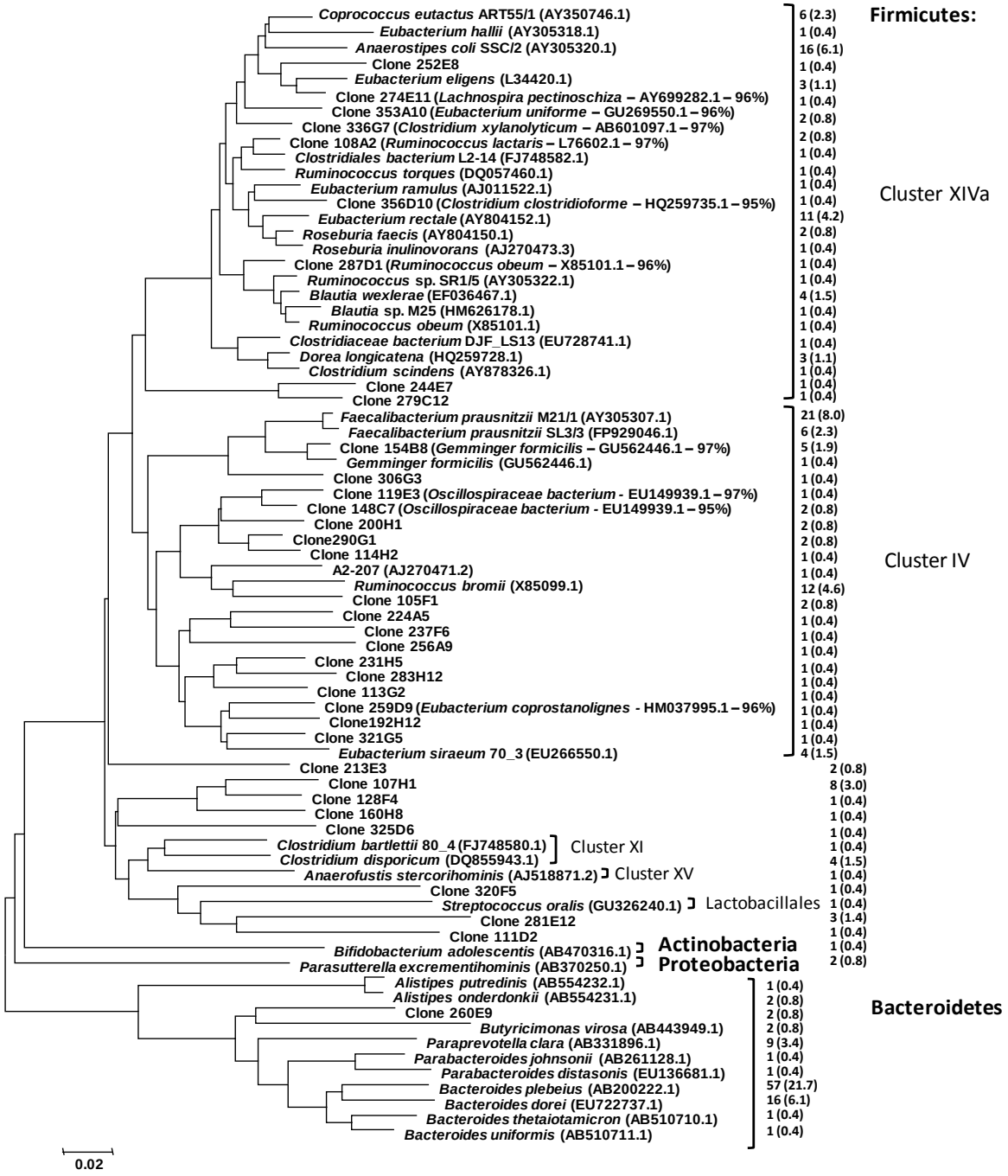


Figure 4.2 Phylogenetic tree of 16S rRNA gene sequences detected in the human faecal sample used for metagenomic library construction.

The tree was constructed using partial 16S rRNA gene sequences of 263 clones, grouped into phylotypes based on at least 98% sequence identity. The number of sequences per phylotype and their percentage is presented. The sequence closest bacterial identity (if ≥ 98%) is shown, based on the results from BLAST search at NCBI database. Where the closest sequence identity was <98% the closest cultured species are indicated in parentheses. Clones with sequence identity ≤95% to cultured species are presented by their clone ID. The branch length index is presented by a bar at the bottom of the figure.

4.2.4 Library gridding and transfer into the *L. lactis* MG1363 host

The mechanical cell disruption with the FastPrep® instrument caused shearing of the metagenomic DNA, leading to good yield of fragments of the correct size; therefore no further size fragmentation was needed before gel-purification of the desired fraction. DNA fragments (5-10 kb) were separated by agarose gel electrophoresis (method 2.4.7) and recovered by electroelution (method 2.4.8) followed by purification and end-repairing (method 2.4.11). The purified DNA fraction was successfully used for ligation (see 2.4.12) with pTRKL2 and subsequent transformation into *E. coli* XL1 Blue (method 2.6.4), with a transformation efficiency of $3.0x10^6$ CFU.μg^{-1} DNA. The insert frequency based on ratio of white and blue colonies was 32%. The average insert size of the metagenomic library was estimated at 2.5 kb following PCR screening of 24 randomly picked colonies (Figure 4.3).

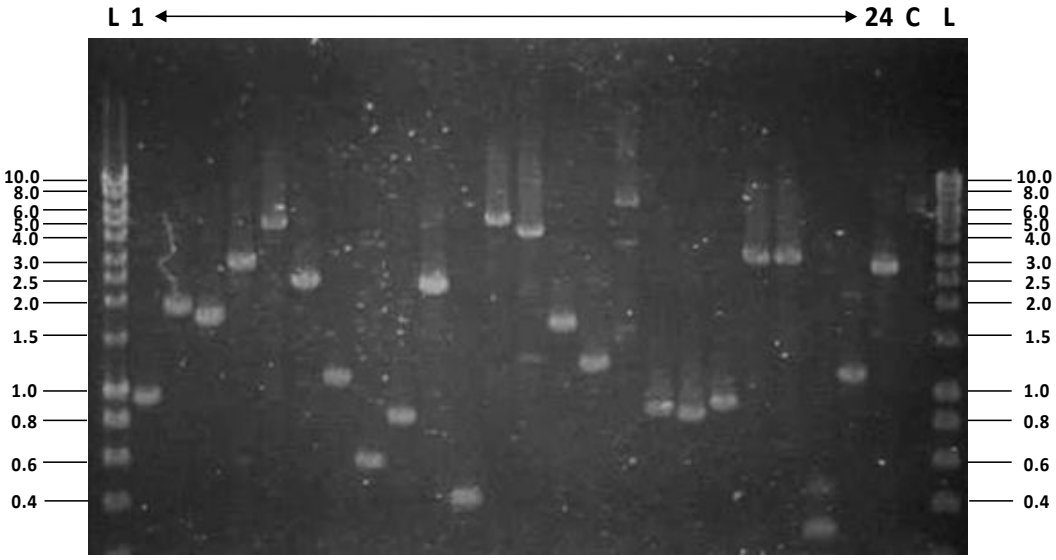


Figure 4.3 Agarose gel electrophoresis of randomly picked clones from *E. coli* XL1 Blue metagenomic library screened by colony PCR.

The insert DNA was amplified from both end using M13_F and M13_R primers according to section 2.5.1. The insert size ranges from approximately 0.4 kb to 7 kb, with an estimated mean of 2.5 kb. C- plasmid pTRKL2, L – Hyperladder I [kb].

The metagenomic *E. coli* clones were stored in 384 well microtiter plates using a manual or automatic colony picking technique (Table 4.3). Manual picking was more accurate than the automated procedure. No bacterial growth was observed in 11% of clones picked by the automated system. In the case of manual picking this number was 65-fold lower (Table 4.3). The number of negative clones (blue colonies which do not contain an insert) picked by the robot was also much higher than for manually picked clones (Table 4.3). The total number of clones assumed to carry an insert (~2.5 kb) in the *E. coli* library was 6146, which gives 15.4 Mb of cloned metagenomic DNA.

In order to generate a library in *L. lactis* MG1363, freshly grown *E. coli* XL1 Blue clones were pooled (method 2.7.2) and plasmid DNA was extracted. Recombinant plasmid DNA was then transformed into *L. lactis* MG1363 (see section 2.6.2) and 4608 clones were picked and stored for functional screening. The average insert size was established based on PCR screening of 15 random colonies and it was estimated at approximately 2.5 kb with 75% insert frequency, which gives 11.5 Mb of cloned metagenomic DNA.

		Quality control of the library		
Picking technique	**Number of clones**	**No growth[1]**	**No insert[2]**	**Positive clones[3]**
Manual	3456	6 (0.17%)	6 (0.17%)	3444
Automatic	3456	394 (11%)	360 (10%)	2702
Total	6912	400	366	6146

1 – The growth of clones was established based on OD600 measurement, the threshold was 0.1.

2 – The lack of insert was determined based on β-galactosidase plate assay (blue colonies).

3 – Final result shows the final number of positive clones excluding clones with no growth and no insert.

Table 4.3 Number of clones from *E. coli* metagenomic library.

The number of clones with no growth and no insert was determined, respectively, in batches picked manually or automatically.

4.4 Functional screening

The next aim of this research was to examine hydrolytic activities of cloned genes from the human gut microbiota using functional screening of the metagenomic libraries. The aim was to detect glycoside hydrolase enzymes involved in dietary carbohydrate degradation. Several commercially available substrates which chemically resemble the diet-derived carbohydrates were used. These substrates enable screening for a variety of enzymatic activities (see table 1.2). Starch from potato was used to screen for amylolytic activities; carboxymethyl cellulose was used as a soluble derivative of cellulose to detect cellulolytic activity; xylan derived from oat spelt and was used to screen for xylanolytic activity; lichenan is a commercially available β-glucan with higher proportion of β-1,4 to β-1,3 linkages than in cereal β-glucans; polygalacturonic acid from citrus fruit was used to screen for the pectinolytic activities (see Table 2.4). The *E. coli* and *L. lactis* libraries were screened for the ability to degrade these substrates and specific detection method was used according to section 2.8.1. Congo Red (Teather and Wood 1982), ruthenium red (Jayasankar and Graham 1970) or iodine solution was used to detect hydrolysis zones around active clones, indicating specific polysaccharide degradation. Additionally, two commercially available fluorescence substrates to screen for arabinofuranosidase and rhamnopyranosidase activities were used and plates were examined under UV light. The function-based screening applied a BioRobotics system which arrayed all the *E. coli* XL1 Blue (6912) and *L. lactis* MG1363 (4608) clones on trays with substrate-containing medium (method 2.7.3). The initial screening led to identification of positive colonies whose enzyme activity was confirmed by secondary screening. All positive clones were selected by inspection of the substrate-containing plates.

4.4.1 Functional screening of *E. coli* library

In total 55 clones were obtained in *E. coli* during primary screening (Table 4.4), including 24 starch degrading clones, 16 carboxymethyl cellulose (CMC) degrading clones, one clone with lichenase and one clone with arabinofuranosidase activity. Moreover, 12 clones showed dual activity on starch and CMC-containing plates and one clone was active on CMC-, lichenan- and xylan-containing plates. During

secondary screening all positive clones were screened again according to method 2.8.1 in order to confirm the enzyme activity. The results confirmed enzyme activity for 16 *E. coli* clones, including three clones with amylase activity, two clones with CMCase activity and one clone with arabinofuranosidase activity (Table 4.4). Nine clones showed dual activity on starch and CMC-containing plates and one clone was active on CMC, xylan and lichenan containing plates. No positive clones were detected on polygalacturonic acid- and rhamnopyranoside-containing plates.

The discrepancy between primary and secondary screening indicates a high rate of detection of false positive clones during the primary screening. Furthermore, the recovery of positive clones differed upon repeat screening of eight 384-well plates. During the first screen, 36 clones were selected for further analysis, but only 13 positive clones were selected upon repeat screening (Figure 4.4). The difference between clones recovered from the same set of plates at different times of screening was substantial and raised the question of the reliability of the detection method. However, all the clones that showed any zones of hydrolysis were selected for secondary screening. During secondary screening, differences in enzymes activities were noted (Figure 4.5), with some of the clones producing very bright and typical clearing zones, while other clones generated much smaller and indistinctive halos. The background host *E. coli* XL1 was tested during this study and showed no activity on any of the substrates used for the functional screening. The inserts from all positive clones from the secondary screening assay were sequenced and analysed.

	Number of *E. coli* clones	
Substrate	**Primary screening**	**Secondary screening**
Starch	24	3 (0.04%)
CMC	16	2 (0.03%)
Starch/CMC	12	9 (0.13%)
CMC/ xylan/lichenan	1	1 (0.01%)
Lichenan	1	0
Polygalacturonic acid	0	0
Arabinofuranoside	1	1 (0.01%)
Rhamnopyranoside	0	0
Total	55	16

Table 4.4 Number of *E. coli* clones selected during primary and secondary screening on different substrate-containing plates.

The frequency of positive clones for different activities is shown in brackets. CMC – carboxymethyl cellulose.

A **B**

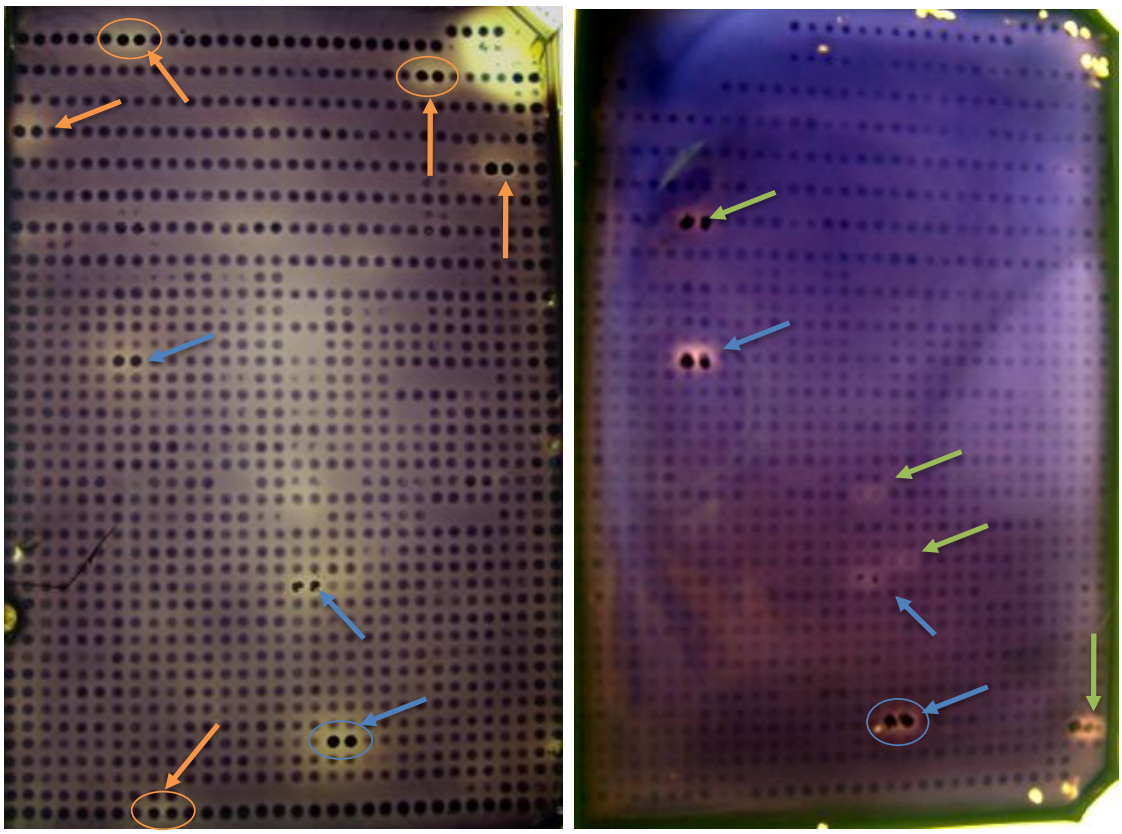


Figure 4.4 Example of reproducibility during primary screening for amylase-positive clones using starch-containing trays and iodine solution as the detection method.

A and B represent the same set of clones (arrayed in duplicate) screened independently during the study at two different times. Clones selected for further analysis are marked with arrows. Three clones (blue arrows) were selected from both plates. An irreproducible halo pattern was observed for five clones on plate A during the first screening (orange arrows) and for four clones during the repeat screening on plate B (green arrows). The right top corner of plate A contains amylase positive control. The clones whose enzyme activities were confirmed during secondary screening and were selected for further analysis are circled in plates A and B.

4.4.2. Functional screening of the *L. lactis* library

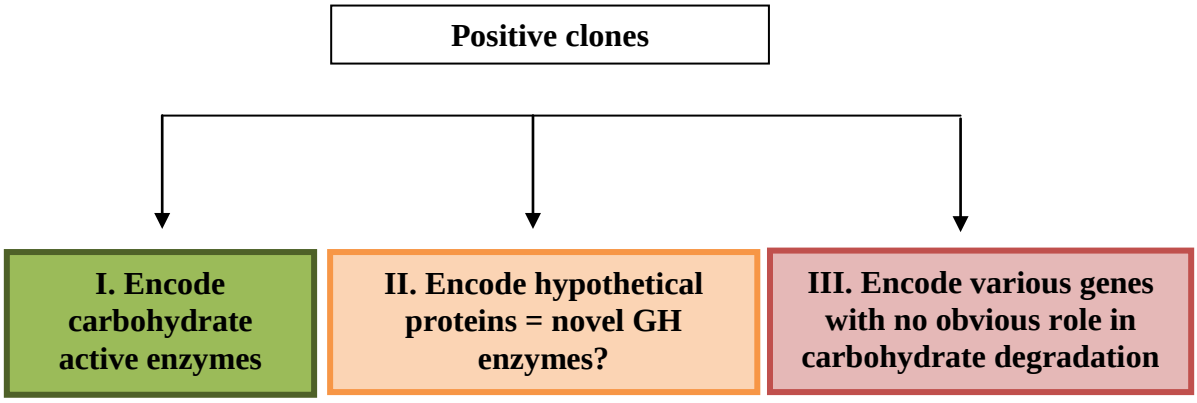
Functional screening of the *L. lactis* library led to identification of three clones which all showed enzyme activity on CMC-, xylan- and lichenan-containing plates. Enzyme activity was confirmed in all of the clones by secondary plate screening (method 2.8.1). Based on the clearance zones the enzyme activity was higher than in most of the *E. coli* clones. The restriction digestion profile and sequencing showed the presence of the same insert DNA in these three clones (P20A8, P20P7 and P25N22). Clone bias has likely been introduced during growth of the *E. coli* library prior to clone pooling and plasmid purification. Significant growth differences were observed between freshly grown *E. coli* clones before pooling. The OD_{600} ranged from 0.2 to 1.2, with the average between all clones 0.4. Hence, an unequal representation of pooled plasmid DNA led to the sequence redundancy observed for the positive *L. lactis* clones. All positive clones from the secondary screening of the *E. coli* metagenomic library (Table 4.4) were re-transformed into *L. lactis* and the enzyme activity was determined by using substrate-containing plates. The results showed that none of the *L. lactis* clones expressed the enzyme activities previously observed in *E. coli* (data not shown).

4.5 Individual clone analysis

Each clone confirmed as positive during the secondary screening was sequenced and subsequently analysed using bioinformatics tools which allowed gene prediction and taxonomic classification (see method 2.5.2). Positive clones were categorised according to their carbohydrate degradation profile and based on the predicted genes (Figure 4.5). Firstly, there was a group of clones containing ORFs with predicted glycoside hydrolase enzymes and domains characteristic of known GH proteins. Analysis of the insert sequences detected genes from several glycoside hydrolase families, including family 9, 13, 43 and 51. These enzymes are known to be involved in the degradation of starch, cellulose and arabinose-containing polysaccharides. Secondly, there are clones encoding hypothetical proteins which could potentially be novel glycoside hydrolase enzymes. Thirdly, there are clones which encode ORFs of miscellaneous functions not obviously related to carbohydrate degradation, including regulatory factors, transport and conjugation proteins. Clone P20A8, which was

recovered during screening in *L. lactis* MG1363 and was active on CMC-, lichenan and xylan-containing plates, is also described; a detailed analysis of this interesting clone is presented in chapter 5.

A total of 63 kb of non-redundant metagenomic DNA was sequenced, with an average insert size of 3.8 kb. Sixty (40 completed and 20 truncated ORFs) protein-encoding genes were predicted within the sequences of all the clones. The majority showed high similarity to sequenced genes from gut bacteria. Twenty-five percent of the predicted ORFs were most similar to sequences annotated as hypothetical protein or predicted protein in the NCBI-nr database. The predicted genes mostly derived from the predominant genera of the human gut microbiota, including *Roseburia, Ruminococcus, Lachnospiraceae, Bacteroides* and *Alistipes*. The observed high degree of similarity to genes from sequenced genomes is not surprising, since the genome information of many abundant gut bacteria are available. In addition, several clones showed low identity to cultured and sequenced microorganisms, hence the insert in those clones likely originated from novel bacteria.



Category	Clone ID	Substrate	Activity	Source organism	Length [bp]	Accession numbers
I	P1E14	CMC	+	F(R)	3637	JN807448
	P3B15	AF	+++	F (L)	3470	JN807445
	P3H22	S	+	B	3202	JN807446
	P5H21	CMC	+	B	4130	JN807447
	P20A8*	CMC/XL/L	+++/++/+++	F(L)	2643	JN807460
II	P1I16	S/CMC	++/++	F(R)	2754	JN807450
	P1P9	S/CMC	++/+	B	5693	JN807449
	P3C3	S/CMC	+/++	F/B	2202	JN807452
	P5D24	S/CMC	+/+	B	5531	JN807456
	P5E1	S/CMC	+/+	B	3949	JN807451
	P5J3	S/CMC	++/++	F(R)	4246	JN807453
	P7I21	S	+	F/B	5402	JN807454
	P9N7	CMC/XL/L	+++/++/+++	B	2655	JN807455
III	P1D18	S/CMC	+/+	F(L)	2918	JN807458
	P3N11	S/CMC	++/++	B	3927	JN811090
	P5M23	S/CMC	+/+	F/B	2994	JN807459
	P8I17	S	++	F(L)	4049	JN807457

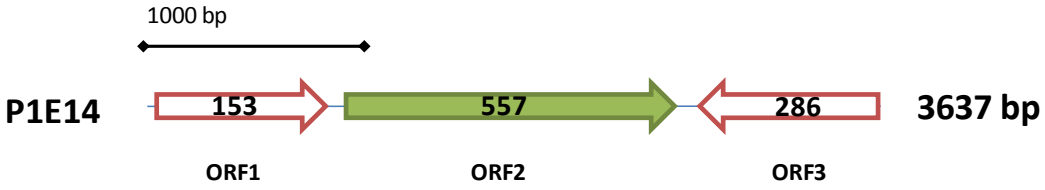
Figure 4.5 Categories of positive clones selected during secondary functional screening of the metagenomic library.

Based on sequence analysis, three groups of clones in relationship to carbohydrate metabolism were defined. The table presents the clone ID, substrate for which activity was found, the results of the enzyme activity assay on substrate-containing plates and predicted bacterial origin. S = starch, CMC = carboxymethyl cellulose, XL = xylan, L = lichenan, AF = arabinofuranoside. +++ = strong activity, ++ = moderate activity, + = weak activity. F = Firmicutes, L =Lachnospiraceae, R = Ruminococcaceae, B = Bacteroidetes, v = various, * = clone from *L. lactis* screening.

4.5.1 Clones with homology to predicted glycoside hydrolase enzymes – category I

Clone P1E14

The insert of clone P1E14 encoded three genes with a variable degree of similarity to predicted proteins of *Ruminococcus bromii* L2-63 (Figure 4.6). *Orf2* encoded a putative glycosidase with an α-amylase catalytic domain, which contains three crucial active-site residues (D235, E270 and D341) within conserved domains identified previously in α-amylase family enzymes (Oslancova and Janecek 2002) (Figure 4.7). The N-terminus of the ORF2 encoded protein has a characteristic lipoprotein signal peptide with a cleavage site between glycine and cysteine. Upstream of *orf2*, one gene was detected showing homology to a component of well characterised maltose transport system which is composed of a periplasmic maltose-binding protein and two ABC type sugar transport system proteins with homology to pore forming subunits MalF and MalG (Bordignon *et al.*, 2010). ABC transporters in the maltose transport system show a modular architecture, comprising transmembrane domains and nucleotide-binding domains. The presence of a transmembrane structure was confirmed for ORF1, which featured membrane spanning segments (Bordignon *et al.*, 2010). The maltose/maltodextrin system enables the bacterium to utilise maltose and maltodextrins, which are derived from enzymatic cleavage of starch or glycogen. ORF3 encodes a predicted anthranilate synthase component I which catalyses the conversion of chorismate and glutamine to anthranilate, glutamate and pyruvate (Balderas-Hernandez *et al.*, 2009). Surprisingly, clone P1E14 was selected during functional screening on CMC-containing plates and showed weak enzymatic activity. No activity was observed on starch-containing plates, suggesting that ORF2 might have different substrate specificity than predicted from the bioinformatic analysis. The BlastP results noted 76% similarity to a predicted glucosidase from *R. bromii* L2-63; therefore *orf2* might encode a glycoside hydrolase with a novel enzyme activity.



Closest Blast P match:			
ORF	**ORF1#**	**ORF2**	**ORF3#**
Accession	**CBL15128.1**	**CBL15129.1**	**CBL15002.1**
Source organism	***Ruminococcus bromii* L2-63**	***Ruminococcus bromii* L2-63**	***Ruminococcus bromii* L2-63**
Predicted function	**ABC-type maltose transporter (287 aa)**	**Glycosidase (557 aa)**	**Anthranilate synthase component I (486 aa)**
% ID/ Sim	**93/94**	**64/76**	**74/87**
Score	**280**	**758**	**457**
Match length [aa]	**151**	**567**	**286**

- truncated ORF

Figure 4.6 Schematic organization of the ORFs in the insert of clone P1E14 detected on CMC-containing plates.

The truncated ORFs are shown as open arrows. The numbers in the arrows represent the length of ORFs in amino acids. The green arrow (ORF2) is a putative glycoside hydrolase enzyme involved in carbohydrate utilisation. The red arrows are proteins involved in carbohydrate transport (ORF1) and synthesis of anthranilate (ORF3).

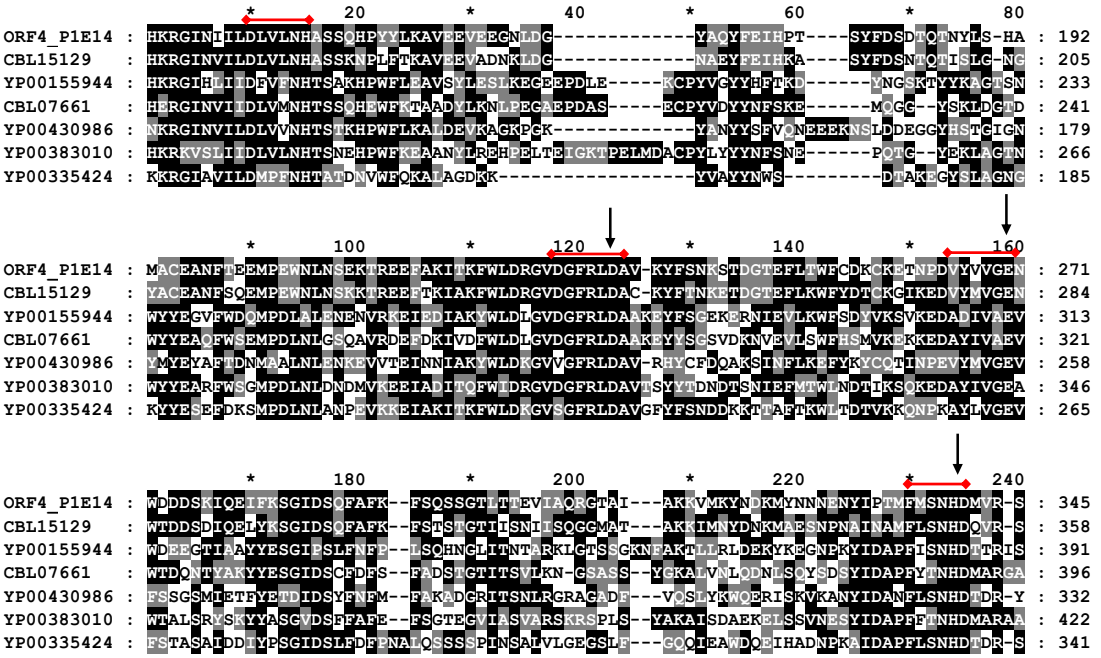


Figure 4.7 Amino acid sequences alignment of the catalytic domains of α-amylases from different bacteria.

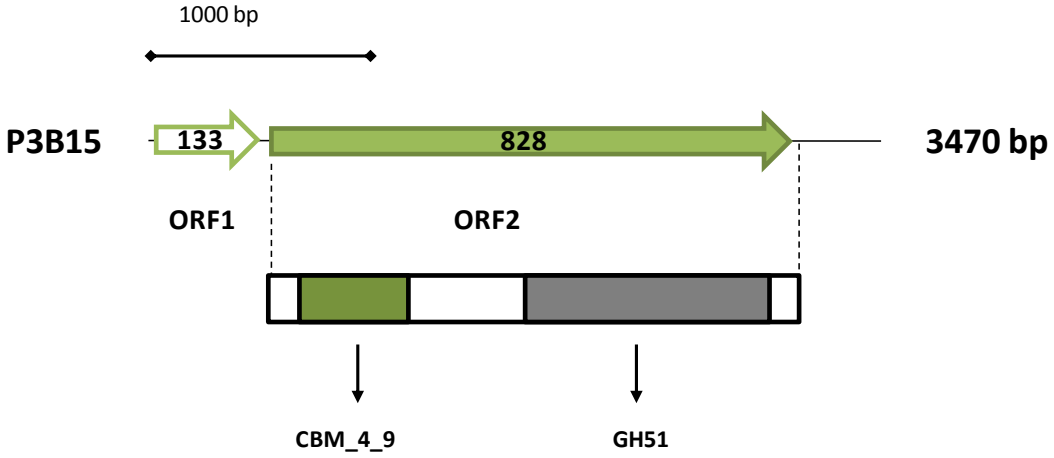
The conserved regions are shown with red bars, while catalytic active sites are marked with arrows (Oslancova and Janecek 2002). ORF4_P1E14 – ORF4 from clone P1E14; CBL15129.1 - *Ruminococcus bromii* L2-63; YP_00155944.1 - *Clostridium phytofermentans* ISDg; CBL07661.1 - *Roseburia intestinalis* M50/1; YP_004309866.1 - *Clostridium lentocellum* DSM 5427; YP_003830102.1 - *Butyrivibrio proteoclasticus* B316; YP_003354240.1 - *Lactococcus lactis* subsp. *lactis* KF147.

Clone P3B15

The arabinofuranosidase active clone P3B15 was detected using a fluorescent method. Two *orfs* were identified in the insert and they showed a very high level of identity to genes from *Roseburia intestinalis* (Figure 4.8). The truncated *orf1* showed a high degree of similarity to a putative xylosidase/ arabinosidase from *Roseburia intestinalis* L1-82. Genome information of *Roseburia intestinalis* L1-82 allowed the entire sequence of this ORF to be predicted which, based on sequence similarities, had been classified as a member of glycoside hydrolase family 43. Enzymes from this family are highly active in the breakdown of branched xylan, which requires the action of β-xylanases, β-xylosidases and α-arabinofuranosidases (Whitehead 1995, Sakka *et al.,* 1993, Suryani *et al.,* 2004). A significant number of enzymes from the GH43 family are bifunctional proteins with β-xylosidase and α-arabinofuranosidase activity, encoded by bacteria and fungi (Whitehead 1995, Sakka *et al.,* 1993).

The complete *orf2* showed similarity to a putative secreted arabinosidase from *Roseburia intestinalis* L1-82 (ZP_04743179.1). The deduced length of the ORF2 protein was 828 aa and the predicted protein showed a modular architecture. The protein contains an α-L-arabinofuranosidase C-terminal domain, which is characteristic of the enzymes belonging to glycoside hydrolase family 51. These enzymes catalyse the hydrolysis of non-reducing terminal α-L-arabinofuranosidic linkages in L-arabinose-containing polysaccharides and are found in highly active cellulolytic bacteria such as *Clostridium cellulovorans* or *Clostridium thermocellum* (Kosugi *et al.,* 2002, Taylor *et al.,* 2006). BLAST matches of ORF2 from clone P3B15 showed similarity to putative arabinofuranosidase enzymes from different bacterial species from closely related *Roseburia* strains as well as more distant species of thermophilic bacteria of the genera *Thermotoga* and *Dictoglomus* (Figure 4.10). The conserved PGG locus as well as two glutamate residues (E372 and E416), which act as an acid/base residue and enzymatic nucleophile in the active site of the enzyme (Margolles and de los Reyes-Gavilan 2003, Taylor *et al.,* 2006) were found in ORF2 of clone P3B15 (Figure 4.10). Upstream of the catalytic domain, carbohydrate binding domain CBM_4_9 was identified which was shown to bind to cellulose and xylan polymers and is linked to catalytic domains of different glycoside hydrolase families (Gaudin *et al.,* 2000, Johnson *et al.,* 1996, Kataeva *et al.,* 2001).

Analysis of the corresponding genomic region of *Roseburia intestinalis* L1-82, showed the presence of a predicted xylosidase gene. Genome information as well as experimental data shows that *Roseburia* species are highly active degraders of xylan-containing dietary fibres (Mirande *et al.*, 2010, Chassard *et al.*, 2007).



Closest Blast P match:		
ORF	**ORF1#**	**ORF2**
Accession No	**ZP_04743177.1 (465 aa)**	**ZP_04743179.1 (828 aa)**
Source organism	***Roseburia intestinalis* L1-82**	***Roseburia intestinalis* L1-82**
Predicted function	**Xylosidase/arabinosidase**	**α-arabinofuranosidase**
% ID/ Sim	**98/98**	**99/99**
Score	**268**	**1711**
Match length [aa]	**133**	**828**

- truncated ORF

Figure 4.8 Schematic overview of clone P3B15 detected on arabinofuranoside-containing plates.

The table presents the closest Blast P match. The truncated ORF1 is shown as open arrow. The numbers in the arrows represent the length of the ORFs in amino acids. The domain structure of ORF2 is shown: CBM_4_9 – carbohydrate binding module, GH51 – catalytic domain glycoside hydrolase family 51 (Pfam database). The green arrows represent genes involved in carbohydrate degradation.

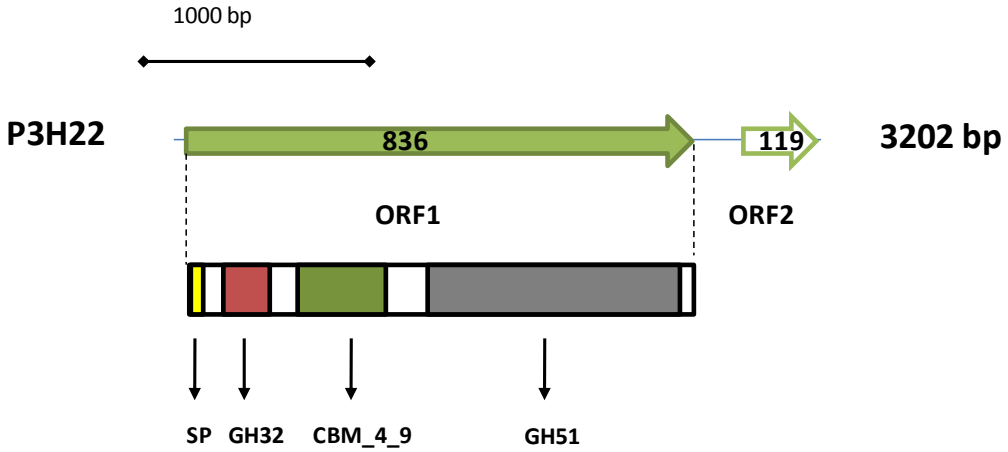
Clone P3H22

The metagenomic clone P3H22 showed the presence of an insert containing two *orfs* highly similar to genes detected in *Bacteroides plebeius* DSM 17135 (Figure 4.9). The first ORF in the insert was 836 aa long and showed a high degree of identity to a hypothetical protein BACPLE_03733 (ZP_032100421) from *Bacteroides plebeius* DSM 17135. A modular structure with several domains was detected in the deduced protein encoded by ORF1. At the N-terminus a predicted signal peptide region was noted with a cleavage site between alanine and histidine. A glycosyl hydrolase family 32 domain with potential β- fructosidase activity was identified. The β-fructosidase superfamily includes enzymes involved in hydrolysis of nonreducing β-D-fructosidic bonds present in sucrose, fructooligosaccharides (FOS) and inulin (Scott *et al.,* 2011, Alberto *et al.,* 2004). The human diet is rich in fructose-containing carbohydrates; however the human genome lacks genes that encode enzymes that cleave fructosidic bonds. The genomes of known gut bacteria such as *Bifidobacterium longum, Bifidobacterium lactis* and *Bifidobacterium breve* (Russell *et al.,* 2011) and *Bacteroides thetaiotaomicron* (Xu *et al.,* 2003) encode fructosidases which enable degradation of fructose-containing sugars. The GH32 domain is followed by a CBM_4_9 domain featuring binding properties to polysaccharides such as cellulose and xylan. The C-terminal domain showed similarity to α-arabinofuranosidases from family GH51 with a conserved PGG locus and glutamate residues (E562 and E619) in the catalytic site of the enzyme (Figure 4.10). The BlastP results using the amino acid sequence of ORF1 showed that the protein is well conserved across different species of *Bacteroides* (Figure 4.10). A similar multi-domain protein is also found in other species of the Bacteroidetes phylum – *Prevotella bergenis* (ZP_06006743.2), *Paraprevotella xylaniphila* (ZP_08319121.1) and *Dysgonomonas gadei* ATCC BAA-286 (ZP_08472155.1).

The second *orf* in clone P3H22 is a truncated gene with homology to a predicted β-galactosidase (BACPLE_03734) from *Bacteroides plebeius* DSM 17135. The deduced length of the entire ORF based on the genome information was 783 aa. The predicted protein had a multi-domain structure with glycoside hydrolase family 35 (PF01301) and discoidin domains (Cheng *et al.,* 2009). ORF2 showed a similarity to the N-terminal end of BACPLE_03734 protein (ZP_032100431) with partial

sequence of the GH35 domain. The GH35 family includes enzymes that display β-galactosidase and exo-β-glucosaminidase activities (Gamauf *et al.,* 2007). The discoidin domain (DS), also known as F5/F8 type C domain, has been found in many eukaryotic and prokaryotic proteins. The domain binds a variety of molecules including phospholipids and carbohydrates required for nutrient assimilation and cellular adhesion. A variety of bacterial glycoside hydrolase enzymes contain DS domains, which were found to bind insoluble polysaccharides including lichenan, cellulose and chitin (Cheng *et al.*, 2009).

Clone P3H22 showed weak activity on starch-containing plates. The sequence analysis does not confirm the presence of amylase-encoding genes. The domains detected in ORF1 belong to GH enzymes with unknown activity on starch. Weak arabinofuranosidase activity was detected for clone P3H22 when tested on fluorescent substrate-containing plates. The architectural pattern of ORF1 from clone P3H22 was similar to ORF1 from clone P3B15. They share 33% identity in their amino acid sequence over whole length and bear CBM 4_9 and C-terminal α-arabinofuranoside domains.



Closest Blast P match:		
ORF	**ORF1**	**ORF2#**
Accession No	**ZP_03210042.1 (836 aa)**	**ZP_03210043.1 (733 aa)**
Source organism	***Bacteroides plebeius* DSM17135**	***Bacteroides plebeius* DSM17135**
Predicted function	**α-arabinofuranosidase**	**β-galactosidase**
% ID/Sim	**96/98**	**95/96**
Score	**1674**	**244**
Match length [aa]	**830**	**119**

- truncated ORF

Figure 4.9 Schematic overview of the clone P3H22 detected on starch-containing plates.

The table presents the closest Blast P match. The truncated ORF2 is shown as open arrow. The numbers in the arrows represent the length of the ORFs in amino acids. The domain structure of ORF1 is shown: SP – signal peptide, GH32 – catalytic domain glycoside hydrolase family 32, CBM_4_9 – carbohydrate binding module and GH51 – catalytic domain glycoside hydrolase family 51. The green arrows represent genes involved in carbohydrate metabolism.

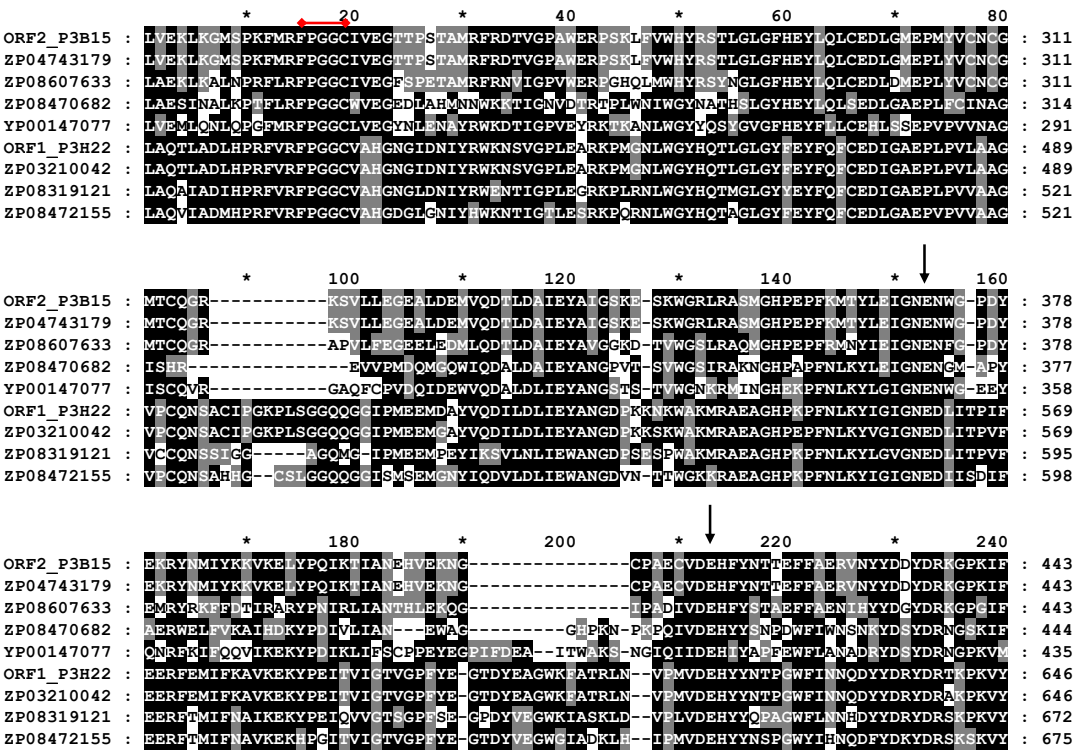


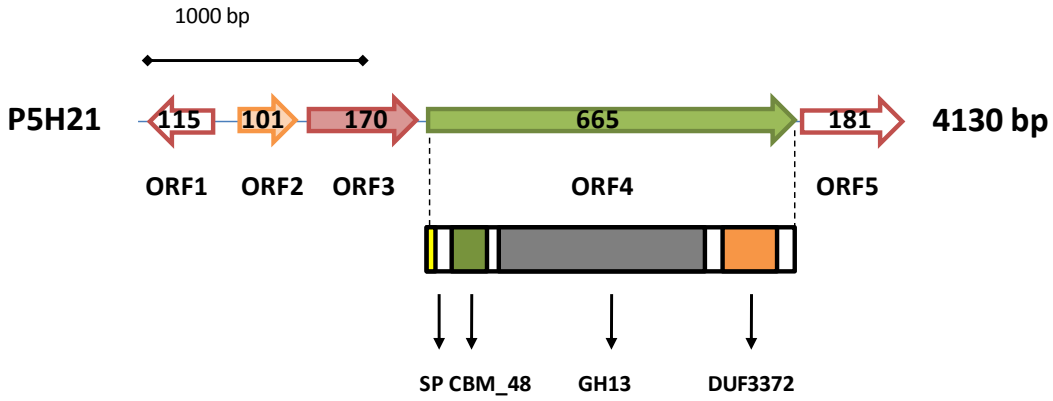
Figure 4.10 Amino acid sequence alignment of α-L-arabinofuranosidase partial catalytic domains from different bacteria.

The sequences represent BlastP results from the search using protein sequence of ORF2 from clone P3B15 and ORF1 from clone P3H22 as queries. The conserved motif characteristic to arabinofuranosidase enzymes is shown with red bar; two arrows indicate conserved glutamate residues. ORF2_P3B15 – ORF2 from clone P3B15; ORF1_P3H22 – ORF1 from clone P3H22; ZP_04743179.1 - *Roseburia intestinalis* L1-82; ZP_08607633.1 - *Lachnospiraceae bacterium* 3_1_57FAA_CT1; ZP_08470682.1 - *Dysgonomonas mossii* DSM 22836; YP_001470776.1 - *Thermotoga lettingae* TMO; ZP_03210042.1 – *Bacteroides plebeius* DSM 17135; ZP_08319121.1 - *Paraprevotella xylaniphila* and ZP_08472155.1 - *Dysgonomonas gadei* ATCC BAA-286.

Clone P5H21

The insert of clone P5H21 includes five *orfs* with high sequence identity to genes from *Bacteroides uniformis* ATCC 8492 (Figure 4.11). A similar gene organisation was present in genomes of other *Bacteroides* strains including *Bacteroides* sp. D20 and *Bacteroides* sp. 4_1_36. The genes of insert from clone P5H21 encodes following putative proteins: an urea transporter (ORF1), a hypothetical protein BACUNI_ 01236 (ORF2), a predicted endodeoxyribonuclease RuvC protein (ORF3), a pullulanase type I protein (ORF4) and a putative ABC transporter (ORF5). Pullulanases are starch de-branching enzymes that hydrolyse the α-1,6-glucosidic linkage of α-glucans: pullulan, amylopectin, and glycogen. ORF4 is highly related to a type I pullulanase from *Bacteroides uniformis* ATCC8492 and showed a multi-domain structure. The N-terminus of ORF4 contains a signal peptide, followed by a carbohydrate binding module CBM_48, α-amylase catalytic domain (PF00128) and C-terminal domain of unknown function DUF3372. The carbohydrate binding module family 48 (PF02922) was found in de-branching enzymes such as isoamylases and pullulanases that metabolize polysaccharides with α-1,6-linked glucose residues. The CBM_48 domain enables binding to carbohydrate and is associated to GH13 modules (Christiansen *et al.,* 2009). The CBM_48 domain was reported as part of cellulose synthase enzymes from red algae involved in the synthesis of cellulose microfibrils. The involvement in cellulose binding was not confirmed (Matthews *et al.,* 2010; Christiansen *et al.,* 2009). The domain was also found in β-units of AMP-activated kinases which regulate carbohydrate metabolism in eukaryotes (Matthews *et al.,* 2010). The α-amylase catalytic domain contained four conserved regions found in GH13 family enzymes and the consensus sequence YNWGYD characteristic of pullulanase type I enzymes (Mikami *et al.*, 2006). Three important catalytic residues, D349, E378 and D478, were also identified (Figure 4.12) (Oslancova and Janecek 2002). The C-terminal DUF3372 domain is not functionally characterised (PFAM 11852) but its presence was detected in a putative bacterial pullulanases and oligo-α 1,6–glucosidases. The modular architecture of ORF4 was found to be characteristic of pullulanases from *Bacteroides*, *Prevotella* and *Lactobacillus* (based on BlastP results). The amino acid sequence of several related enzymes from different bacterial genera of gut microbiota also showed the

presence of conserved regions and residues (Figure 4.12). Clone P5H21 was selected on CMC-containing plates showing weak CMCase activity. Analysis of ORFs in the insert of clone P5H21 does not provide a clear explanation of the CMCase activity found. The predicted pullulanase protein encoded by ORF4 belongs to well characterised GH13 family with the ability to hydrolyze α-1,4 and α-1,6 glycosidic bonds, whereas cellulose is a polymer with β-1,4 glucose unit. The clone was tested on starch-containing plates and showed weak activity confirming the sequence based prediction for ORF4.



Closest Blast P match:					
ORF	ORF1#	ORF2	ORF3	ORF4	ORF5#
Accession	ZP_02069819.1	ZP_02069820.1	ZP_02069821.1	ZP_02069822.1	ZP_02069823.1
Source organism	*Bacteroides uniformis* ATCC8492	*Bacteroides uniformis* ATCC8492	*Bacteroides uniformis* ATCC8492	*Bacteroides uniformis* ATCC8492	*Bacteroides uniformis* ATCC8492
Predicted function	Urea transporter (295 aa)	Hypothetical protein (101 aa)	Endodeoxyribo-nuclease (189 aa)	Pullulanase type I (665 aa)	ABC-type transporter (491 aa)
% ID/ Sim	94/96	99/100	100/100	99/100	99/99
Score	209	207	352	1385	363
Match length [aa]	115	100	170	665	181

- truncated ORF

Figure 4.11 Schematic overview of the insert from clone P5H21 detected on CMC-containing plates.

The truncated ORFs are shown as open arrows. The numbers in the arrows represent the length of ORFs in amino acids. The green arrow is a putative glycoside hydrolase enzyme involved in carbohydrate utilisation. The red arrows are putative proteins not involved in carbohydrate degradation. The orange arrow represents a hypothetical protein. The domain structure of ORF4 is shown: SP – signal peptide, CBM_48: carbohydrate binding module family 48; GH13 - catalytic domain from glycoside hydrolase family 13, DUF3372 – domain of unknown function.

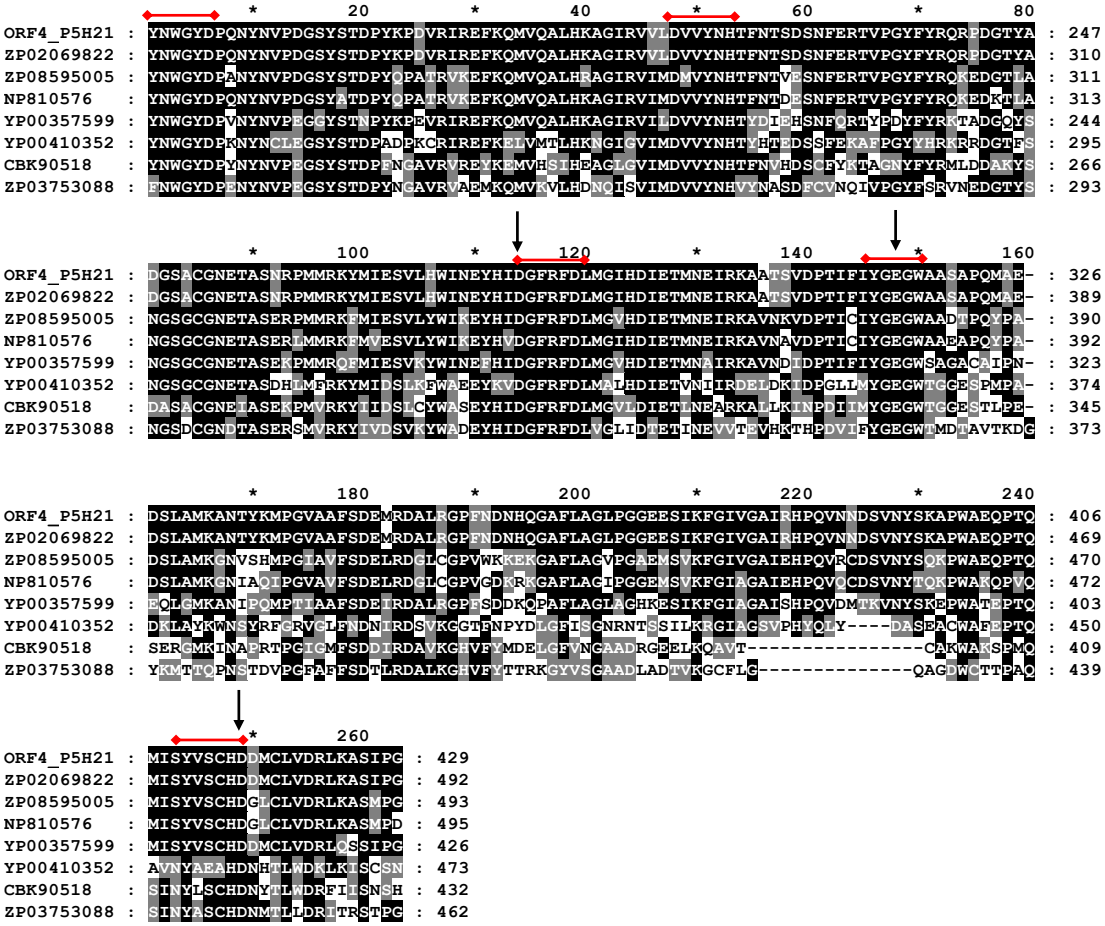


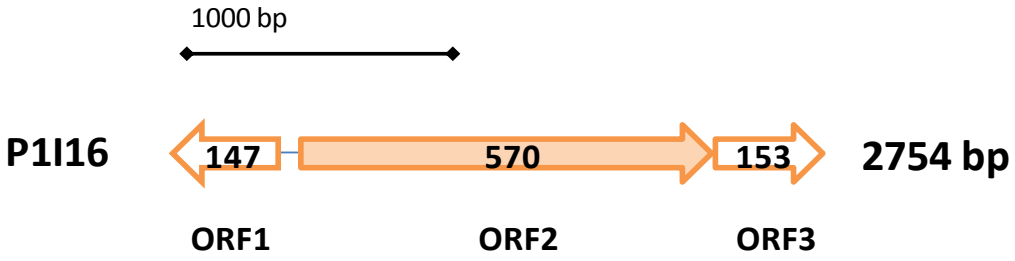
Figure 4.12 Amino acid sequence alignment of the catalytic domains of α-amylase from different bacteria.

The conserved regions are shown with red bars, while catalytic active sites are marked with arrows (Oslancova and Janecek 2002). ORF4_P5H21 – ORF4 from clones P5H21; ZP02069822 – *Bacteroides uniformis* ATCC 8492; ZP_08595005 - *Bacteroides ovatus* 3_8_47FAA; NP810576 - *Bacteroides thetaiotaomicron* VPI-5482; YP_003575993.1 – *Prevotella ruminocola* 23; YP_00410352 - *Ruminococcus albus* 7; CBK90518 - *Eubacterium rectale* DSM 17629; ZP_03753088 - *Roseburia inulinivorans* DSM 16841.

4.5.2 Analysis of clones encoding hypothetical proteins – potentially novel GH enzymes – category II

Clone P1I16

Clone P1I16 was detected initially on starch-containing plates showing moderate enzymatic activity. It was also tested on CMC-containing plates resulting in moderate enzyme activity. Insert sequencing revealed the presence of one *orf* encoding a hypothetical protein and two truncated *orfs* (Figure 4.13). The result of BLASTP search for all ORFs showed a low degree of similarity to database entries. ORF2 showed a match to a hypothetical protein from *Faecalibacterium prausnitzii* SL3/3, with a sequence identity of 38%. ORF2 showed the presence of a signal peptide with a predicted cleavage site between alanine and glutamate but no conserved domains were detected within its structure. The BlastP results also reported a partial match with a low degree of similarity (28%) to a cellulosome enzyme with dockerin domain from *Clostridium thermocellum* strains (ZP_06249385.1, YP_001038220.1). The dockerin is a protein domain associated with endoglucanase catalytic domains essential in forming the cellulosome complex involved in cellulose degradation (Bayer *et al.,* 2008). The tandem repeats of calcium binding residues characteristic to dockerin domain were not recognised in the sequence of ORF2, therefore it is unlikely that ORF2 encodes dockerin-bearing enzyme. ORF3 was a truncated gene with similarities to a putative β-1,4-xylanase from *Ruminococcus* sp. 18P13. The deduced length of this protein is 633 aa and several conserved domains were predicted. At the N-terminus a CBM4_9 domain was identified followed by a glycoside hydrolase family 10 domain involved in xylan utilisation. A similar multi-domain architecture was described for *xynB* gene of *Clostridium cellulovorans* (Han *et al.,* 2004). The degree of similarity to hypothetical proteins or predicted proteins was low for the ORFs detected in this insert. Hence, it seems likely that the insert originated from a novel bacterium of the human gut. The partial similarity to a cellulosomal endoglucanase serves as a good indicator for a potentially novel glycoside hydrolase enzyme.



Closest Blast P match:			
ORF	**ORF1#**	**ORF2**	**ORF3#**
Accession No	**No hit**	**CBL01310.1**	**CBL24798.1**
Source organism	**-**	***Faecalibacterium prausnitzii* SL3/3**	***Ruminococcus obeum* A2-162**
Predicted function	**Hypothetical protein**	**Hypothetical protein (563 aa)**	**Hypothetical protein (613 aa)**
% ID/Sim	**-**	**38/48**	**32/45**
Score	**-**	**334**	**56.6**
Match length [aa]	**-**	**564**	**148**

- truncated ORF

Figure 4.13 Schematic overview of the insert from clone P1I16 detected on starch- and CMC-containing plates.

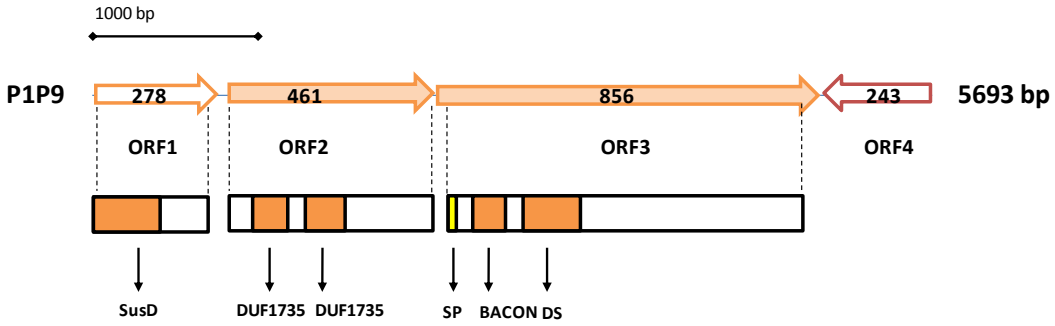
The truncated ORFs are shown as open arrows. The numbers in the arrows represent the length of the ORF in amino acids. The orange arrows represent ORFs encoding hypothetical proteins.

Clone P1P9

Clone P1P9 contained an insert encoding four ORFs with a variable degree of identity to hypothetical proteins from *Bacteroides caccae* ATCC43185 and *Bacteroides plebeius* DSM17135 (Figure 4.14). The first two ORFs showed similarity to a hypothetical protein from *Bacteroides caccae* ATCC43185. The ORF1 is a truncated gene whose deduced protein length based on *B. caccae* ATCC43185 genome information is 643 aa. The C-terminal end of this protein contains a SusD family domain (PF07980) present in a significant number of putative uncharacterized proteins (2898 sequences in the PFAM database showed the same domain structure). This domain is found in the SusD protein, which is encoded in the SUS operon of *B. thetaiotaomicron.* The operon consists of genes that code for proteins involved in starch binding (SusC-F) and hydrolysis (SusA, SusB and SusG). SusD, in association with SusC, is involved in substrate binding (Cho and Salyers 2001, Bakolitsa *et al.*, 2010). The insert of clone P1P9 does not show the same organisation as the SUS operon. However, sus-like clusters have been identified in the genome of *B. thetaiotaomicron* which contains 57 paralogues of the *susD* gene. Many of these clusters are linked to genes which are probably involved in utilisation of different carbohydrates (Xu *et al.*, 2007). ORF2 shared identity to a hypothetical protein from *B. caccae* ATCC43185, which contains two domains of unknown function - DUF 1735 (PF08522). The tandem dyad DUF1735 domains were found in predicted GH43 enzymes from *Bacteroides* species. The GH43 family includes mostly enzymes with α-L-arabinofuranosidase (EC 3.2.1.55), endo-α-L-arabinanase (EC 3.2.1.99) and xylolytic activities (Gilbert *et al.*, 2008). The catalytic domain for GH43 family was not identified in ORF2. The sequence of ORF3 has high identity to a hypothetical protein from *B. plebeius* DSM17135. Analysis of the ORFs showed the presence of an N-terminal lipoprotein signal peptide with a predicted cleavage site between glycine and cysteine. The gene carries a BACON domain which stands for Bacteroides-Associated Carbohydrate-binding Often N-terminal domain (PF13004) and discoidin (DS) domain (PF00754). The BACON domain was proposed to be involved in mucin binding in *Bacteroides* species based on bioinformatics data, but experimental data are needed to confirm these assumptions (Mello *et al.*, 2010). The BACON domain can be attached to glycoside hydrolase catalytic domains from families 5 (cellulases), 13 (amylases), 16 (laminarinases) and

42 (β-galactosidases). The DS domain was found to bind a variety of carbohydrates and is followed by the catalytic domains from different GH enzymes (Cheng *et al.*, 2009). The C-terminal part of the protein encoded by ORF3 does not show the presence of any predicted domains belonging to GH families. ORF4 showed a high degree of identity to cysteinyl-tRNA synthetase (CysRS) which catalyzes the transfer of cysteine to the 3'-end of its tRNA.

Clone P1P9 was initially detected on starch-containing plates and showed moderate enzymatic activity. It was also plated on CMC-containing plates and showed weak carboxymethyl cellulase activity. Bioinformatic analysis of predicted ORFs does not recognise domains or catalytic sites for known GH enzymes. Hence, there is the possibility that the insert of clone P1P9 encodes a potentially novel glycoside hydrolase. Based on the sequence analysis and functional screening, it was hypothesised that protein encoded by ORF3 could be a novel GH enzyme. The possibility that the protein encoded by ORF2 is a novel carbohydrate active enzyme should also be considered, since a number of glycoside hydrolase enzymes from *Bacteroides* bear DUF1735 domain.



Closest Blast P match:				
ORF	ORF1#	ORF2	ORF3	ORF4#
Accession	ZP_01960231.1	ZP_01959758.1	ZP_03207951.1	ZP_03207952.1
Source organism	*Bacteroides caccae* ATCC43185	*Bacteroides caccae* ATCC43185	*Bacteroides plebeius* DSM17135	*Bacteroides plebeius* DSM17135
Predicted function	Hypothetical protein (643 aa)	Hypothetical protein (445 aa)	Hypothetical protein (856 aa)	Cysteinyl-tRNA synthetase (493 aa)
% ID/Sim	57/74	30/49	94/96	99/99
Score	332	223	1687	432
Match length [aa]	281	453	850	226

- truncated ORF

Figure 4.14 Schematic overview of the insert from clone P1P9 detected on starch- and CMC-containing plates.

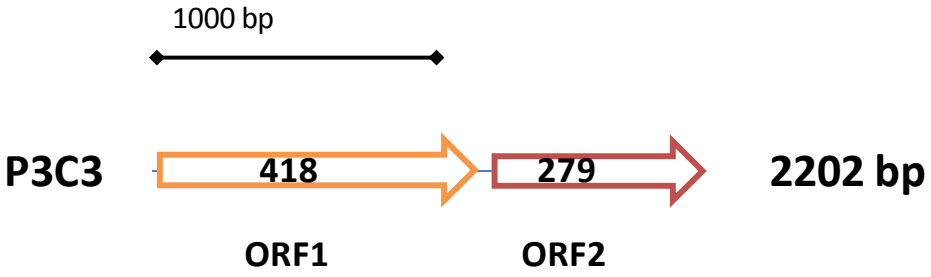
The truncated ORFs are shown as open arrows. The numbers in the arrows represent the length of the ORF in amino acids. The orange arrows represent ORFs encoding hypothetical proteins. The red arrow represents a protein unlikely to be involved in carbohydrate metabolism. The domain architecture is shown for the hypothetical protein encoded by ORF1, ORF2 and ORF3.

Clone P3C3

Clone P3C3 was selected on CMC-containing plates and showed moderate enzymatic activity. It also showed weak amylase activity when tested on starch. Bioinformatic analysis of the insert sequence detected two ORFs (Figure 4.15) showing partial matches to a putative lipoprotein from *Bryantella formatexignes* and to a putative surface protein Pls (plasmin-sensitive protein) from *Streptococcus agalactiae* COH1, respectively.

ORF1 showed the presence of a domain (PF01522) which was found in a polysaccharide deacetylase family enzyme (Millward-Sadler *et al.,* 1994). The domain was reported in a nodulation protein (NodB) from *Rhizobium* as being involved in synthesis of host specific acylated chito-oligosaccharide factors which initiate nodule formation and allow rhizobia to enter the plant. This family of polysaccharide deacetylases also includes chitin deacetylase from yeast and endoxylanases which hydrolyse glycosidic bonds in xylan. The domain was found in xylanase XylD from *Cellulomonas fimi* and since XylD did not catalyse deacetylation of chitobiose, it was suggested that the NodB-like sequence in this xylanase originated from *Rhizobium* and was transferred to *C. fimi* (Millward-Sadler *et al.,* 1994). The polysaccharide deacetylase domain detected in ORF1 was also found adjacent to the catalytic domain of cellulases from GH family 9 and xylanases from GH10 and GH11 families. The genomes of gut-associated *Bacteroides* species contain many genes encoding polysaccharide deacetylases. The host epithelial glycans contains O-acetylated sugars such as sialic acid, which protects them from cleavage by bacterial glycoside hydrolases. The production of deacetylases by gut bacteria modifies epithelial glycan making it more available for bacterial degradation (Xu *et al.,* 2007). It is possible that ORF1 bears a potentially novel GH catalytic domain with glucanase activity. The BlastP results showed that further matches to ORF1 were annotated as putative hydrolases, however experimental confirmation is needed to test this hypothesis. ORF2 showed the presence of a highly TS-rich region and partial similarity to a surface protein Pls. The Pls protein is crucial for bacterial colonisation and survival in the host as well as invasion of epithelial cells. This serine rich high molecular surface protein was found mainly in bacteria such as *Staphylococcus* and *Streptococcus* (Bensing and Sullam 2010).

The BLAST results showed moderate similarity to predicted proteins but did not detect conserved domains involved in enzymatic carbohydrate utilisation and which are characteristic to GH enzymes. The results could possibly indicate that ORF1 may be involved in degradation of carbohydrates, and then ORF2 may be involved in its binding but based on the presented results further investigation is needed.



Closest Blast P match:		
ORF	**ORF1#**	**ORF2#**
Accession No	**ZP_05344538.3**	**ZP_00784981.1**
Source organism	***Bryantella formatexigens* DSM 14469**	***Streptococcus agalactiae* COH1**
Predicted function	**putative lipoprotein (650 aa)**	**Surface protein (243 aa)**
% ID/Sim	**72/82**	**40/63**
Score	**645**	**79.7**
Match length [aa]	**419**	**188**

- truncated ORF

Figure 4.15 Schematic overview of the insert from clone P3C3 detected on CMC- and starch-containing plates.

The truncated ORFs are shown as open arrows. The numbers in the arrows represent the length of the ORF in amino acids. The orange arrow represents an ORF potentially encoding a potentially novel GH protein. The red arrow is an ORF encoding a protein likely to be involved in carbohydrate binding.

Clone P5D24

In clone P5D24, several ORFs were detected that showed a high level of sequence identity to predicted proteins from *Alistipes putredinis* DSM 17216 (Figure 4.16). ORF1 encodes a predicted membrane protein with six transmembrane regions and a DUF368 domain identified during Blast search. ORF2 encodes a putative shikimate dehydrogenase enzyme involved in the shikimate pathway which links carbohydrate metabolism and aromatic amino acid biosynthesis in microorganisms and plants. ORF3 and ORF4 are predicted as hypothetical proteins with domains of unknown function (PF09413). ORF5 showed a high degree of identity to a putative DNA polymerase III subunit alpha involved in replicative synthesis in bacteria. The only candidate for a potential involvement in carbohydrate utilisation in clone P5D24 is encoded by ORF1. However, the results derived from *in silico* analysis provide insufficient information on the role of this hypothetical transmembrane protein.

Closest Blast P match:					
ORF	ORF1#	ORF2	ORF3	ORF4	ORF5#
Accession	ZP_02425823.1	ZP_02425822.1	ZP_02425821.1	ZP_02425820.1	ZP_02425819.1
Source organism	*Alistipes putredinis* DSM 17216	*Alistipes putredinis* DSM 17216	*Alistipes putredinis* DSM 17216	*Alistipes putredinis* DSM 17216	*Alistipes putredinis* DSM 17216
Predicted function	membrane protein (314 aa)	shikimate dehydrogenase (244 aa)	hypothetical protein (72aa)	hypothetical protein (55 aa)	polymerase III subunit alpha (1287 aa)
% ID/Sim	98/99	99/99	100/100	100/100	99/99
Score	547	503	146	115	2004
Match length [aa]	278	244	72	55	962

- truncated ORF

Figure 4.16 Schematic overview of the insert from clone P5D24 detected on starch- and CMC-containing plates.

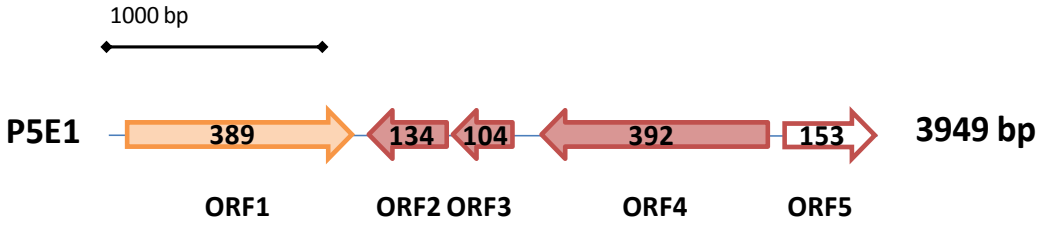
The truncated ORFs are shown as open arrow. The numbers in the arrows represent the length of the ORF in amino acids. The orange arrows represent ORFs encoding hypothetical proteins or genes possibly involved in carbohydrate metabolism. The red arrows are ORFs encoding proteins unlikely to be involved in carbohydrate degradation.

Clone P5E1

Clone P5E1 was detected on starch and CMC-containing plates and showed weak but detectable enzyme activity. Five ORFs encoded by the insert with a high degree of identity to predicted genes from *Bacteroides* sp. 9_1_42FAA were detected (Figure 4.17). A similar genomic organisation was also found in other *Bacteroides* species including *B. vulgatus* and *B. dorei.* ORF1 showed similarity to a conserved hypothetical protein (ZP06086514.1) from *Bacteroides* sp. 9_1_42_AA. A signal peptide at the N-terminal end and outer membrane domain at the C-terminal end was present in this protein. The outer membrane C-terminal domain represents a predicted transmembrane β-8,10-barrel found in outer membrane proteins such as OmpA, OmpX and NspA, and in the outer membrane enzyme PagP. Outer membrane proteins are responsible for structural integrity of the bacterial cell and transport of various molecules. They are also receptors for colicins and phages during conjugation (Papamichail and Delihas 2006).

ORF2 and ORF3 showed similarity to a predicted HEPN (higher eukaryotes and prokaryotes nucleotide-binding) domain-containing protein and a nucleotidyl-transferase (NT), respectively. The well conserved complex HEPN-NT probably plays a role in bacterial antibiotic resistance by transferring a nucleotidyl group to drugs or other toxic substances (Grynberg *et al.,* 2003). ORF4 was predicted as a Na^+/H^+-dicarboxylate symporter (PF00375), which transports a variety of dicarboxylates and inorganic anions across the membrane using energy derived from the movement of sodium ions (Strickler *et al.,* 2009). *Orf5* was a truncated gene encoding a putative 6-phosphogluconate dehydrogenase. The enzyme is involved in conversion of 6-phosphogluconate to ribulose-5-phosphate, which is part of the pentose phosphate pathway that generates NADPH and pentoses. The insert shows the presence of several predicted proteins unlikely to be involved in carbohydrate degradation. It seems possible that ORF1, which showed a presence of an outer membrane domain, may encode a novel catalytic domain involved in carbohydrate metabolism. SusG which is a part of *sus*-gene cluster is an outer membrane protein with hydrolytic activity (Martens *et al.,* 2009). In addition, several studies identified outer membrane proteins as candidates for involvement in cellulose adhesion and utilisation. Kim *et al*. (2011a) described cellulase/ endoglucanase activity of a metagenomic clone encoding a putative outer membrane protein from *Vibrio*

alginolyticus 12G01. Outer membrane proteins were also predicted to be involved in cellulose binding and utilisation by the soil bacterium *Cytophaga hutchinsonii* (Xie *et al.*, 2007). This assumption was consistent with the observation made by Jun *et al.* (2007) who showed that different outer membrane proteins in *Fibrobacter succinogenes* were required for cellulose binding and degradation.



Closest Blast P match:					
ORF	ORF1	ORF2	ORF3	ORF4	ORF5#
Accession	ZP_06086514.1	ZP_06086513.1	ZP_06086512.1	ZP_06086511.1	ZP_06086510.1
Source organism	*Bacteroides* sp. 3_1_33FAA	*Bacteroides* sp. 3_1_33FAA	*Bacteroides* sp. 3_1_33FAA	*Bacteroides* sp. 3_1_33FAA	*Bacteroides* sp. 3_1_33FAA
Predicted function	Hypothetical protein (389 aa)	HEPN domain-containing protein (134 aa)	DNA polymerase (104 aa)	Na^+/H^+-dicarboxylate symporter (392 aa)	6-phosphogluconate dehydrogenase (478 aa)
% ID/Sim	100/100	100/100	100/100	99/100	98/99
Score	811	278	208	766	311
Match length [aa]	389	134	104	392	153

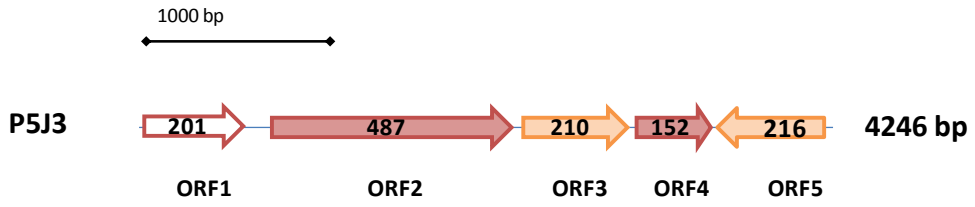
- truncated ORF

Figure 4.17 Schematic overview of the insert from clone P5E1 detected on starch- and CMC-containing plates.

The truncated ORF is shown as open arrow. The numbers in the arrows represent the length of the ORF in amino acids. The orange arrow represents an ORF encoding a hypothetical protein. The red arrows are ORFs encoding proteins unlikely to be involved in carbohydrate degradation.

Clone P5J3

Clone P5J3 was detected on starch and CMC-containing plates and showed moderate enzyme activity. Bioinformatics analysis predicted the presence of five *orfs* which encode proteins with similarities to putative proteins or hypothetical proteins (Figure 4.18). ORF1 and ORF2 represent putative proteins of a two-component signal transduction system, which are involved in regulation of processes in response to changes in the environment. The first component is a two-domain response regulator protein (PF00072) (Casino *et al.*, 2010). The second component of this system is a sensor histidine kinase involved in auto-phosphorylation and transfer of the phosphoryl group to the signal receiver domain on ORF1. These two ORFs are not likely to be of direct relevance to carbohydrate metabolism. ORF3 showed regions of low complexity sequence rich in arginine and histidine. The BlastP results showed no matches. ORF4 is a predicted peptidase with a periplasmic serine protease domain (PF00089) which is found in many bacterial enzymes. These proteins protect bacteria from thermal and other stresses and may be important for the survival of bacterial pathogens. The protein encoded by *orf5* showed similarity to a hypothetical protein but no putative conserved domains were detected.



Closest Blast P match:					
ORF	ORF1#	ORF2	ORF3	ORF4	ORF5
Accession	ZP_08420204.1	ZP_08420205.1	-	ZP_06144018.1	ZP_02038441.1
Source organism	*Ruminococcaceae bacterium* D16	*Ruminococcaceae bacterium* D16	-	*Ruminococcus flavefaciens* FD-1	*Bacteroides capillosus* ATCC 9799
Predicted function	DNA-binding response regulator (233 aa)	sensory transduction histidine kinase (411 aa)	Hypothetical protein	peptidase S1 and S6 chymotrypsin (466 aa)	hypothetical protein (130 aa)
% ID/Sim	74/85	46/66	-	50/70	44/60
Score	307	411	-	107	91.7
Match length [aa]	200	467	-	119	122

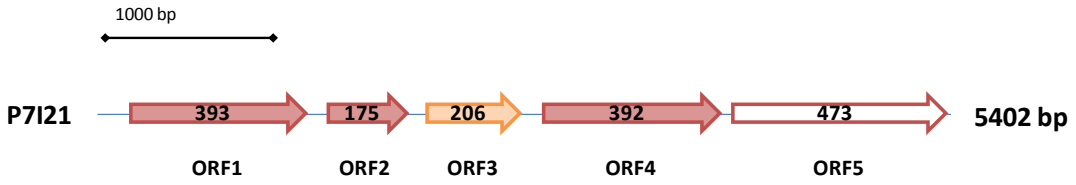
- truncated ORF

Figure 4.18 Schematic overview of the insert from clone P5J3 detected on starch- and CMC-containing plates.

The truncated ORF1 is shown as open arrow. The number in the arrow represents the length of the ORF in amino acids. The orange arrows represent ORFs encoding hypothetical proteins. The red arrows are ORFs encoding proteins not expected to be involved in carbohydrate degradation.

Clone P7I21

Clone P7I21 was detected on starch-containing plates and showed weak amylase activity. Several *orfs* with a low degree of identity to putative proteins from cultured microorganisms were detected (Figure 4.19). For ORF1, 3, 4 and 5, no putative conserved domains were detected and the level of similarity to cultured microorganisms was very low. ORF2 showed the presence of a signal peptidase type I domain. The domain is found in membrane bound serine proteases that cleave the amino-terminal signal peptide extension from proteins and translocate them across biological membranes. No relevance to carbohydrate metabolism can be linked to the ORFs predicted in insert of clone P7I21; however the weak amylase activity could indicate the presence of novel glycoside hydrolase.



Closest Blast P match:					
ORF	ORF1	ORF2	ORF3	ORF4	ORF5#
Accession	ZP_01254855.1	ZP_04299582.1	YP_003276898.1	ZP_07642083.1	YP_003250214.1
Bacterial origin	*Psychroflexus torquis* ATCC 700755	*Bacillus cereus* MM3	*Comamonas testosteroni* CNB-2	*Streptococcus mitis* SK597	*Fibrobacter succinogenes* S85
Predicted function	putative phosphoglucomutase (571 aa)	signal peptidase (189aa)	hypothetical membrane protein (299 aa)	malonyl CoA-acyl transacylase (318 aa)	GLUG domain protein (1994 aa)
% ID/Sim	27/49	33/51	24/44	29/50	36/55
Score	38.5	62	36.6	41.2	53.5
Match length [aa]	130	174	100	103	121

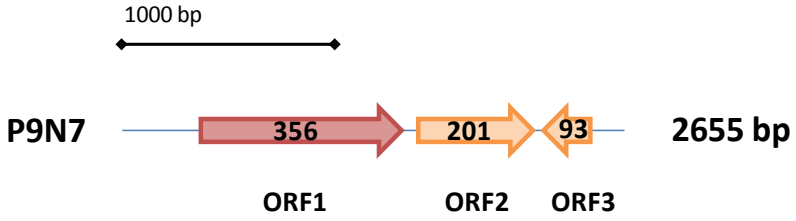
- truncated ORF

Figure 4.19 Schematic overview of the insert from clone P7I21 detected on starch-containing plates.

The open arrow represents truncated ORF. The numbers in the arrows represent the length of the ORFs in amino acids. The orange arrows represent ORFs encoding hypothetical proteins. The red arrows are ORFs encoding predicted proteins with insufficient information to assess a likelihood of homology to enzymes involved in carbohydrate degradation

Clone P9N7

Clone P9N7 harboured three ORFs with a degree of identity to predicted proteins from various microbial sources (Figure 4.20). ORF1 has similarities to a predicted site recombinase with a domain characteristic to DNA breaking-rejoining enzymes. ORF2 and ORF3 showed identity to hypothetical proteins with no conserved domains being detected. Clone P9N7 was selected on CMC-, xylan- and lichenan-containing plates and showed high enzyme activity. This could indicate that the insert encodes a novel and highly potent enzyme with different activities. The sequencing data do not predict domains which might be involved in carbohydrate degradation; therefore further investigation is needed to confirm the novelty of the insert.



Closest Blast P match:			
ORF	ORF1	ORF2	ORF3
Accession No	ZP_08514714.1	ZP_07918052.1	ZP_03643004.1
Source organism	*Allistipes* sp. HGB5	*Bacteroides* sp. D2	*Bacteroides coprophilus* DSM18228
Predicted function	Site recombinase (394 aa)	Hypothetical protein (755 aa)	Hypothetical protein (89 aa)
% ID/Sim	56/71	37/53	54/69
Score	316	92.8	97.4
Match length [aa]	335	201	89

Figure 4.20 Schematic overview of the insert from clone P9N7 detected on CMC-, xylan- and lichenan-containing plates.

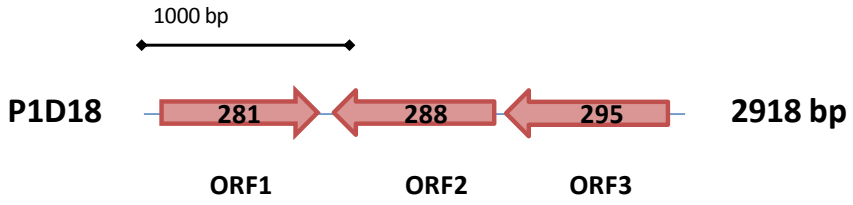
The numbers in the arrows represent the length of the ORF in amino acids. The orange arrows represent ORFs encoding hypothetical proteins. The red arrows are ORFs encoding proteins unlikely to be involved in carbohydrate degradation.

4.5.3 Clones encoding genes unrelated to glycoside hydrolase enzymes – category III

A number of clones which were selected during functional screening of the metagenome library showed a presence of genes encoding putative proteins with functions not directly related to carbohydrate hydrolysis. Therefore they are likely to be false-positive clones.

Clone P1D18

The insert of clone P1D18 encoded three genes (Figure 4.21). ORF1 and ORF2 shared a high a degree of identity to putative transcriptional regulators with helix-turn-helix domains. Both ORFs contained a domain specific to RpiR (ribose phosphate isomerase regulator). The RpiR regulates gene expression of protein RpiB (ribose phosphate isomerase), which interconverts ribose phosphate to the ribulose phosphate derivative during the non- oxidative phase of the pentose phosphate pathway (Sorensen and Hove-Jensen 1996). Downstream of the RpiR proteins, ORF3 was detected showing identity to a putative glucokinase protein. The glucokinase encoded by ORF3 is a key enzyme in sugar degradation via the Embden-Meyerhof pathway which catalyzes the irreversible phosphorylation of glucose to glucose-6-phosphate.



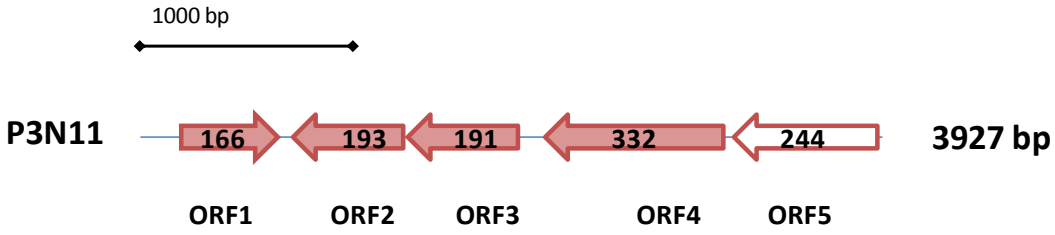
Closest Blast P match:			
ORF	ORF1	ORF2	ORF3
Accession	ZP_07956285.1	ZP_07956286.1	ZP_07956287.1
Source organism	*Lachnospiraceae bacterium* 5_1_63FAA	*Lachnospiraceae bacterium* 5_1_63FAA	*Lachnospiraceae bacterium* 5_1_63FAA
Predicted function	RpiR family Helix turn helix domain (283 aa)	RpiR family Helix turn helix domain (288 aa)	glucokinase (295 aa)
% ID/ Sim	98/99	100/100	99/99
Score	520	595	585
Match length [aa]	259	288	295

Figure 4.21 Schematic overview of the insert from clone P1D18 detected on starch- and CMC-containing plates.

The numbers in the arrows represent the length of the ORFs in amino acids. The red arrows are ORFs encoding predicted proteins which are unlikely to be involved in carbohydrate degradation.

Clone P3N11

Clone P3N11 showed the presence of a gene cluster encoding conjugative transposon proteins which mediate conjugation processes between bacteria (Figure 4.22).



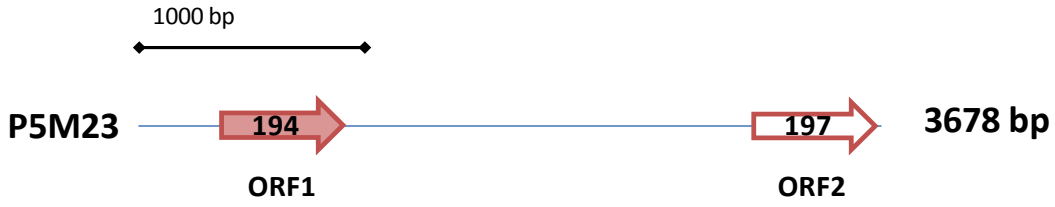
Closest Blast P match:					
ORF	ORF1	ORF2	ORF3	ORF4	ORF5
Accession	ZP_03301615.1	ZP_04848316.1	ZP_03300495.1	ZP_05547998.1	ZP_04849205.1
Source organism	*Bacteroides dorei* DSM 17855	*Bacteroides* sp. 1_1_16	*Bacteroides dorei* DSM17855	*Parabacteroides* sp. D13	*Bacteroides* sp. 1_1_6
Predicted function	lysozyme related protein (167aa)	conjugative transposon protein TraQ (165 aa)	conjugative transposon protein TraO (190 aa)	conjugative transposon protein TraN (327 aa)	conjugative transposon protein TraM (432 aa)
% ID/ Sim	69/80	66/80	75/84	77/84	71/83
Score	228	229	288	286	270
Match length [aa]	165	164	190	189	180

Figure 4.22 Schematic overview of the insert from clone P3N11detected on starch- and CMC-containing plates.

The numbers in the arrows represent the length of the ORFs in amino acids. The red arrows are ORFs encoding predicted proteins which are unlikely to be involved in carbohydrate degradation.

Clone P5M23

Two short ORFs were predicted in the insert from clone P5M23 (Figure 4.23). ORF1 showed a low degree of identity to a putative transcriptional regulator from *Butyrivibrio fibrisolvens* 16/4. ORF2 showed the similarity to a phage DNA replication protein from *Ruminococcus bromii* L2-63. Two detected ORFs are unlikely to be relevant to carbohydrate degradation.



Closest Blast P match:		
ORF	ORF1	ORF2#
Accession No	CBK74213.1	CBL16209.1
Source organism	*Butyrivibrio fibrosolvens* 16/4	*Ruminococcus bromii* L2-63
Predicted function	Transcriptional regulator (116 aa)	Phage DNA replication protein (287 aa)
% ID/ Similarity	33/51	65/80
Score	75	198
Match length [aa]	38.1	277

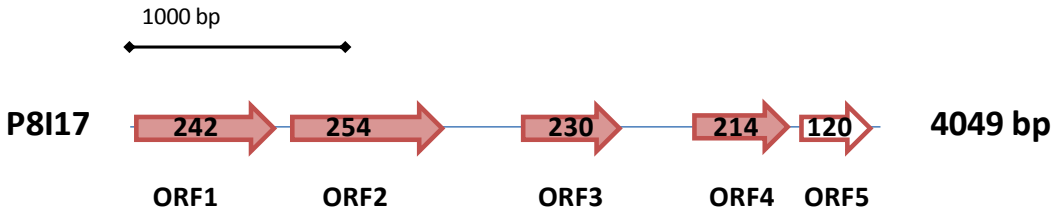
- truncated ORF

Figure 4.23 Schematic overview of the insert from clone P5M23 detected on starch and CMC-containing plates.

The truncated ORF is shown as open arrow. The numbers in the arrows represent the length of the ORFs in amino acids. The red arrows are ORFs encoding predicted proteins which are unlikely to be involved in carbohydrate degradation.

Clone P8I17

Clone P8I17 included five genes encoding proteins with high degree of identity to predicted proteins from *Lachnospiraceae bacterium* 5_1_63AA (Figure 4.24). ORF1 was a predicted glucosamine-6-phosphate isomerase. The enzyme is responsible for the conversion of D-glucosamine 6-phosphate into D-fructose 6-phosphate which is the last step in the pathway for N-acetylglucosamine (GlcNAC) utilisation in bacteria (Oliva *et al.*, 1995). ORF2 contained an iron-sulphur binding domain found in many bacterial ferrodoxin proteins that mediate electron transfer in a range of metabolic reactions. A MATE (Multi Antimicrobial Extrusion) efflux protein was encoded by *orf3* and based on its transmembrane profile acted as a drug/sodium antiporter. ORF4 showed similarity to a putative cell wall hydrolase with a domain found in enzymes involved in cell wall hydrolysis that occurs during germination. The C-terminal domain is characteristic to enzymes found in *Bacillus subtilis*. The cell wall hydrolases such as lysozyme are classified to GH family 21 and cleave β-1,4-glycosidic bonds between N-acetylmuramic acid and N-acetyl-D-glucosamine residues in a bacterial peptidoglycan. Truncated *orf5* encoded a putative uncharacterized ABC transporter predicted based on the genome sequence to be involved in uptake of amino acid, peptides and inorganic ions.



Closest Blast P match:					
ORF	ORF1	ORF2	ORF3	ORF4	ORF5#
Accession	ZP_07956664.1	ZP_07956663.1	ZP_07956662.1	ZP_07956661.1	ZP_07956660.1
Source organism	*Lachnospiraceae bacterium* 5_1_63AA	*Lachnospiraceae bacterium* 5_1_63AA	*Lachnospiraceae bacterium* 5_1_63AA	*Lachnospiraceae bacterium* 5_1_63AA	*Lachnospiraceae bacterium* 5_1_63AA
Predicted function	glucosamine-6-phosphate isomerase (242 aa)	4Fe-4S binding domain-containing protein (285 aa)	MATE efflux family protein (449 aa)	cell wall hydrolase (211 aa)	ABC transporter (340 aa)
% ID/Sim	99/100	100/100	99/99	99/100	98/98
Score	495	531	461	434	241
Match length [aa]	242	254	229	211	120

- truncated ORF

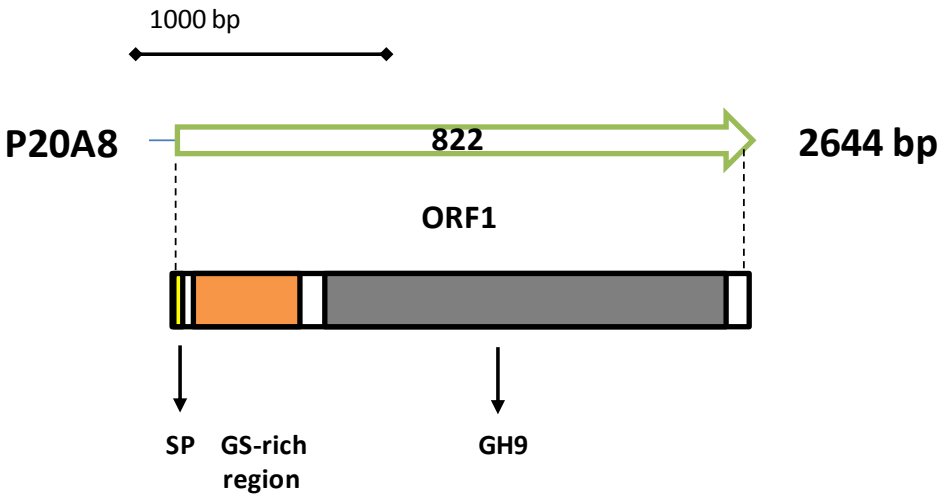
Figure 4.24 Schematic overview of the insert from clone P8I17 detected on starch-containing plates.

The truncated ORF is shown as open arrow. The numbers in the arrows represent the length of the ORFs in amino acids. The red arrows are ORFs encoding predicted proteins which are unlikely to be involved in carbohydrate degradation.

4.5.4 Clone P20A8 recovered from *L. lactis* metagenome library

Bioinformatic analysis showed the presence of one truncated ORF in a *L. lactis* metagenomic clone selected on CMC-, xylan- and lichenan-containing plates (Figure 4.25). *Orf1* showed high degree of identity to a gene from *Coprococcus eutactus* ART55/1 encoding a putative protein containing a GH9 catalytic domain and an Ig-like domain (CBK83841.1). The level of enzyme activity detected with *L. lactis* clones was very high on selective plates. Interestingly, this clone was not recovered during *E. coli* screening, possibly because of the difference in gene expression between the two hosts. The plasmid DNA from *L. lactis* clone (P20A8) was re-transformed into *E. coli* XL1 Blue (see method 2.6.4) and showed very weak enzyme activity on selective plates. The result demonstrated a significant difference of gene expression in these Gram-positive and Gram-negative heterologous hosts

The enzymes with conserved catalytic domain of the GH9 family are not well represented amongst human gut bacteria. Therefore, the further sequence- and function- based analysis of full-length genes encoding predicted enzymes from GH9 family from *Coprococcus* strains is reported in chapter 5.



Closest Blast P match:	
ORF	ORF1#
Accession No.	CBK83841.1
Source organism	*Coprococcus eutactus* ART55/1
Predicted function	Bacterial surface protein containing Ig-like domain
% ID/Sim	100/100
Score	1679
Match length [aa]	822

- truncated ORF

Figure 4.25 Schematic overview of the insert from clone P20A8 detected on CMC-, xylan- and lichenan-containing plates.

The number in the arrow represents the length of the ORF in amino acids. The truncated ORF is shown as open arrow. The green arrow is ORF encoding predicted proteins which is involved in carbohydrate degradation. Domain structure of predicted part of a putative protein from *C. eutactus* ART55/1 is presented: SP – signal peptide, GH9 – glycoside hydrolase catalytic domain family 9.

4.6 Discussion

In this study, several metagenomic clones originating from the human gut microbiota and likely to be involved in the degradation of dietary carbohydrates were identified. Several genes from category II clones appear to encode novel glycoside hydrolase enzymes (category II). Previous studies reported that the gut microbiota contain glycoside hydrolase enzymes from different GH families that enable the degradation of dietary polysaccharides (Li *et al.,* 2009, Tasse *et al.,* 2010). In the present study, positive clones were able to degrade a variety of polysaccharides: starch, carboxymethyl cellulose, xylan, lichenan and arabinofuranoside. The *in silico* analysis showed the presence of non-overlapping inserts, mostly derived from abundant gut bacteria. However, for several clones the insert derived from uncultured microorganisms or microorganism whose genome information is not available, based on a low degree of similarity to previously characterised genes.

The small insert metagenomic library was prepared initially in *E. coli* and was subsequently transferred into *L. lactis*. The phylogenetic analysis of metagenomic DNA confirmed typical microbial diversity of the faecal sample and was consistent with previous studies (Tap *et al.,* 2009, Walker *et al.,* 2011). The size of the *E. coli* library was 6,146 clones (excluding all the negative clones), covering in total 1.5 x 10^4 kb of metagenomic DNA, with an average insert size of 2.5 kb, which was equivalent to 5 bacterial genomes, assuming an average genome size of 3 Mb. The size of the *L. lactis* library was smaller and consisted of 3,456 clones (excluding all the negative clones estimated from PCR screening), covering in total 8.6 x 10^3 kb. The reported metagenomic libraries from the literature created much higher number of clones that were screened for various enzyme activities. Tasse *et al.* (2010) constructed a library of 156,000 *E. coli* clones covering in total 5.46 x 10^6 kb of metagenomic DNA with an average insert size of 35 kb. Walter *et al.* (2005) constructed a library of 5,760 clones with an average insert size of 55 kb (3.2 x 10^5 kb of metagenomic DNA). The library from rumen microbiota produced by Ferrer *et al.* (2005) comprise of 200,000 clones with an average insert size of 5.5 kb, which corresponded to 1.1 x 10^6 kb of environmental DNA. The frequency of finding positive clones in the present work was found relatively high compared to other studies. Here, the genes encoding putative amylases were found in a frequency 1 per 1.3 Mb compared to Tasse *et al.* (2010) work where it was 1 per 70 Mb. The

frequency of xylanase-encoding genes was 10-fold higher in the present study than reported by Tasse *et al.* (2010) (1 per 15 Mb vs. 1 per 150 Mb). The glucanase-encoding genes were detected in the present study in a frequency of 1/15 Mb compared to 1/200 Mb reported by Tasse *et al.* (2010) and to 1/110 Mb reported by Walter *et al.* (2005). Nine clones with CMCase activity detected by Ferrer *et al.* (2005) corresponded to a frequency of 1 per 10 Mb which was approximately 10-fold lower than the frequency of CMCase positive clones found here. These differences are likely due to technical differences between projects. For example, the copy number between the different cloning vectors could have led to differences in expression level.

Functional metagenomic screening relies on reproducible assays for the function of interest. In this work, phenotypic screening was based on visual inspection of substrate-containing plates. This method was reported to be successful in many metagenomic projects, which searched for glycoside hydrolases from different environmental samples (Kim *et al.*, 2011; Hess *et al*,. 2011). The present work showed a poor reproducibility between repeated screens that has led to recovery of false-positive clones. A previous study also reported several pitfalls of the applied detection method using skimmed milk agar to functionally screen a gut metagenomic library for proteases, which led to identification of false-positive clones (Jones *et al.*, 2007). Another study (Litthauer *et al.*, 2010) demonstrated problems of using tributyrin agar to detect lipolytic activity. In this study positive clones were subjected to a second round of screening, which improved the rate of detection of truly positive clones.

This work also tried to establish the suitability of using *Lactococcus lactis* as a heterologous host for the metagenomic application (further discussed in chapter 6), which led to recovery of one highly active fibrolytic clone that showed a high identity to a gene originating from *Coprococcus eutactus* ART55/1. Interestingly, this clone was not detected in the *E. coli* library and was exclusively expressed in the *L. lactis* host which confirms the gene expression differences between two the hosts. Further analysis of this clone was conducted and is presented in the next chapter.

Chapter 5

From sequence to function – analysis of cellulase-encoding genes from *Coprococcus* species

5.1 Introduction

A highly active fibrolytic positive clone with strong similarity to a gene from *Coprococcus eutactus* ART55/1 was detected during functional screening of the human gut metagenome using *L. lactis* as a heterologous host (chapter 4). The ORF responsible for activity carries a GH9 catalytic domain. Studies on *Coprococcus* species have focused mainly on their production of butyrate (Louis *et al.,* 2004, Duncan *et al.,* 2002). It has also been reported that a lower abundance of *Coprococcus*-related bacteria was observed amongst individuals with irritable bowel syndrome (Malinen *et al.,* 2010). A recent phylogenetic study highlighted importance of *Coprococcus*-related bacteria (Peris-Bondia *et al.,* 2011). There are no reports, however on the cellulolytic activity of *Coprococcus* strains or on their contribution to carbohydrate breakdown in the human gut.

Bacterial cellulose degradation requires the action of free cellulases or high-molecular-weight complexes called cellulosomes (Bayer *et al.,* 2008). The bacterial cellulases are classified into thirteen glycoside hydrolase families, and family 9 is one of the most abundant. Glycoside hydrolase family 9 (GH9) contains mainly endoglucanases with a few processive endoglucanases (exo- and endo- activity), and is represented by well-characterized cellulases from *Clostridium cellulolyticum*, *Thermobifida fusca* and *Clostridium thermocellum* (Mingardon *et al.,* 2011). Although the GH9 family contains more than 350 bacterial enzymes, there are few examples of this family encoded by human gut bacteria. The GH9 family in the CAZy database contains seven predicted cellulases from the highly active cellulolytic strain of *Ruminococcus champanellensis* 18P13 (Chassard *et al.,* 2011, Robert and Bernalier-Donadille 2003) and two predicted cellulases from *Coprococcus eutactus* ART55/1. The importance of cellulose degradation by human gut isolates was already highlighted by Robert *et al*. (2003) however our knowledge of cellulose breakdown by the human gut microbiota is still limited.

Due to the scarcity of information on cellulose breakdown by human gut bacteria, it was considered important to explore the function of the predicted cellulases encoded by *Coprococcus* strains.

5.2 Identification of glycoside hydrolase family 9 genes from **Coprococcus**

The sequence of *L. lactis* clone P20A8 from metagenomic library showed 100% identity to a predicted protein from *Coprococcus eutactus* ART55/1 with a GH9 catalytic domain and an Ig-like domain (CBK83841.1, named ART_GH9/L). The deduced full length protein, based on *C. eutactus* ART55/1 genome information, was 1782 aa. The BLAST P results showed that ART_GH9/L was homologous to a hypothetical protein COPEUT_00372 from *Coprococcus eutactus* ATCC 27759 (ZP_02205610.1), and shared 78% identity over the whole length of the gene. The GH9 catalytic domains of the two coprococcal proteins were also highly homologous to a GH9 domain of a hypothetical protein CLOL250_00652 from *Coprococcus* sp. L2-50 (ZP_02073894.1, named L250_GH9/L). The C-terminal part of L250_GH9/L was however different from ART_GH9/L and COPEUT_00372, hence L250_GH9/L and ART_GH9/L were chosen for further analysis.

In addition, two genes also carrying GH9 domains but which were significantly shorter were identified (based on CAZy database entries) in the draft genomes of *C. eutactus* ART55/1 and *Coprococcus* sp. L2-50 (CBK83282.1.1 and ZP_02074498.1). In summary, two pairs of GH9 encoding genes from *Coprococcus* species were used for further study (Table 5.1).

Name	Name	Accession number	Bacterial origin	Length [aa]	Domain structure	G+C %
Bacterial surface proteins containing Ig-like domains	ART_GH9/L	CBK83841.1	*C. eutactus* ART55/1	1782	SP - GH9 -Ig	46.02
Glycosyl hydrolase family 9	ART_GH9/S	CBK83282.1	*C. eutactus* ART55/1	757	SP - CelD - CBM4_9 - GH9	50.04
Hypothetical protein CLOL250_00652	L250_GH9/L	ZP_02073894.1	*Coprococcus* sp. L2-50	1792	SP-CBM2 - GH9 - Ig -Ig - Ig - Ig	45.40
Hypothetical protein CLOL250_01268	L250_GH9/S	ZP_02074498.1	*Coprococcus* sp. L2-50	762	SP - CelD - CBM4_9 - GH9	44.60

Table 5.1 Overview of GH9 enzymes from *Coprococcus* strains used in the present study.

Name: L – long gene, S – short gene, SP – signal peptide, GH9 – glycoside hydrolase family 9 catalytic domain, Ig – Ig-like domain, CelD – Ig-like domain, CBM – carbohydrate binding module.

5.3 Sequence and predicted domain structure of putative protein ART_GH9/L

The deduced length of the full ORF ART_GH9/L was 1782 amino acids (Appendix 3.1). Bioinformatic analysis showed a multi-domain structure for the deduced protein with several features characteristic of glycoside hydrolase family 9 enzymes (Figure 5.1). The N-terminal part of the *Coprococcus eutactus* ART55/1 protein had a typical signal peptide pattern (1-32 aa) with a cleavage site between alanine and glutamate (Natale *et al.,* 2008). The region between amino acids 349-759 contains a domain found in glycoside hydrolase family 9 (PF00759). The signature pattern for GH9 enzymes (741-759 aa) contains a glutamate residue (aa 752), which acts as a proton donor in the hydrolysis reaction. Two conserved aspartate residues (aa 404 and 407), are thought to deprotonate water involved in nucleophilic attack were present (Pereira *et al.,* 2009, Kurokawa *et al.,* 2002, Pereira *et al.,* 2010). Amino acid alignment of the catalytic domain of the ART_GH9/L protein demonstrated its homology to the catalytic domains of glycoside hydrolase family 9 enzymes from *Cl. thermocellum*, *Cl. cellulovorans*, *Cl. cellulolyticum* and *Ruminococcus* sp. 18P13 (Figure 5.2). A modular architecture is characteristic for well characterised enzymes of the GH9 family. The endoglucanase Cel9B from *Ruminococcus albus* 8 (Devillard *et al.,* 2004), CelE from *Cl. cellulolyticum* (Gaudin *et al.,* 2000) and CelK from *Cl. thermocellum* (Kataeva *et al.,* 2001) showed the presence of a family 4 carbohydrate binding module (CBM4) and an Ig-like domain preceding the family 9 catalytic domain. The cellulase from *Cl. thermocellum* –Cel9I – was reported to be a tri-modular non-cellulosomal enzyme consisting of a GH9 catalytic domain joined to a tandem of CBM3 modules. Many GH9 enzymes carry a C-terminal dockerin domain which allows formation of the cellulosome complex (Bayer *et al.,* 2008). These accessory domains (to the catalytic domain) are recognized in most microbial cellulases. The main function of CBMs is to bind cellulose fibres. These modules play an important role in cellulose hydrolysis. The removal of the CBM leads to reduction or abolition of binding to insoluble substrates leading to partial or complete loss in catalytic activity. (Gaudin *et al.,* 2000, Kataeva *et al.,* 1999, Burstein *et al.,* 2009). A short TGS-rich linker was observed in the protein ART_GH9/L that joins the catalytic domain and Ig-like domain. Threonine-rich linkers were identified for a cellulosomal scaffoldin enzyme CipV from *Acetivibrio cellulolyticus*. The threonine residues serve as glycosylation site. They form numerous polar interactions with the

catalytic domain (Burstein *et al.,* 2009). Ig-like domains are found not only associated with GH9 enzymes, but are also found in other GH families. A cellulase from *Ruminococcus* sp. 80/3 was assigned to the GH5 family and showed the presence of an Ig-like domain at the C-terminal side of the catalytic domain (see section 3.5.3). A recent study showed that an Ig-like domain may be necessary for protein folding by maintaining proper orientation of the active site and catalytic residues (Liu *et al.,* 2010). The C-terminal of the ART_GH9/L protein showed the presence of two internal repeats (aa 1501-1667), which share 60% identity. This region showed 32% identity to a *Ruminococcus albus* F-40 endoglucanase with a dockerin domain. The sequence of ART_GH9/L was also similar to the C-terminal region of dockerin type 1 protein from *Ruminococcus albus* 7 and 8 with 37% and 40% identity, respectively. The clostridial and ruminococcal dockerin domains consist of a well-conserved 22-residue repeat with calcium-binding aspartate or asparagine residues (Pages *et al.,* 1997, Ohara *et al.,* 2000). The protein ART_GH9/L does not contain the dockerin motif found in cellulosomal dockerin type proteins, therefore the presence of a dockerin domain in this protein is not confirmed. Visual analysis of the amino acid sequence of ART_GH9/L showed a stretch of a polar and a hydrophobic region (276 amino acid) at the N-terminal end of the protein. The predominant amino acids were serine (20%) and glycine (12%). A second GS-rich region was identified downstream of the Ig-like domain. This region of 253 amino acids contained 38 serine residues and 35 glycine residues. Among known proteins, glycine-serine rich repeats were found previously in a cell-associated α-amylase from *Butyrivibrio fibrisolvens* 16/4 (Ramsay *et al.,* 2006) and were predicted to attach the protein to the peptidoglycan of the cell wall. Glycine-serine rich regions were also observed for a surface located carbohydrate-binding protein, CbpC from *Streptomyces coelicolor* A3(2). The protein was shown to bind crystalline chitin and cellulose and deletion of the GS-regions led to protein inactivation. It was also hypothesised that the GS-rich domain bridges over a peptidoglycan layer to the extracellular space (Walter and Schrempf 2008). The reported proteins contained a C-terminal signal motif recognised by a sortase enzyme (LPXTG) for subsequent covalent linkage to the cell wall peptidoglycan. Inspection of the ART_GH9/L sequence did not recognise a cell-wall sorting motif. The C-terminal end of the ART_GH9/L protein (115 amino acids) was however highly charged and was composed of 25% lysine (K-rich region). Based on those observations, it is predicted

that ART_GH9/L is non-cellulosomal enzymes, secreted to the extracellular space by a bacterial secretory mechanism.

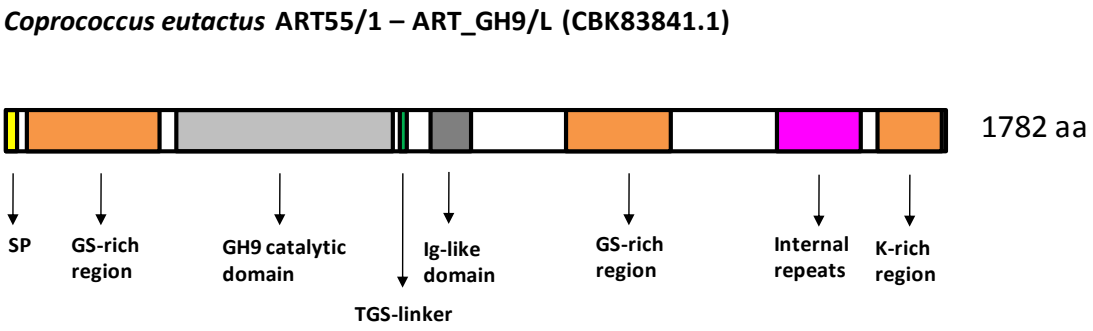


Figure 5.1 Multi-domain structure of ART_GH9/L cellulase from *Coprococcus eutactus* ART55/1.

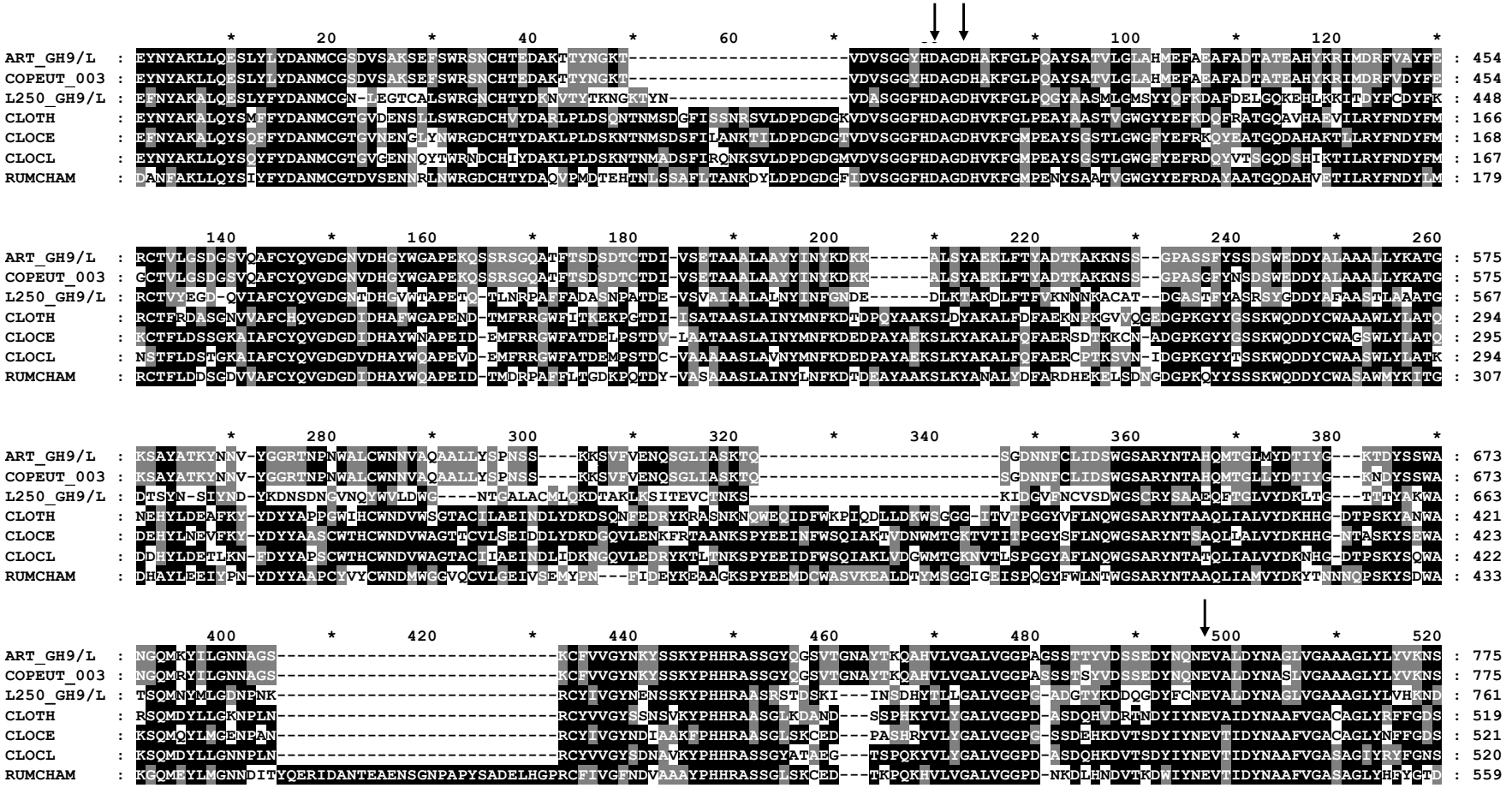


Figure 5.2 Amino acid sequence alignment of the catalytic domains of family 9 genes of the top BlastP matches from ART_GH9/L and L250_GH9/L.

The conserved residues involved in the catalytic reaction are marked with arrows. ART_GH9/L (CBK83841.1) - *Coprococcus eutactus* ART55/1; COPEUT_003 (ZP_02205610.1) - *Coprococcus eutactus* ATCC 27759; L250_GH9/L (ZP02073894.1) - *Coprococcus sp.* L2-50; CLOTH (YP_001039204.1) - *Clostridium thermocellum* ATCC 27405; CLOCE (YP_003844052.1) - *Clostridium cellulovorans* 743B; CLOCL (YP_002505111.1) - *Clostridium cellulolyticum* H10; RUMCHAM (CBL16391.1) - *Ruminococcus* sp. 18P13.

5.4 Sequence and predicted domain structure of putative protein L250_GH9/L

The deduced length of L250_GH9/L protein from *Coprococcus sp.* L2-50 was 1809 aa, according to GeneBank data (Appendix 3.2). Visual inspection of the upstream region to the proposed start codon (M1) did not identify a potential ribosomal binding site. A second methionine residue (M18) was separated from the predicted start codon by 18 nucleotides. The methionine M18 is most likely to be the translational start codon. A potential ribosomal binding site with the sequence GGAAG was spaced from the proposed ATG codon (M18) by an optimal eight nucleotides. A putative -10 region promoter sequence TATAAT was found upstream of the proposed translational start codon. A conserved TG motif characteristic of Gram-positive bacterial genes was also recognised (Voskuil and Chambliss 1998). The -35 putative promoter sequence TTGTGG was separated from the -10 region by 17 nucleotides. The methionine M18 is most likely to be the translational start also because of the presence of a signal peptide, which was not identified with M1 as start codon. Therefore, the deduced length of the L250_GH9/L protein was defined as 1792 amino acids. Protein L250_GH9/L displayed a multi-domain architecture (Figure 5.3, Appendix 3), however it differed markedly from protein ART_GH9/L. A signal peptide sequence was recognised for the proposed ORF with cleavage between alanine residues (A35 and A36). A polar and hydrophobic region (139 amino acids) was observed at the N-terminal end of the protein, which is comprised of 24 residues of serine and 21 residues of glycine. An N-terminal carbohydrate-binding module, CBM_2, was identified, which shows homology to previously characterised CBM_2 domains from different GH families (Figure 5.4). Cellulose, xylan and chitin binding properties were reported for the CBM_2 family (Simpson *et al.*, 2000). Conserved tryptophan residues, which interact with the surface of the polysaccharide, were detected for the CBM_2 domain from protein L250_GH9/L (Figure 5.4) (Simpson *et al.,* 2000). The region aa 338-761 showed the presence of a glycoside hydrolase family 9 catalytic domain with two conserved aspartate residues (D398 and D401). The glutamate residue which acts as a catalytic proton donor was identified at position E738 (Figure 5.2). At both sides of the GH9 domain, TGS-rich linkers were found which joined the accessory modules to the catalytic domain.

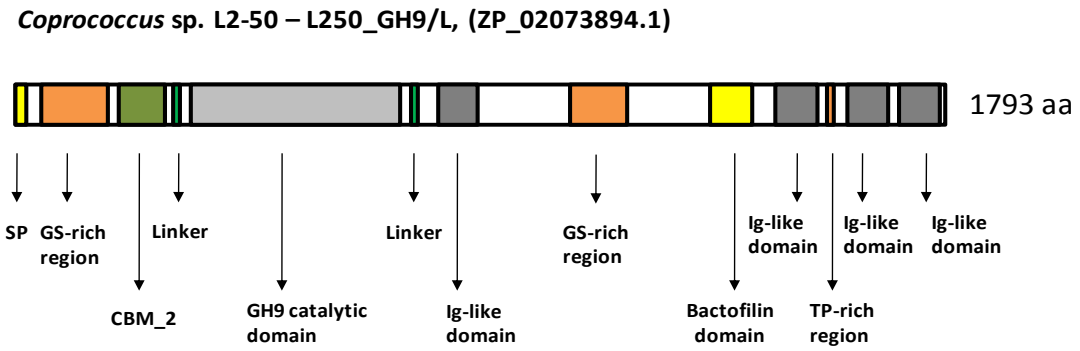


Figure 5.3 Modular architecture of cellulase L250_GH9/L from *Coprococcus* sp. L2-50.

SP – signal peptide, CBM – carbohydrate-binding module.

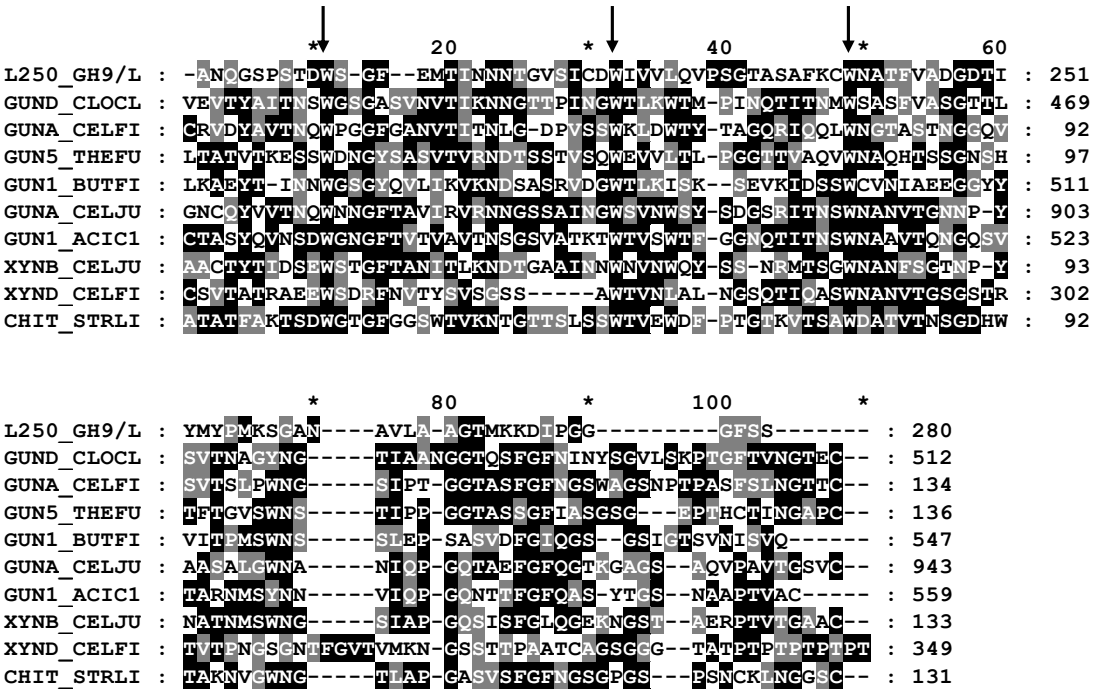


Figure 5.4 Amino acid sequences alignment of the CBM_2 domain.

Locus tags: L250_GH9/L – *Coprococcus* sp. L250; GUN – endoglucanase; XYN – xylanase; CHIT – chitinase; CLOCL – *Clostridium cellulovorans*; CELFI – *Cellulosomas fimi*; THEFU – *Thermomonospora fusca*; BUTFI – *Butyrivibrio fibrisolvens*; CELJU – *Cellvibrio japonicus*; ACIC1 – *Acidothermus cellulolyticus*; STRLI – *Streptomyces lividans*. Conserved tryptophan residues are marked with arrows.

At the C-terminal side of the catalytic domain, four Ig-like domains (group 2) were identified with conserved residues of tryptophan and glycine (Figure 5.5). A GS-rich region of amino acids was observed in the middle part of the protein. A short domain at position 1368-1452 aa showed similarity to the DUF583 domain. Bactofilin proteins (PF04519) contain a DUF583 domain and proline-rich segments at the N- and C-terminal ends, which are known to mediate protein-protein interactions (KAY *et al.*, 2000). This widely conserved and distributed class of cytoskeletal proteins mediates polar localization of a cell wall synthase in *Caulobacter crescentus* during cell division. Bactofilins were shown to be versatile structural elements with functions in a range of different cellular pathways such as stalk biogenesis, bacterial motility and cell wall synthesis (Kuhn *et al.,* 2010). The amino acid sequence of the predicted DUF583 domain from the L250_GH9/L protein was aligned with the DUF583 domain from bactofilin proteins and showed the presence of conserved glycine residues and hydrophobic amino acids (Figure 5.6). A proline rich sequence was also identified in the L250_GH9/L protein, but it was separated from the DUF583 domain by a 171 amino acid region containing an Ig-like domain. Kuhn *et al*. (2010) demonstrated the role of bactofilin in cell movement and locomotion. The functional role of the DUF583 domain of bactofilin proteins in L250_GH9/L is unclear. Predicted bactofilin proteins with DUF583 domain were also found in genomes of other fibrolytic gut bacteria: *Roseburia intestinalis*, *Roseburia inulinivorans*, *Eubacterium rectale* and *Eubacterium siraeum* (information derived from Pfam database).

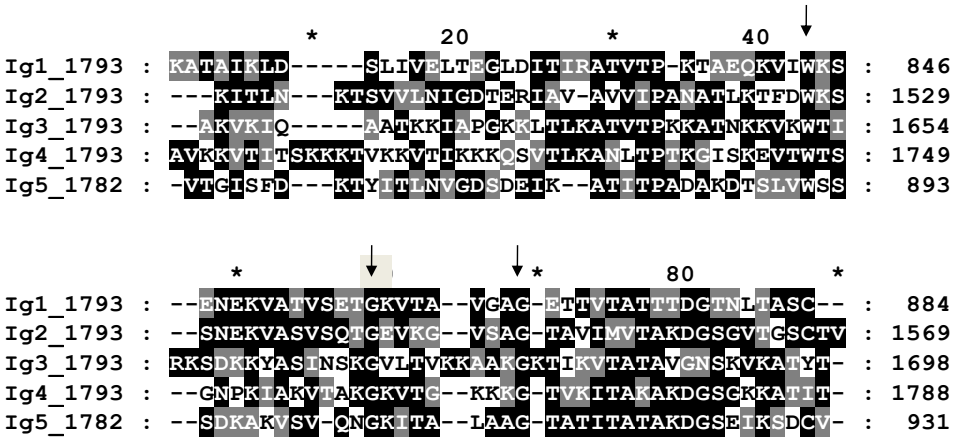


Figure 5.5 Amino acid alignment of a Ig-like domains from L250_GH9/L (Ig1, Ig2, Ig3, Ig4) and ART_GH9/L (Ig5).

Conserved tryptophan and glycine residues are marked with arrows.

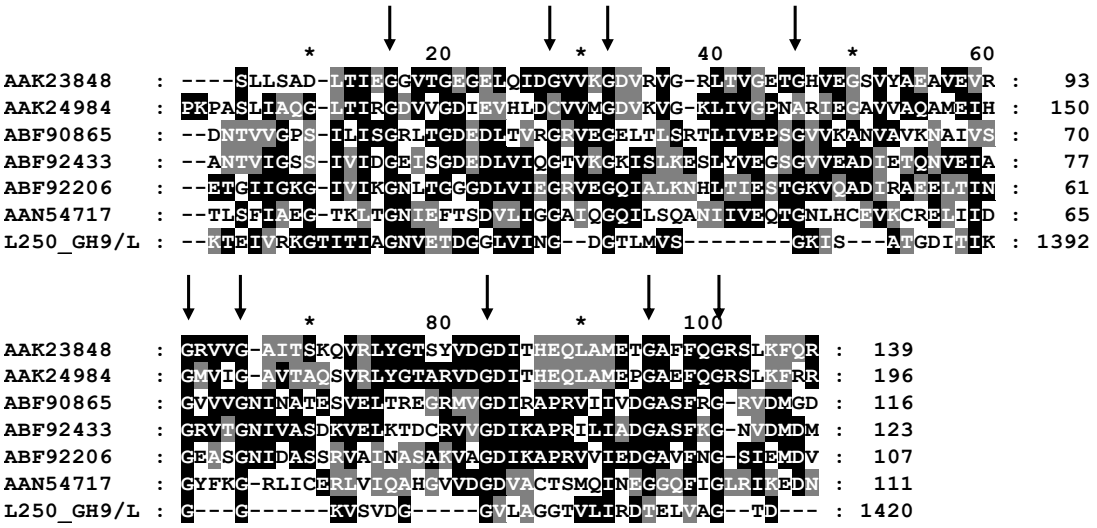


Figure 5.6 Amino acid sequences alignment of bactofilin DUF583 domain homologues.

AAK23848 and AAK24984 – *Caulobacter crescentus*; ABF90865, ABF92433, ABF92206 – *Myxococcus xanthus*; AAN54717 – *Shewanella oneidensis*; L250_GH9/L (ZP_02073894.1) – *Coprococcus* sp. L2-50. The conserved glycine residues are marked with arrows.

5. 5 Molecular analysis of ART_GH9/S and L250_GH9/S proteins

Two genes from *C. eutactus* ART55/1 and *Coprococcus* sp. L2-50, named ART_GH9/S (Appendix 3.3) and L250_GH9/S (Appendix 3.4), which also encode predicted GH9 enzymes, showed a high level of identity to each other (60%) and a homologous multi-domain structure (Figure 5.7). The translational products of the full ORFs were proteins of 757 amino acids and 762 amino acids, respectively (Appendix 3.3 and 3.4). The N-terminal region of both proteins has a pattern for a typical lipoprotein signal peptide with a cleavage site between glycine and cysteine (Natale *et al.,* 2008). The cysteine residue is modified following translation by lipid attachment (Natale *et al.,* 2008). Following the signal peptide region, two domains were identified: a carbohydrate binding domain 4_9 (Johnson *et al.,* 1996) and an N-terminal Ig-like domain found in cellulases (Liu *et al.,* 2010). The ability of binding to different polysaccharides, including xylan, β-1,3-glucan, β-1,3-1,4-glucan, β-1,6-glucan and amorphous cellulose, was demonstrated for CBM4_9 domains. In addition to sugar binding properties, CBMs are involved in their disruption by arranging polysaccharide fibres and structural modification of the hydrogen bond network of polysaccharides (Gaudin *et al.,* 2000, Johnson *et al.,* 1996, Kataeva *et al.,* 2001, Kataeva *et al.,* 1999). The catalytic domain of ART_GH9/S and L250_GH9/S was homologous to a number of GH9 cellulases (Figure 5.8). The signature pattern for GH9 family cellulases with the glutamate residue, which acts as catalytic proton donor, and two conserved aspartate residues are present (Figure 5.8).

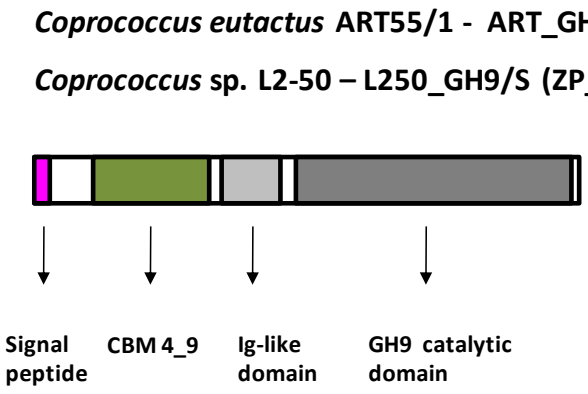


Figure 5.7 Schematic representation of the multi-domain structure of GH9 enzymes from *Coprococcus eutactus* ATR55/1 and *Coprococcus* sp. L2-50.

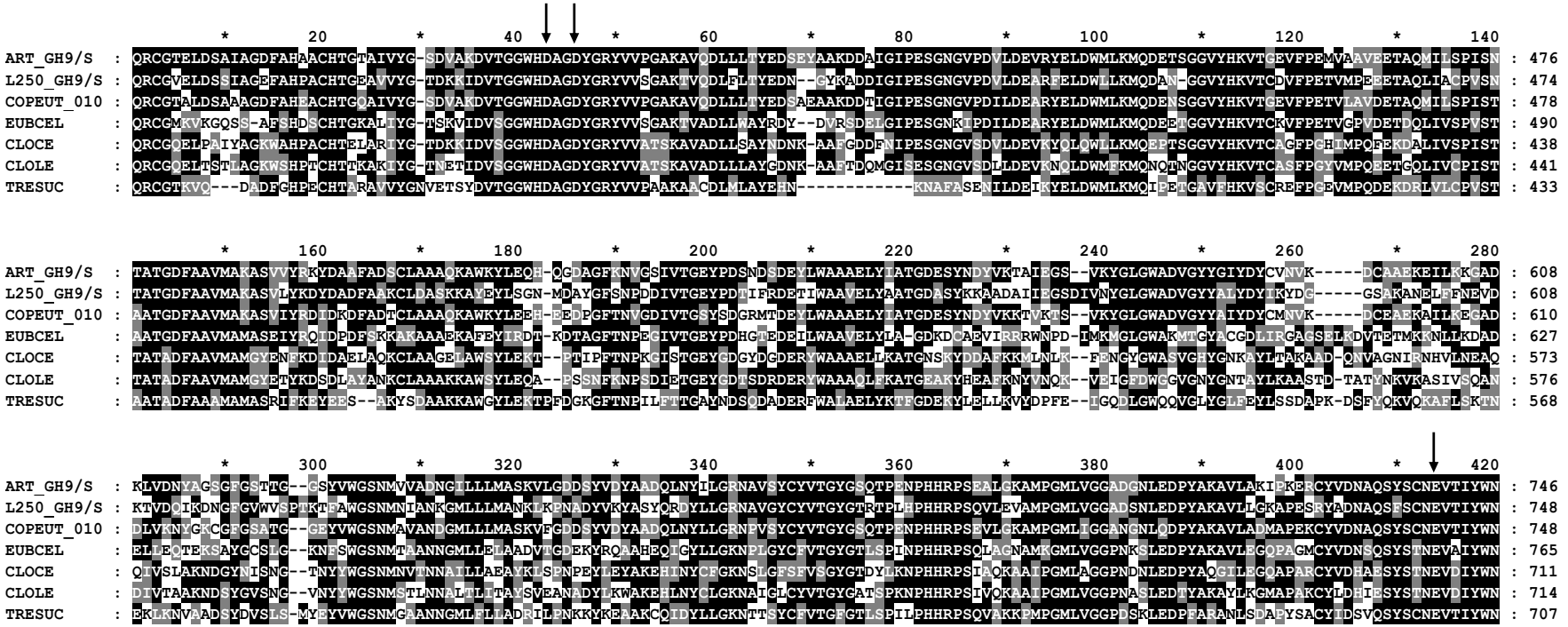


Figure 5.8 Amino acid sequence alignment of the catalytic domains of family 9 genes from the top BlastP matches to ART_GH9/S and L250_GH9/S.

The conserved residues involved in the catalytic reaction are marked with arrows. ART_GH9/S (CBK83282.1) – *Coprococcus eutactus* ART55/1; L250_GH9/S (ZP_02074498.1) – *Coprococcus* sp; L2-50, COPEUT_010 (ZP_02206240.1) – *Coprococcus eutactus* ATCC 27759; EUBCEL (ZP_07838541.1) - *Eubacterium cellulosolvens* 6; CLOCE (YP_00384299.1) – *Clostridium cellulovorans* 743B; CLOLE (YP_004307606.1) - *Clostridium lentocellum* DSM 5427; TRESUC (YP_004366651.1) - *Treponema succinifaciens* DSM 2489.

5.6 Phylogenetic analysis of GH9 enzymes from **Coprococcus** *species*

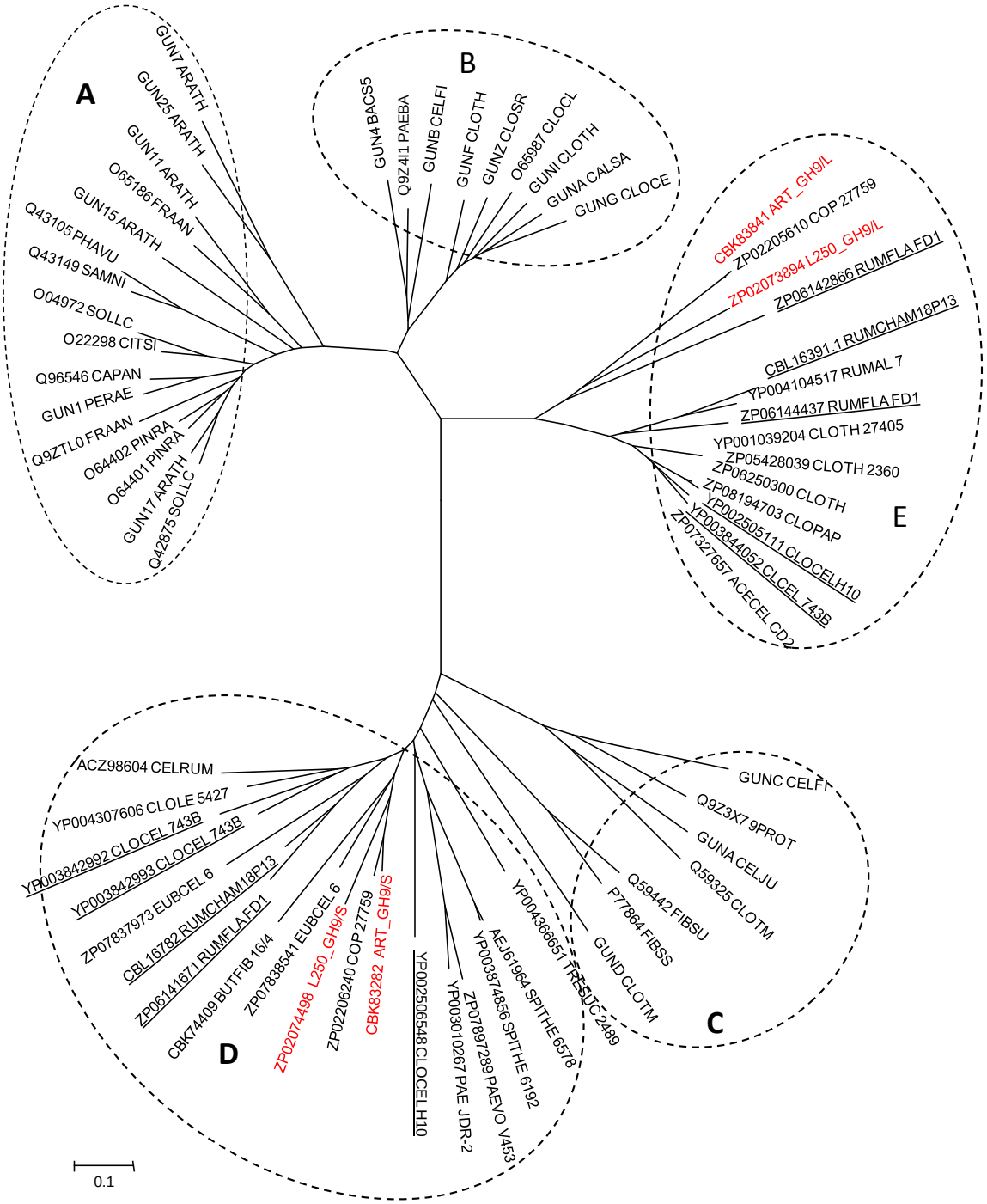
A phylogenetic tree of glycoside hydrolases family 9 based on amino acid sequences of GH9 catalytic domains, was generated. It classified enzymes examined in the present study into different cellulase groups (Figure 5.9, Appendix 3.5). The catalytic domains of enzymes ART_GH9/S and L250_GH9/S were closely related to cellulases of GH9 family from *Eubacterium cellulosolvens* and *Butyrivibrio fibrisolvens*. The other closely related enzymes derived from cellulosome-producing bacterium such as *Clostridium cellulovorans* (Tamaru *et al.,* 2010) and cluster IV ruminococci - *Ruminococcus flavefaciens* FD1 (Berg Miller *et al.,* 2009) and *Ruminococcus champanellensis* 18P13 (Chassard *et al.,* 2011). The modular structure of these enzymes showed similar architecture to protein ART_GH9/S and L250_GH9/S with CBM4_9 and Ig-like domain append to catalytic region. This structural organisation is characteristic of well-studied components of the cellulosome from *Clostridium thermocellum* (CelK) and *Cl. cellulolyticum* (CelE). The coprococcal enzymes share also the same modular structure with other GH9 non-cellulosomal cellulases like CenC from *Cellulomonas fimi* (CenC) or EngO from *Clostridium cellulovorans* (Han *et al.,* 2005). Based on their multi-domain structure these enzymes are typical to theme D cellulases (Devillard *et al.,* 2004). Enzymes lacking the CBM but bearing the Ig-like domain are classified to theme C alongside to endoglucanase CelD of *Cl. thermocellum* cellulosome or cellulase CelC from *Butyrivibrio fibrisolvens* (Ding *et al.,* 1999).

The phylogenetic analysis of catalytic domains of ART_GH9/L and L250_GH9/L separate these enzymes to a newly created group E of cellulases which has not been reported previously in the literature. This newly created cluster of GH9 enzymes requires futher investigation. The closest related enzymes to ART_GH9/L and L250_GH9/L derived from *Ruminococcus flavefaciens* FD-1. The other closely related enzymes derived from cellulosome-producing bacterium *Clostridium thermocellum*, *Cl. papyrosolvens* and *Cl. cellulolyticum*. The phylogenetic analysis showed that the GH9 catalytic domain of coprococcal proteins, together with further enzymes from other organisms, clearly separated them from the rest of typical theme B enzymes. The GH9 enzymes belonging to architectural theme B, contain a GH9 catalytic domain followed by a carbohydrate-binding module and either a second CBM or dockerin domain (Burstein *et al.,* 2009, Gilad *et al.,* 2003). The group E

enzymes showed more diverse structural architecture. Several enzymes (ZP_06250300.1, YP_001039204.1, ZP_05428039.1YP_003844052.1, ZP_07327657.1) showed the presence of the dockerin domain which is characteristic of cellulosomal enzymes involved in the assembly of this highly complex structure. Neither of the coprococcal proteins showed characteristic motifs for a dockerin domain (Burstein *et al.,* 2009). For some of the enzymes, only GH9 catalytic domain was identified. Others, such as coprococcal enzymes bear Ig-like domains. It is possible that proteins from group E may bear novel accessory domains, previously not reported in the literature. Interestingly, group D and E consists of enzymes derived from the same organisms such as *Clostridium cellulovorans* 743B, *Clostridium cellulolyticum* H10 *Ruminococcus champanellensis* 18P13, *Ruminococcus flavefaciens* FD-1 and *Coprococcus* strains. This could indicate that these bacteria possess at least two enzymes involved in cellulose degradation, possibly with different substrate specificities.

Figure 5.9 Phylogenetic analysis of GH9 enzymes from *Coprococcus* species.

Phylogenetic tree was generated using a ClustalW alignment based on representative GenBank sequences of catalytic domains of GH9 family enzymes (for theme A, B and C) or the top BlastP matches (theme D, E). The branch name is the accession number for the protein sequence, followed by bacterial origin (in abbreviation). Coprococcal GH9 enzymes used in present study are shown in red. The underlined enzymes derived from the same bacterium which are present in group D and E. Abbreviations: GUN – endoglucanase, ARATH – *Arabidopsis thaliana*, FRAAN – *Fragana ananassa,* PAHVU – *Phaseolus vulgaris,* SAMNI – *Sambucus nigra,* SOLLC – *Solanium lycopersicum*, CITSI – *Citrus sinensis,* CAPAN – *Capsicum annuum*, PERAE – *Persea americana,* PINRA – *Pinus radiate,* BACS5 – *Bacillus* sp., PAEBA - *Paenibacillus barcinonensis,* CELFI – *Cellulomonas fimi*, CLOTH/ CLOTM - *Cl. thermocellum,* CLOSR - *Cl. stercorarium,* CLOCL – *Cl. cellulolyticum,* CALSA - *Caldocellum saccharolyticum,* CLOCE – *Cl. cellulovorans*, COP - *Coprococcus,* CLOLE - *Clostridium lentocellum,* RUMCHAM - *Ruminococcus champanellensis,* RUMAL – *R. albus*, RUMFLA - *R. flavefaciens*, CLOPAP – *Clostridium papyrosolvens* , ACECEL - *Acetivibrio cellulolyticus,* STRRE - *Strepromyces*, PROT - *Pseudomonas,* CELJU – *Cellvibrio japonius,* FIBSU/ FIBSS – *Fibrobacter succinogenes,* TRESUC - *Treponema succinifaciens*, SPITHE – *Spirochaeta thermophila* , CELRUM - *Cellulosilyticum ruminicola,* EUBCEL - *Eubacterium cellulosolvens,* BUTFIB - *Butyrivibrio fibrisolvens,* PAEVO - *Paenibacillus vortex,* PAE - Paenibacillus. The branch length index is presented by a bar at the bottom of the figure. The length of a branch denotes the genetic distance between the two taxa it connects. 0.1 = means an average of one substitution every ten nucleotide sites. The domain structure of all the enzymes used for the phylogenetic tree construction is presented in Appendix 3.5.


A
GUN7 ARATH
GUN25 ARATH
GUN11 ARATH
O65186 FRAAN
GUN15 ARATH
Q43105 PHAVU
Q43149 SAMNI
O04972 SOLLC
O22298 CITSI
Q96546 CAPAN
GUN1 PERAE
Q9ZTL0 FRAAN
O64402 PINRA
O64401 PINRA
GUN17 ARATH
Q42875 SOLLC
B
GUN4 BACS5
Q9Z4I1 PAEBA
GUNB CELFI
GUNF CLOTH
GUNZ CLOSR
O65987 CLOCL
GUNI CLOTH
GUNA CALSA
GUNG CLOCE
E
CBK83841 ART_GH9/L
ZP02205610 COP 27759
ZP02073894 L250_GH9/L
ZP06142866 RUMFLA FD1
CBL 16391.1 RUMCHAM18P13
YP004104517 RUMAL 7
ZP06144437 RUMFLA FD1
YP001039204 CLOTH 27405
ZP05428039 CLOTH 2360
ZP06250300 CLOTH
ZP08194703 CLOPAP
YP002505111 CLOCELH10
YP003844052 CLCEL 743B
ZP07327657 ACECEL CD2
C
GUNC CELFI
Q9Z3X7 9PROT
GUNA CELJU
Q59325 CLOTM
P77864 FIBSS
Q59442 FIBSU
GUND CLOTM
YP004366651 TRESUC 2489
AEJ61984 SPITHE 6578
YP003874556 SPITHE 6192
ZP07897289 PAEVO V453
YP003010267 PAE JDR-2
YP002506548 CLOCELH10
D
ACZ98604 CELRUM
YP004307606 CLOLE 5427
YP003842992 CLOCEL 743B
YP003842993 CLOCEL 743B
ZP07837973 EUBCEL 6
CBL 16782 RUMCHAM18P13
ZP06141671 RUMFLA FD1
CBK74409 BUTFIB 16/4
ZP07838541 EUBCEL 6
ZP02074498 L250_GH9/S
ZP02206240 COP 27759
CBK83282 ART_GH9/S
0.1

5.7 Heterologous expression of GH9 enzymes in E. coli *and* L. lactis

The GH9 enzymes from *Coprococcus* species were examined further by heterologous gene expression. A set of clones was prepared in order to determine their functionality in *E. coli* and *L. lactis*. Briefly, the four full-length genes encoding the complete GH9 enzymes form *C. eutactus* ART55/1 and *Coprococcus* sp. L2-50, including their predicted promoter sites, were amplified from genomic DNA (see section 2.5.1), cloned in the same orientation into shuttle vector pTRKL2 and transformed into *E. coli* and *L. lactis* (method 2.6.4 and 2.6.2). The fibrolytic activity of *E. coli* and *L. lactis* clones was examined on substrate-containing plates using the Congo Red detection method (2.8.1). The results are shown in Table 5.2 and Figure 5.10.

		Substrate		
Host	**Enzyme**	Lichenan	CMC	Xylan
E. coli XL1 Blue	ART_GH9/S	-	+ (7)	-
	ART_GH9/L	-	+ (6)	-
	L250_GH9/S	+++ (16)	++ (14)	+++ (17)
	L250_GH9/L	+ (8)	++ (13)	+ (6)
L. lactis MG1363	ART_GH9/S	+ (8)	++ (10)	++ (12)
	ART_GH9/L	++ (10)	++ (10)	++ (12)
	L250_GH9/S	+ (8)	++ (10)	++ (14)
	L250_GH9/L	+++ (15)	++ (14)	+++ (16)

- – not detected
+ – weak activity (≤ 9 mm)
++ – moderate activity (10-14 mm)
+++ - strong activity (≥ 15 mm)

Table 5.2 Results of plate assays for cellulolytic enzyme activity in *E. coli* and *L. lactis* clones encoding GH9 enzymes from *Coprococcus* species.

The strength of the enzyme activity was established based on the diameter of the clear zone (in brackets).

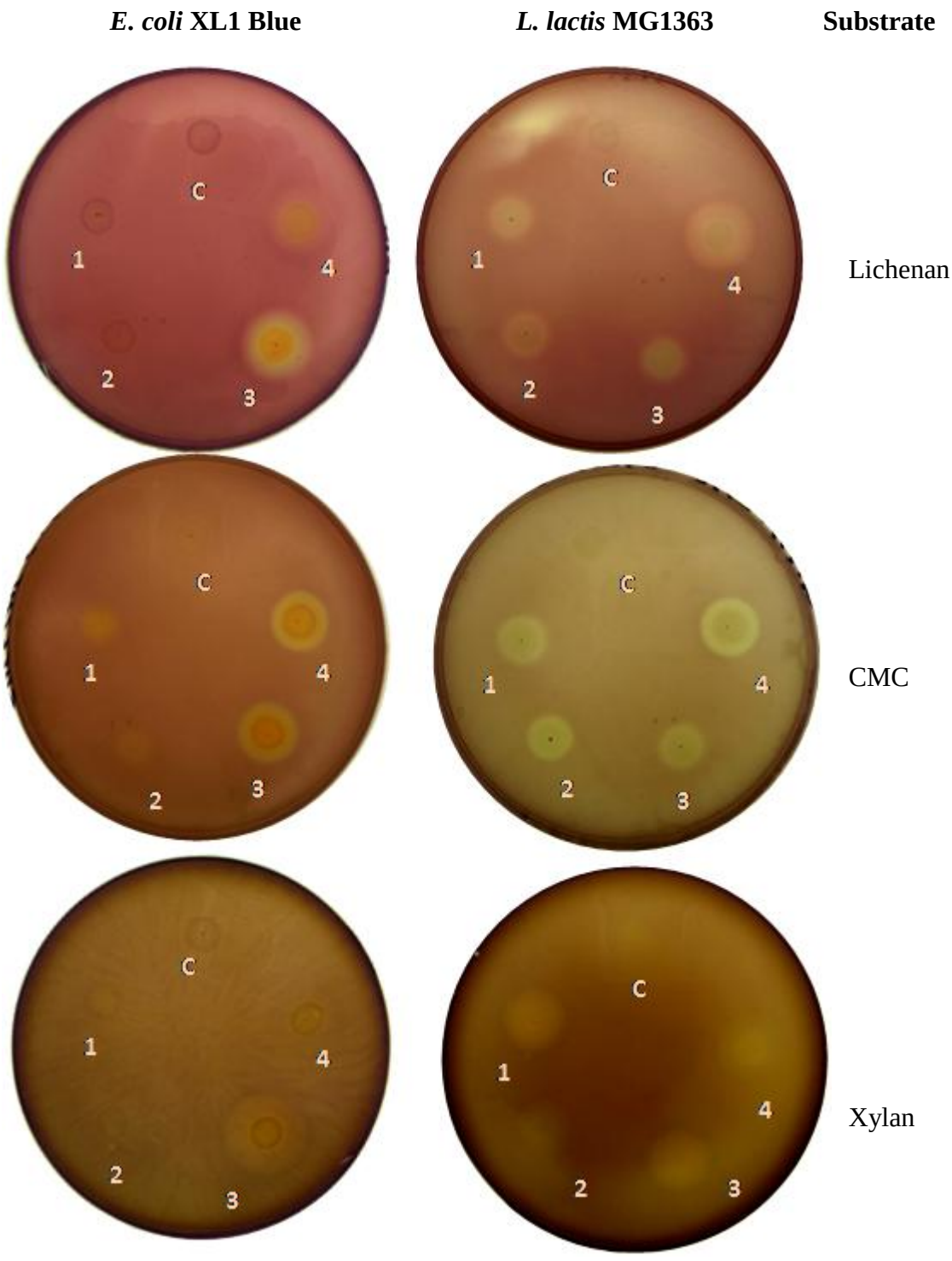


Figure 5.10 Plate test for fibrolytic enzyme activity of *E. coli* and *L. lactis* clones.

A freshly grown overnight culture (10 µl) was pipetted onto the surface of lichenan-, CMC- and xylan-containing plates, allowed to incubate overnight, and stained with Congo Red. In order to increase the contrast, plates were flooded with 50 mM acetic acid to turn the background toward the blue colour instead of pale orange. C- negative control, pTRKL2. 1 – ART_GH9/S, 2 – ART_GH9/L, 3 – L250_GH9/S, 4 – L250_GH9/S.

The plate assay confirmed the expression of the full-length ORF of enzyme ART_GH9/L in *L. lactis* MG1363, but showed limited enzyme activity in *E. coli* XL1 Blue. Differences in other enzyme activities were also observed between the two heterologous hosts. Enzyme ART_GH9/S from *C. eutactus* ART55/1 showed limited carboxymethyl cellulase activity in *E. coli* XL1 Blue. No enzymatic activity was detected on lichenan- and xylan-containing plates. By comparison, detectable enzyme activity was observed on lichenan-, CMC- and xylan- containing plates for *L. lactis* MG1363 clones. The *E. coli* XL1 Blue and *L. lactis* MG1363 clones encoding L250_GH9/S showed clear halos on lichenan-, CMC- and xylan-containing plates, indicating that both hosts functionally expressed the lichenase, carboxymethyl cellulase and xylanase activity. The clearing zone was considerably greater for the *E. coli* construct on lichenan-containing plates. L250_GH9/L showed similar levels of carboxymethyl cellulase activity in both hosts. The xylanase and lichenase activity was greater for the *L. lactis* construct than for the *E. coli* construct. The plate assay indicated that overall both genes from *Coprococcus* sp. L2-50 were functional in both hosts.

5.8 Functional analysis of GH9 enzymes in E. coli *and* L. lactis

In order to assess enzyme activities in both hosts in a quantitative way, reducing sugar assays were performed according to method 2.8.5. Briefly, three independent overnight cultures of *E. coli* (OD_{600} 4.1 - 4.8) and *L. lactis* (OD_{600} 2.0 – 2.7) harbouring the different cloned GH9 genes were used to prepare two fractions: cell-free extract [C] and extracellular supernatant [S] (see method 2.8.4). Enzyme activity was determined by measuring the amount of reducing sugar released by the fractions during incubation with CMC (0.5%) or lichenan (0.5%) as substrates for 2 hours at 37°C. Synergism between the enzymes from *C. eutactus* ART55/1 (ART_GH9/S and ART_GH9/L) and *Coprococcus* sp. L2-50 (L250_GH9/S and L250_GH9/L) was determined by mixing equal volumes of enzyme fractions prepared as described in section 2.8.4.

L250_GH9/S expressed in *E. coli* showed activity on lichenan and CMC as previously observed during the plate assay (Figure 5.11). The lichenase activity was greater than the CMCase activity and was detected in both fractions. The majority of enzyme activity was recovered in the cell free extract. The activity in the supernatant

fraction was 2-fold lower with CMC and 2.5-fold lower with lichenan. Enzyme L250_GH9/L showed greater activity with CMC than lichenan. CMCase activity was mainly detected in the cell-free extract and was 5-fold higher than in the supernatant. Lichenase activity was detected only in the cellular fraction but it was 6-fold lower than the CMCase activity. The results are in agreement to the observations from the semi-quantitative plate assay. They also indicate that the proteins L250_GH9/S and L250_GH9/L were mainly associated with the *E. coli* bacterial cell. The enzyme activity detected in the supernatant fraction could be due to cell lysis (Figure 5.11). The enzyme activity for *L. lactis* cultures was detected mainly in the extracellular fraction (Figure 5.11). In comparison to the L250_GH9/S *E. coli* construct, the enzyme from *L. lactis* was considerably less active. CMCase activity was observed for enzyme L250_GH9/S but no lichenase activity was detected for this enzyme. The L250_GH9/L enzyme showed detectable carboxymethyl cellulase activity in both fractions. Lichenase activity was detected only in the supernatant. The secretion of this protein to the extracellular space indicated the recognition of the predicted signal peptide by the *L. lactis* system.

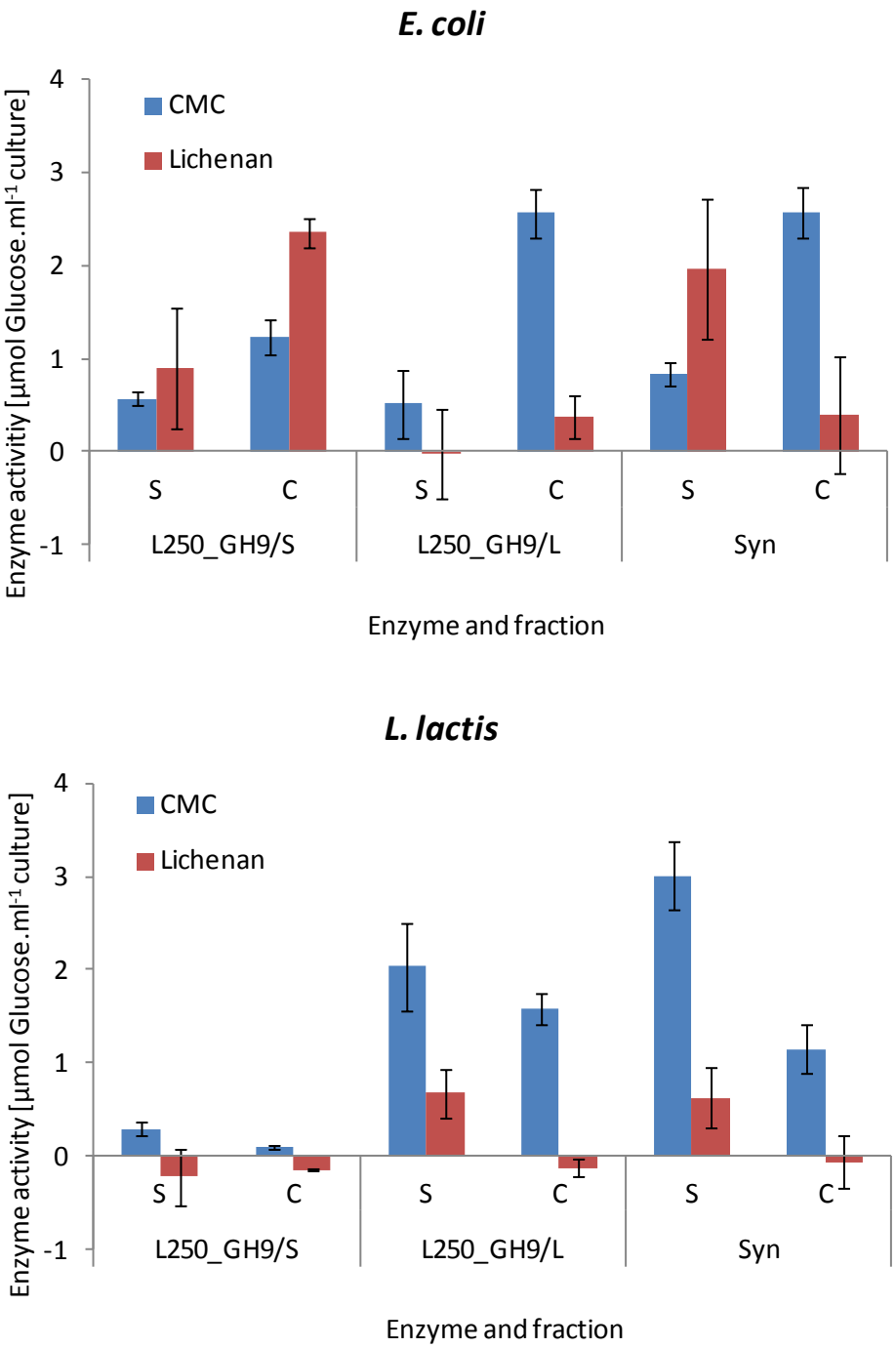


Figure 5.11 Hydrolysis of CMC and lichenan by L250_GH9/S, L250_GH9/L and combined enzyme complex [Syn] expressed by *E. coli* and *L. lactis* culture.

The enzyme activity is expressed as µmol of glucose released by enzymes prepared from one milliliter of *E. coli* or *L. lactis* culture fractionated into supernatant [S] and cell-free extract [C] incubated with 0.5% (w/v) substrate for 2 hours at 37°C. Error bars represent the standard deviation of three replicates.

The enzyme activity of ART_GH9/S and ART_GH9/L was examined only in *L. lactis* (Figure 5.12) since preliminary experiments with the reducing sugar assay showed no assayable activity in *E. coli* cultures (data not shown). This agrees with the plate assay results, which demonstrated limited clearing zones on CMC-containing plates and no hydrolysis zones on lichenan- and xylan-containing plates. Detectable carboxymethyl cellulase activity was observed for the ART_GH9/S and ART_GH9/L enzymes in the activity test (Figure 5.12). The CMCase activities were mainly detected in the supernatant fraction and were much greater than in the cell-free extract, for ART_GH9/S and ART_GH9/L. Low lichenase activity was observed for both enzymes.

A synergistic effect of the enzymes occurs when the combined activity is greater than the sum of the individual enzyme activities. Cellulases from different microbial systems, including different GH families and modes of action were previously shown to act synergistically with increases activity towards recalcitrant substrates (Vazana *et al.,* 2010). During the present study the synergy between coprococcal enzymes was assessed, however there is no conclusive evidence for synergy occurring between examined enzymes (Figure 5.11 and Figure 5.12).

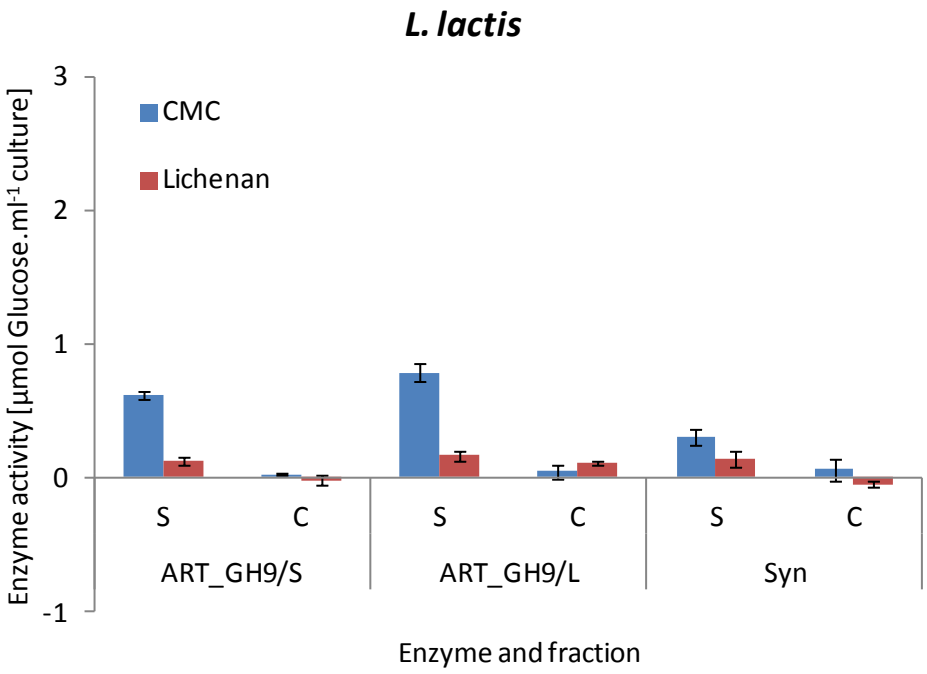


Figure 5.12 Hydrolysis of CMC and lichenan by ART_GH9/S, ART_GH9/L and combined enzyme complex [Syn] expressed by *L. lactis* culture.

The enzyme activity is a µmol of glucose released by enzymes prepared from 1 ml of *L. lactis* culture fractionated into supernatant [S] and cell-free extract [C] incubated with 0.5% (w/v) substrate for 2 hours at 37°C. Error bars represent the standard deviation of three replicates.

5.9 Zymogram analysis of Coprococcus *enzymes*

The enzymatic activities of L250_GH9/S and L250_GH9/L expressed in *E. coli* and *L. lactis* were further confirmed by a zymogram assay. ART_GH9/S and ART_GH9/L enzymes were tested only in *L. lactis*.

The cell-free extract and supernatant fraction was analysed on a CMC-containing SDS-PAGE gel according to method 2.8.6 (Figure 5.13). The estimated molecular weights for the selected proteins from *Coprococcus* strains were calculated based on their sequence. L250_GH9/S and ART_GH9/S were 83 kDa and 82 kDa, respectively. The molecular weight for L250_GH9/L and ART_GH9/L was estimated at 189 kDa and 186 kDa, respectively.

The fractions from *E. coli* (Figure 5.13) showed carboxymethyl cellulase activity bands for proteins L250_GH9/S and L250_GH9/S. In the supernatant fraction of enzyme L250_GH9/S, two bands were visible with a size of around 90 kDa. The third band, which was less visible, was observed about 29 kDa. The strongest band was predicted to be protein L250_GH9/S, since the corresponding band was detected in the cell-free extract and its size matches the estimated molecular weight. The additional bands may be a product of L250_GH9/S degradation (lower band) or protein aggregation (higher band). Protein L250_GH9/L active bands of high MW were observed in both fractions, however the stronger band was observed in the cell-free fraction, in accordance with reducing sugar assay. Several smaller bands were visible on the gel suggesting proteolytic cleavage of the major product.

Analysis of the enzymes expressed in *L. lactis* showed detectable bands only for L250_GH9/L in both fractions with the expected size (Figure 5.14). No bands were observed for L250_GH9/S, ART_GH9/S and ART_GH9/L (data not shown). The CMCase activities of these enzymes expressed in *L. lactis* were considerably lower than CMCase activity of L250_GH9/L, considering the results from the reducing sugar assay (see Figure 5.9 and Figure 5.10). The undetectable activity of these enzymes in *L. lactis* during zymogram analysis could be due to assay condition leading to issues with the proteins renaturation. In contrast to *E. coli*, the proteolytic cleavage is unlikely in this *L. lactis* strain since it is protease negative, therefore proteolytic inactivation is not considered as an issue during this zymogram assay.

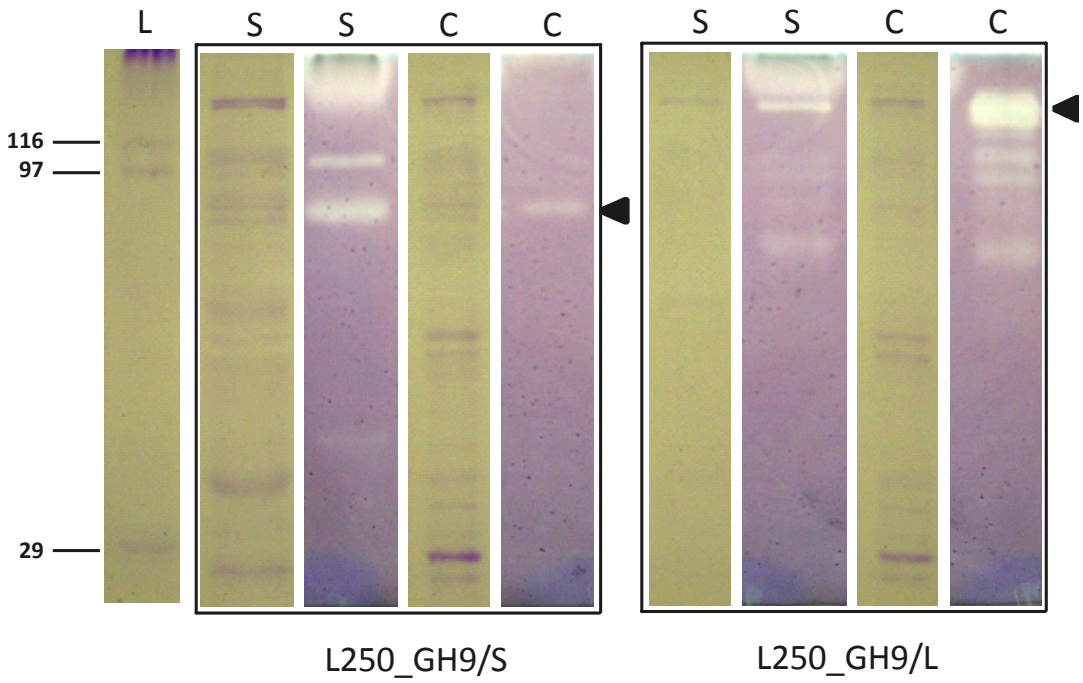


Figure 5.13 Coomassie staining and zymogram on CMC of the coprococcal proteins expressed in *E. coli*.

Arrow – protein bands, S – supernatant, C - cell-free extract, L – ladder [kDa].

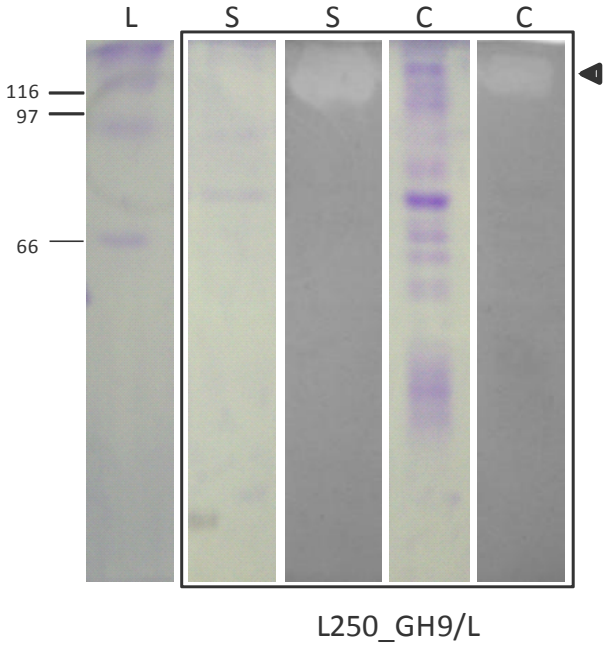


Figure 5.14 Coomassie staining and zymogram on CMC of the coprococcal protein expressed in *L. lactis*

Arrow –L250_GH9/L protein band, S – supernatant, C - cell-free extract fraction. L – ladder [kDa].

5.10 Analysis of possible factors affecting the expression of GH9 enzymes in **E. coli** *and* **L. lactis.**

Differences were observed in measurable enzyme activities between the two heterologous hosts *L. lactis* and *E. coli*. An analysis of the upstream regions of the genes predicted possible promoter recognition sites. Hypothetical promoter sequences are highlighted in the sequence for each GH9 gene from *Coprococcus* strains and are presented in Appendix 3. A transcriptional study is needed to obtain conclusive evidence for the authenticity of these regions. The strength of the promoter depends on the spacer length between the -10 and -35 region, which optimally is 17 ± 2 nucleotides, and the degree of substitutions in the consensus region (Browning and Busby 2004). The predicted promoters for coprococcal genes showed a conserved -10 region which was usually spaced by 17 nucleotides from -35 region. The TG motif, characteristic for genes from Gram-positive bacteria, was also present which could possibly enhance expression of a gene in *L. lactis* (Voskuil and Chambliss 1998). There are also further factors that may be involved in successful transcription of the recombinant genes such as upstream elements and alternative sigma factors (Browning and Busby 2004). The success of heterologous gene expression may be limited by differences in codon usage between the foreign gene and the native system. Analysis of the codon usage in *E. coli* showed infrequent use of some codons for arginine (AGA/AGG, CGA), isoleucine (ATA) and leucine (CTA). A subset of these codons was reported to cause problems during translation by introducing errors leading to frame shifts and a decrease in expression (Jana and Deb 2005). The codon usage pattern of ART_GH9/S and ART_GH9/L showed the presence of these rare codons, indicating potential issues during expression of these enzymes in *E. coli*. In contrast, L250_GH9/S and L250_GH9/L, which were expressed and active in *E. coli*, lacked these rare codons or their frequency was lower (Appendix 3.6, Table 5.3).

The bacterial protein post-translational targeting (cytoplasm vs. extracellular) is dependent on the presence of a signal peptide domain at the N-terminal end of the protein. Enzymes ART_GH9/L and L250_GH9/L showed the presence of a secretory signal peptide with a hydrophobic core. Gram-positive and Gram-negative bacteria produce a signal peptidase type I enzyme, which cleaves the signal peptide (SP), and translocates the unfolded protein through the plasma membrane by Sec-dependent

translocation (Natale *et al.*, 2008). The ART_GH9/L and L250_GH9/L enzymes were found predominantly exported to the supernatant by *L. lactis*, demonstrating the recognition of SP by the host secretory system. In *E. coli* L250_GH9/L was mainly detected in the cell-free extract, which suggests that this protein was not readily exported. The secretory proteins should stay in unfolded confirmation to be readily exported via Sec-dependant pathway. It was hypothesised that coprococcal protein L250_GH9/L expressed in *E. coli*, folded too rapidly in the cytoplasm to avoid degradation and therefore it was not secreted. This phenomenon is a bottleneck of heterologous expression in *E. coli* which led to partial or complete aggregation of the protein as insoluble inclusion bodies (Baneyx and Mujacic 2004). The proteins: ART_GH9/S and L250_GH9/S carry a typical lipoprotein signal peptide. L250_GH9/S expressed in *E. coli* was mainly found in the cell-free extract as previously reported L250_GH9/L, showing limited recognition of a lipoprotein motif by the secretory system of *E. coli*. The *L. lactis*-expressed proteins ART_GH9/S and L250_GH9/S were mainly found in the supernatant fraction which confirms the recognition of the carried signal peptide by the host secretory system.

		ART_GH9/S			ART_GH9/L			L250_GH9/S			L250_GH9/L			*E.coli*		*L. lactis*	
AA	**Code**	No	/1000	F	No	/1000	F	No	/1000	F	No	/1000	F	/1000	F	/1000	F
Arg	**AGA**	11	14.51	1.00	17	9.53	0.65	9	11.80	0.56	12	6.63	0.34	2.72	0.05	10.60	0.29
Ile	**ATA**	19	25.07	0.61	48	26.92	0.50	0	0.00	0.00	14	7.73	0.16	5.34	0.09	15.10	0.20

Table 5.3 Rare codons found in *E. coli*, *L. lactis* and GH9 enzymes from *Coprococcus eutactus* ART55/1 and *Coprococcus* sp. L2-50.

The *E. coli* and *L. lactis* codons usage pattern was obtained from the Codon Usage Database (http://www.kazusa.or.jp/codon/). The GH9 enzyme codon analysis was done by the on-line tool available at: http://www.bioinformatics.org/sms2/index.html. No - total number of occurrences of that codon in the sequence, /1000 –frequency of the codon per 1000 codons in the coding regions of analysed data (coprococcal genes and *E. coli*/ *L. lactis* genomes). F is a frequency of the codon in the set of all codons for the same amino acid. These codons in ART_GH9/S and ART_GH9/L enzymes were shown to be problematic during *E. coli* expression. AA = amino acid, Arg = arginine, Ile = isoleucine. The full list of codon usage in *E. coli* and *L. lactis* can be found in Appendix 3.6.

5.11 Discussion

This chapter described the sequence and functional characteristics of predicted GH9-containing enzymes encoded by coprococcal strains. The results presented in this work indicate that *Coprococcus* strains possess enzymes for the degradation of different polysaccharides including CMC, lichenan and xylan. The modular structure of L250_GH9/S and ART_GH9/S showed similarity to previously examined GH9 enzymes belonging to theme D (group B2) (Devillard *et al.*, 2004). The architecture of ART_GH9/L and L250_GH9/L showed the presence of a GH9 catalytic domain and was classified to newly created theme E. This cluster consists of cellulosomal protein from *Clostridium thermocellum*, *Ruminococcus albus* and *R. flavefaciens*, which show the presence of a dockerin domain at the C-terminal end (Fontes and Gilbert 2010). The characteristic dockerin domain motif with highly conserved calcium-binding residues was not detected in the coprococcal enzymes. Other accessory domains including CBM and Ig-like domain were part of the multi-domain structure of coprococcal proteins ART_GH9/L and L250_GH9/L. Both enzymes showed the presence of a GS-rich region at the N- and C-terminal ends of the proteins. In addition, L250_GH9/L showed the presence of a TP-rich region at the C-terminus of the protein. The GS- and TP-rich domains found in coprococcal proteins might serve as novel carbohydrate binding domains (Ramsay *et al.*, 2006).

The results presented in this work also showed differences in gene expression in *E. coli* XL1 Blue and *L. lactis* MG1363. The genes from *Coprococcus* sp. L2-50 were expressed in both hosts. However, genes from *Coprococcus eutactus* ART55/1 showed detectable enzyme activity only in *L. lactis* MG1363. Previous reports demonstrated the benefits of applying multiple cloning hosts for functional studies since the *E. coli* system was estimated to express only 40% of the *in silico* analysed genes (Gabor *et al.*, 2004, Martinez *et al.*, 2004, Craig *et al.*, 2010). Analysis of codon usage between the surrogate host and the recombinant gene showed the possible impact on expression of GH9 genes in *E. coli*. Differences in codon usage between *E. coli* and ART_GH9/S/ ART_GH9/L genes were observed. The presence of rare arginine and isoleucine codons in the cloned genes may have affected protein accumulation levels and mRNA stability, as reported previously (Baneyx 1999).

The results also demonstrated differences in enzyme localisation between both hosts. The enzyme activity for *E. coli* clones was mainly detected in the cell-free extract,

suggesting that predicted signal peptides for coprococcal proteins were not recognised by the host secretory machinery. In contrast, the lactococcal clones exported the product to the extracellular space similarly to previously reported studies. (Le Loir *et al.*, 2005, Morello *et al.*, 2008).

Chapter 6

General Discussion

6.1 Choice of a heterologous expression host - **Lactococcus lactis *vs.* Escherichia coli**

A previous study showed that alternative heterologous hosts can increase the diversity and efficiency of functional screening of the metagenomic libraries (Wang *et al.,* 2000). It was also estimated that only 40% of genes are readily expressed in *Escherichia coli*. 73% of bacterial genes originating from Firmicutes – the dominant phylum of human gut microbiota – were predicted to be expressed in *E. coli* (Gabor *et al*., 2004). The remaining 27% of genes would not be detected due to failure in gene expression; therefore *Lactococcus lactis* was used as an alternative host during this study. It was hypothesised that genes from gut Firmicutes will be readily expressed in *L. lactis* since the %G+C content is similar. However there are other factors that can limit gene expression including transcription and translation initiation signals, protein-folding, post-translational modification, protein secretion or toxicity of the recombinant gene (Taupp *et al.,* 2011). The results presented in this work showed that *E. coli* had several advantages over *L. lactis* and in the future other Gram-positive Firmicutes host should be considered for the metagenomic screening. The example could be a well-studied *Bacillus subtilis* which has been used for numerous cloning studies (Zweers *et al*., 2008).

The major limitation in preparing the genomic library directly in *L. lactis* MG1363 was the transformation efficiency which was substantially different from *E. coli*. It was shown in this study that the transformation efficiency in *E. coli* was 1000-fold higher than in *L. lactis*. The poor cloning efficiencies in *L. lactis* were observed in a study conducted by Geertsma and Poolman (2007) which led finally to an initial cloning step in *E. coli* and subsequent transfer into *L. lactis*. A similar approach was applied in many other studies that used several different expression hosts for constructing metagenomic libraries (Martinez *et al.,* 2004, Craig *et al.,* 2010). The preparation of highly efficient electrocompetent *L. lactis* cells was shown to produce up to 10^8 CFU.μg^{-1} DNA (Papagianni *et al.,* 2007). However, direct restriction-ligation cloning produced as little as 10^2 CFU.μg^{-1} DNA (Geertsma and Poolman 2007). The metagenomic library in this work also had to be prepared initially in *E. coli,* followed by transfer into the *L. lactis* host.

The efficiency in terms of active clones recovered in the *E. coli* and *L. lactis* metagenomic libraries varied. One positive clone, respectively in *E. coli* and *L. lactis* was detected showing high fibrolytic activity on the CMC-, lichenan and xylan-containing plates. The efficiency for these activities was therefore higher in *L. lactis* as a lower number of clones was screened in this host. However, further fifteen positive clones were detected in the *E. coli* library only. These results indicate that *E. coli* is the more efficient heterologous host for the metagenomic functional study compared to *L. lactis.*

This study showed significant differences in gene expression between the two hosts which depends on several factors including promoter recognition, codon usage, post-translational modification and protein targeting. *L. lactis* was found to be the more stringent host to express genes derived from human gut microbiota which affected the efficiency of functional screening in this host. The different gene expression between the two hosts led to the recovery of metagenomic clone P20A8 exclusively in *L. lactis*. Only limited enzyme activity was observed when the plasmid DNA from this clone was transformed into *E. coli*. Further analysis showed that codon usage possibly affected the expression of the gene in the latter host. No enzyme activity was detected, when the *E. coli* metagenomic clones were transformed into *L. lactis*. Screening of the *Ruminococcus* sp. 80/3 genomic library did not recover positive clones in *L. lactis*, possibly due to the lower number of screened colonies in this host. Several clones from *Ruminococcus* sp. 80/3 with high enzyme activity observed in *E. coli* were not able to replicate in *L. lactis*, possibly by having toxic effects on the host cells. Other studies also reported differences in gene expression in *E. coli* and *L. lactis* showing higher number of prokaryotic proteins being expressed in the former host (Surade *et al.,* 2006, Linares *et al.,* 2010). Among all the metagenomic clones from the *E. coli* screening eight correspond to bacteria from the Bacteroidetes phylum and seven from the phylum Firmicutes. This indicates that genes originating from different bacterial groups were expressed in *E. coli.*

L. lactis was reported to be a widely used heterologous host mainly for industrial applications, which requires secretion of the protein (Morello *et al.,* 2008). Here the differences in protein secretion were noted for coprococcal proteins expressed in *E. coli* or *L. lactis*. The latter host was shown to export coprococcal cellulases whose signal peptides were recognised by the host secretory system. In contrast, coprococcal proteins expressed in *E. coli* were mainly associated with the cell-free

extract. The results also showed that coprococcal proteins expressed in *L. lactis* were protected from the proteolytic degradation, in contrast to *E. coli*-expressed proteins (section 5.9).

In conclusion, whereas both bacteria were reported as successful heterologous hosts in a number of studies (Kunji *et al.,* 2003, Mergulhão *et al.,* 2005, Geertsma and Poolman 2007), in the present study *E. coli* was found to be the better host for the functional screening of genomic and metagenomic libraries, as it led to a higher number of transformants and positive clones. However, one of the most active clones was found only in the alternative host *L. lactis*, which shows that the usage of alternative cloning hosts is a practical strategy to maximise the recovery of positive clones from metagenomic libraries. The substantial differences in gene expression and protein export observed during this study further suggest that the employment of alternative hosts can be useful to functional screening of metagenomic libraries. Therefore employment of other alternative heterologous hosts should be considered for potentially new functional-based studies.

6.2 Functional study of the human gut microbiota

The metabolic activities involved in carbohydrate utilisation of the gut microbiota were examined in the present work by applying functional screening of genomic and metagenomic libraries. The results showed a wide range of enzymes encoded by gut bacteria that enable efficient dietary carbohydrate degradation. These enzymes are necessary for bacterial fermentation which has an impact on human health (Flint *et al.,* 2008). The lack of specific bacterial groups was reported as detrimental for human health and can be associated with obesity, diabetes and other chronic diseases. Absence of *Ruminococcus bromii* in two volunteers was reported by Walker *et al.* (2011), which affected efficient starch degradation in these subjects. *R. bromii* was shown as a highly amylolytic bacterium also in another study (Abell *et al.,* 2008) and it belongs to cluster IV ruminococci, which are primarily colonisers of the particulate fraction in the faecal sample (Walker *et al.,* 2008). In the present study, one metagenomic clone encoded a predicted amylase with similarity to a *R. bromii* protein (P1E14). Clones derived from the *Ruminococcus* sp. 80/3 genomic library showed that this bacterium encodes several glycoside hydrolase enzymes required for dietary fibre degradation. This was in agreement with a previous study

(Wegmann *et al.* unpublished data). *Ruminococcus* sp. 80/3 was shown to utilize complex β-glucan polymers like cellulose, lichenan and xylan (weak activity). A clone encoding a predicted GH5 cellulase was associated with cellulose and lichenan degradation whereas clones encoding enzymes from the GH2 and GH3 family can be involved in the removal of galactose and glucose groups from the non-reducing end of the dietary β-glucan polymers. These results show that cluster IV ruminococci have a wide range of enzymes required for dietary polysaccharides metabolism which allow their colonisation and breakdown of the fibre particles in the gut.

The other group of highly specialised carbohydrate utilisers are Bacteroides species, whose genomes contain many genes associated with carbohydrate degradation (Xu *et al.,* 2003, Xu *et al.,* 2007). In the present study, several clones originating from Bacteroides were detected in the metagenomic library. These clones encode putative pullulanase and an arabinofuranosidase. These enzymes were shown to display multi-domain architecture with carbohydrate-binding modules. It was also noted that the genomic region from which the inserts derived often contain additional GH enzymes. The arabinofuranosidase enzyme from *Bacteroides plebeius* was similar to a *Roseburia intestinalis* gene. According to previous reports different groups of bacteria have adapted to a carbohydrate-rich ecosystem by possessing similar enzyme activities required for their survival. It was also shown that the occurrence of horizontal gene transfer is frequent between bacteria and allows adaptation to the environment by acquiring new features (Lozupone *et al.,* 2008).

The present study emphasised the role for *Coprococcus* strains in carbohydrate utilisation. This group of bacteria belong to the predominant phylotypes (2-4%, Arumugam *et al.,* 2011) from the human gut which was shown to be present in the active fraction of gut bacteria based on 16S rRNA analysis (Peris-Bondia *et al.,* 2011). A highly active fibrolytic clone containing GH9 catalytic domain, initially selected in the *L. lactis* metagenomic library, was likely to originate from *Coprococcus eutactus* ART55/1. The subsequent analysis showed the presence of further GH9 enzymes encoded by this group of gut bacteria. According to the CAZy website, GH9 enzymes are underrepresented in the human gut. The coprococcal GH9 enzymes showed multi-domain architecture and their catalytic modules are related to cluster IV ruminococci GH9 enzymes. A previous study reported that highly active

cellulolytic bacteria isolated from the human gut belong to the cluster IV ruminococci (Robert and Bernalier-Donadille 2003). *Coprococcus* belongs to cluster XIVa which includes highly active xylanolytic (Mirande *et al.*, 2010) and amylolytic bacteria (Ramsay *et al.*, 2006). This is however the first report on the coprococcal ability to utilise various plant cell wall polysaccharides.

This study also confirmed that an expression based approach can provide evidence on the function of hypothetical proteins or predicted proteins with no assigned function. Here, 25% of the predicted ORFs showed homology to hypothetical proteins, which will be further characterised to confirm their activity. These hypothetical proteins originated from predominant groups of the human gut bacteria or novel phylotypes (based on the degree of identity to previously sequenced bacterial genes). Some of the hypothetical genes showed a modular architecture with domains of unknown function, which have been reported for conserved proteins (Mello *et al.*, 2010).

The results presented in this work showed that the gut microbiota encodes a wide range of hydrolytic enzymes from various glycoside hydrolase families. The different enzymatic activities are required to efficiently degrade carbohydrates that reach the colon. The majority of clones encoded potentially novel enzymes previously not associated with carbohydrate metabolism. Thus metagenomics is a powerful tool to a gain better understanding on the metabolic potential of the human gut microbiota.

6.3 Future work

The results described in this work revealed a vast range of research possibilities that would be interesting to pursue and explore in the future.

- Characterisation of novel glycoside hydrolase enzymes. The collection of clones derived from the metagenomic library (category II clones) should be subject to mutagenesis in order to identify the gene responsible for the detected activity. The candidate genes could be over-expressed and the protein purified to allow the biochemical characterisation of the enzyme (including activity assays with different substrates). The novel β-glucosidase enzyme derived from the genomic library of *Ruminococcus* sp. 80/3 should

be further characterised by possibly using different method for over-expression and protein purification. Many of the detected glycoside hydrolase enzymes in this study showed a multi-domain structure. Previous reports established methods to express isolated domains as tagged constructs that allow purification and subsequent biochemical analysis of the domain of interest (Devillard *et al.,* 2004, Xu *et al.,* 2004). Especially, the GH9 enzymes from *Coprococcus* showed a modular architecture which should be subject to binding studies. It is possible that new carbohydrate binding modules are encoded by these enzymes.

- *Coprococcus in vitro* study. The present study showed that *Coprococcus* strains encode glycoside hydrolase enzymes which allow degradation of variety of dietary polysaccharides. A growth study of *Coprococcus* on different substrates should be conducted; the total sugar assay and reducing sugar assay should be performed in order to determine the enzyme activity. The effect of co-culturing of *Coprococcus* strains and other fibrolytic gut bacteria could be examined to determine possible synergy between bacteria during degradation of different plant-derived polysaccharides.
- Screen both metagenomic libraries in *E. coli* and *L. lactis* for further enzymes activities involved in carbohydrate, xenobiotics and lipid metabolism. The functional screening could be complemented with additional enzymes activities such mannosidases, frucotosidases which enrichment was predicted by sequence based analysis (Qin *et al*., 2010; Arumugam *et al*., 2011).

Appendices

Appendix 1

Sequence of genes encoding glycoside hydrolase enzymes detected during functional screening of *Ruminococcus* sp. 80/3 genomic library. Domain search was done by using on-line database PFAM and SMART accessed on 20 July 2011 (Finn *et al.*, 2010, Letunic *et al.*, 2009).

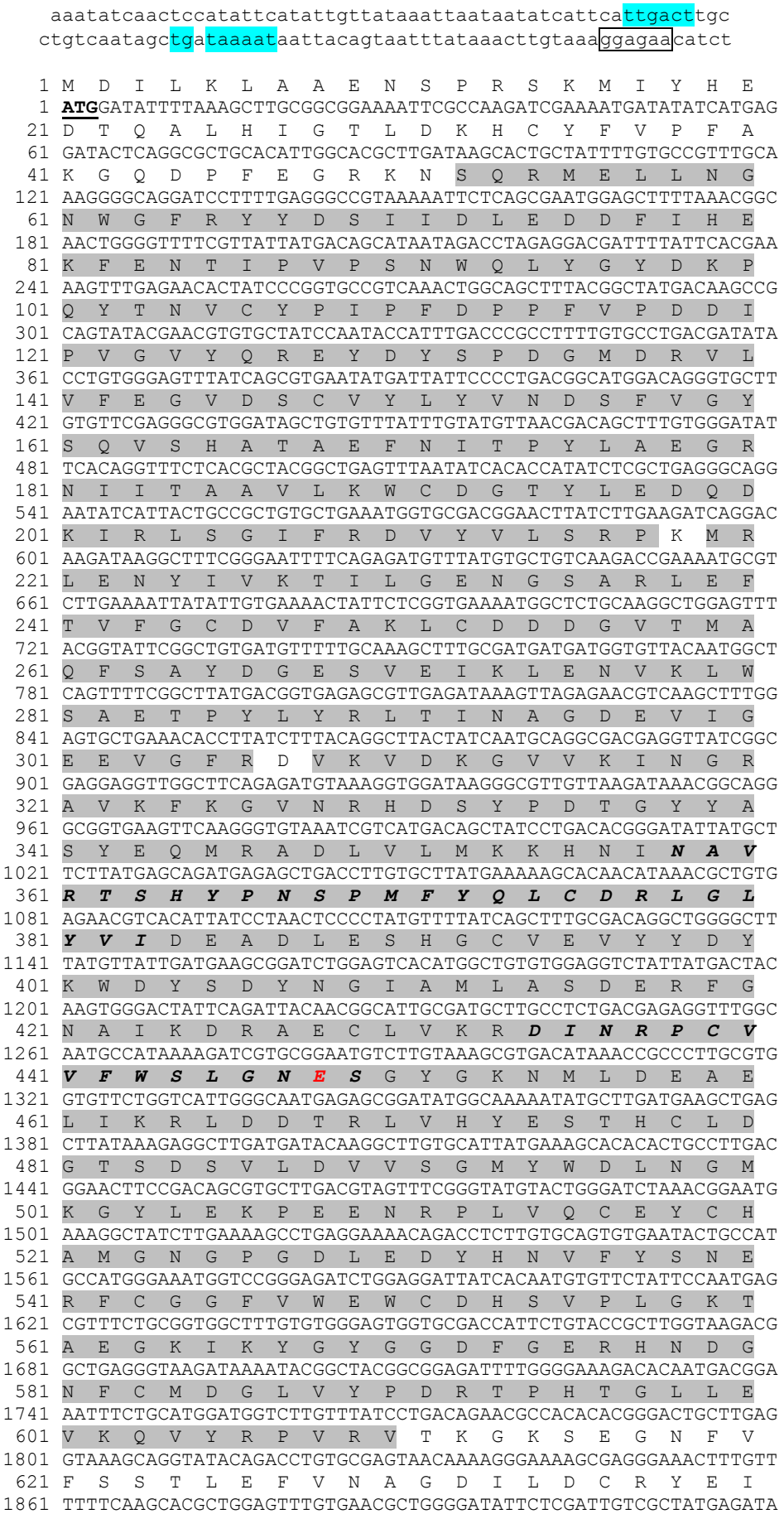
catttattctttctaagatgaatgttgaaaaggcaaatgcggaaatcaaagaaaaact
tggaatatccgaaaatgaagttgtgatggaatcacagaactaaaatggaggaaacaat

1 M K Y T L D W L T D P E V F A V N R I A
1 **ATG**AAATATACACTCGACTGGCTGACCGATCCGGAAGTGTTTGCAGTCAACAGAATCGCA
21 A H S D H R I Y A N H L E A D A D N S S
61 GCACATTCCGACCACAGAATCTATGCAAACCATTTGGAAGCCGATGCAGACAATTCCTCT
41 L I Q N L N G T W K F A W A K N P S E W
121 CTGATTCAGAATCTGAACGGCACCTGGAAATTTGCGTGGGCAAAGAATCCGTCTGAATGG
61 E Q K F Y E E S D S T D N F D N I Q V P
181 GAGCAAAAGTTCTATGAAGAAAGCGACAGCACTGACAATTTTGACAACATTCAGGTTCCG
81 G H M E L Q G Y G K P Q Y V N T I Y P W
241 GGACACATGGAATTGCAGGGATACGGCAAACCGCAGTACGTCAATACCATTTATCCTTGG
101 E G Q E H L R P P F I S E K D N P V G S
301 GAGGGACAGGAGCATCTCCGTCCGCCGTTTATCTCTGAAAAGGACAACCCTGTCGGCAGC
121 Y V R Y F D L N D A L K N K R V F I S F
361 TACGTCAGATATTTTGATTTGAACGATGCTCTGAAAAATAAACGTGTATTCATTTCTTTT
141 Q G V E T A M Y V W L N G E F I G Y S E
421 CAGGGTGTGGAAACTGCTATGTATGTCTGGCTGAACGGCGAATTTATCGGCTACAGCGAG
161 D S F T P S E F E L T P Y I R E K D N K
481 GATTCTTTTACACCGTCTGAATTTGAACTGACACCATATATCCGTGAAAAGGACAACAAA
181 L A V A V F K R S S A S W L E D Q D F W
541 CTTGCTGTTGCAGTGTTCAAAAGAAGTTCTGCAAGCTGGCTGGAAGATCAGGATTTCTGG
201 R F S G I F R D V F L Y A A P S A H V R
601 AGGTTCTCCGGCATTTTCCGTGACGTATTCCTTTATGCGGCACCGTCTGCCCATGTTCGT
221 D M R V I S D Y D G T N G I F S A T L D
661 GATATGAGGGTGATTTCAGATTATGATGGTACAAATGGTATCTTCTCTGCAACGCTGGAT
241 I A G K C S V R S I L T D E N G T V I S
721 ATTGCGGGCAAGTGTTCTGTGAGATCGATCTTGACAGATGAAAATGGAACGGTTATCTCA
261 E S N E K E S V N W T I E N S K P W S A
781 GAATCCAATGAAAAGGAATCTGTAAACTGGACAATTGAAAATAGCAAGCCTTGGAGTGCC
281 E I P N L Y T F T V I L T D E D G N E I
841 GAAATACCGAATCTCTACACGTTTACAGTGATTCTGACAGACGAAGACGGAAATGAAATC
301 E V S R T K V G F R R F E L K N G I M C
901 GAGGTCAGCAGAACCAAAGTCGGATTCCGCCGATTTGAACTGAAAAACGGAATCATGTGT
321 L N G K R I I F K G I N R H E F D A K T
961 TTAAACGGAAAACGTATCATTTTCAAGGGCATCAATCGACACGAATTTGATGCGAAAACA
341 G R A I T K E D M L F D I Q F M K K N N
1021 GGCAGAGCAATCACAAAGGAAGATATGCTGTTTGACATCCAGTTCATGAAGAAAAACAAT
361 I ***N A V R T C H Y P N N S F W Y Q L C D***
1081 ATCAATGCAGTTCGTACCTGTCATTATCCGAACAATTCTTTTTGGTATCAGCTTTGTGAT
381 ***A Y G I Y L I*** D E T N L E T H G T W Q K
1141 GCATATGGCATTTATCTGATTGATGAAACCAATCTGGAAACACACGGTACATGGCAGAAG
401 L G A T D P S W N V P G S L P E W K E A
1201 CTTGGTGCAACAGATCCGTCTTGGAATGTACCTGGTTCACTTCCGGAATGGAAAGAAGCT
421 V L D R A K S M Y E R ***D K N H A S V L I***
1261 GTACTGGACAGAGCAAAATCAATGTATGAACGTGACAAGAATCACGCAAGTGTTCTCATC
441 ***W S C G N E S*** Y C G E G I A A M S E Y F
1321 TGGTCTTGCGGAAACGAATCCTACTGCGGAGAAGGTATTGCTGCAATGTCGGAATATTTC
461 R S V D P T R L V H Y E G V T R V P N H
1381 AGAAGCGTAGACCCGACAAGACTGGTGCACTATGAGGGCGTTACAAGAGTTCCGAATCAT
481 Q Y D S I T D M E S R M Y A K P Q E V E
1441 CAGTATGATTCTATCACGGATATGGAAAGCCGTATGTATGCAAAACCGCAGGAAGTGGAA
501 E Y L K Q N T G R Q Y I S C E Y M H A M
1501 GAATACCTGAAGCAGAATACAGGCAGACAGTATATCAGCTGTGAGTATATGCACGCTATG
521 G N S L G G L S L Y T D L E D K Y E A Y
1561 GGCAATTCTCTTGGCGGTCTGAGCCTTTACACTGACCTTGAGGACAAATATGAGGCTTAT
541 Q G G F I W D Y I D Q A I E T K N D D G
1621 CAGGGTGGTTTCATCTGGGACTACATCGATCAGGCAATCGAAACCAAAAATGACGACGGG
561 K T V L A Y G G D F E D R P S D Y G F C
1681 AAAACGGTTCTGGCTTATGGCGGAGATTTTGAGGACAGACCAAGCGATTACGGATTCTGC
581 T N G V V Y A D R T Y S P K V Q E M K A
1741 ACCAACGGTGTGGTCTATGCCGATCGCACCTATTCTCCAAAGGTTCAGGAAATGAAGGCT
601 L Y S N I R M S I Q N G M L T V E N R N
1801 CTTTACTCCAACATCAGAATGTCTATTCAAAACGGAATGCTTACCGTAGAAAACCGAAAT
621 L F A D T N N L S F V V R L E K N G V V
1861 CTCTTTGCAGATACAAATAACTTGTCCTTTGTTGTACGATTGGAAAAGAATGGCGTTGTT
641 L K S D T F S L N V P A G E S K T M K L

```
1921 CTGAAATCCGATACATTTTCTCTGAATGTTCCTGCCGGAGAAAGCAAGACAATGAAACTG
 661 T  L  H  K  I  D  S  T  G  E  Y  V  Y  H  V  S  A  V  T  A
1981 ACACTTCACAAAATAGACAGTACAGGAGAATATGTTTATCATGTTTCCGCCGTTACCGCA
 681 D  Q  T  L  W  A  D  A  G  H  E  I  A  F  A  Q  E  V  F  E
2041 GATCAGACACTATGGGCAGATGCAGGACACGAAATTGCCTTTGCTCAGGAAGTGTTTGAA
 701 V  K  D  N  Q  I  P  E  T  T  L  Q  K  P  T  I  V  Y  G  D
2101 GTGAAAGATAATCAGATTCCGGAAACAACATTGCAAAAGCCAACGATTGTATACGGCGAT
 721 V  I  I  G  V  H  G  E  N  F  S  M  M  F  D  K  K  E  G  G
2161 GTTATAATTGGCGTTCACGGAGAAAACTTCTCCATGATGTTCGACAAAAAAGAGGGTGGC
 741 I  S  S  L  R  Y  N  G  F  E  Y  I  T  R  T  P  K  V  S  F
2221 ATCAGTTCTCTGAGATACAATGGTTTTGAATATATCACACGCACACCAAAGGTCAGCTTC
 761 W  R  A  M  T  D  N  D  T  G  A  S  E  P  Y  N  L  A  Q  W
2281 TGGAGAGCGATGACAGACAATGACACTGGTGCTTCCGAGCCATACAATCTTGCACAGTGG
 781 Y  S  A  G  K  F  A  K  Y  K  T  V  S  W  L  E  Q  E  D  A
2341 TATTCCGCAGGTAAATTTGCAAAGTATAAAACCGTTTCATGGCTGGAACAGGAAGATGCT
 801 L  K  I  T  F  T  Y  Q  A  A  C  V  P  T  F  E  F  T  V  T
2401 TTGAAAATCACATTCACATATCAGGCTGCCTGTGTTCCGACTTTTGAGTTTACTGTAACC
 821 Y  T  A  H  F  D  G  K  L  G  V  S  V  N  Y  A  G  V  S  G
2461 TATACCGCGCACTTTGACGGAAAGCTGGGCGTAAGTGTAAATTATGCAGGAGTAAGCGGA
 841 M  S  D  M  P  V  L  A  L  D  F  K  M  K  K  Q  L  C  N  F
2521 ATGTCCGATATGCCGGTTCTTGCATTGGACTTCAAGATGAAAAAACAGCTTTGCAATTTT
 861 Q  Y  Y  G  L  G  P  D  E  N  Y  S  D  R  C  K  G  A  R  L
2581 CAGTATTACGGACTCGGCCCTGATGAAAATTACAGCGACCGATGCAAAGGAGCAAGACTG
 881 G  L  W  K  S  T  A  K  E  N  L  S  G  Y  L  N  P  Q  E  C
2641 GGATTGTGGAAATCTACGGCAAAAGAAAATCTTTCCGGATATCTCAATCCTCAGGAATGC
 901 G  N  R  T  G  V  R  T  L  S  V  H  D  D  M  Q  H  G  L  T
2701 GGCAACCGTACCGGTGTAAGAACGCTCTCTGTCCATGATGATATGCAGCATGGTCTGACG
 921 F  Q  K  A  S  A  P  F  E  M  S  V  L  P  Y  S  A  Y  E  L
2761 TTCCAAAAGGCATCTGCTCCGTTTGAAATGAGCGTACTGCCATACAGTGCCTATGAACTC
 941 E  N  A  M  H  L  D  E  L  P  S  V  R  Y  T  W  V  R  I  A
2821 GAAAACGCAATGCATCTGGATGAACTTCCAAGTGTTCGCTACACATGGGTCAGAATCGCT
 961 A  K  Q  M  G  V  G  G  D  D  S  W  G  A  P  V  H  E  E  Y
2881 GCAAAGCAAATGGGTGTCGGCGGCGATGATTCCTGGGGTGCACCGGTGCATGAGGAGTAC
 981 R  I  H  A  D  Q  P  M  K  L  E  F  V  I  M  P  L  R  S  *
2941 AGAATCCATGCAGATCAGCCGATGAAATTGGAATTTGTAATTATGCCCTTAAGATCGTAA
```

Appendix 1.1 Sequence of ORF2 β-galactosidase (3000 bp) from clone pFI2710_3GA, pFI2710_5GA, pFI2710_11GA and pFI2710_12GA. The open reading frame and translation (999 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 2997 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH2 catalytic domain are in italics and bold. The conserved glutamate residue is in red. Domains are indicated as follows:

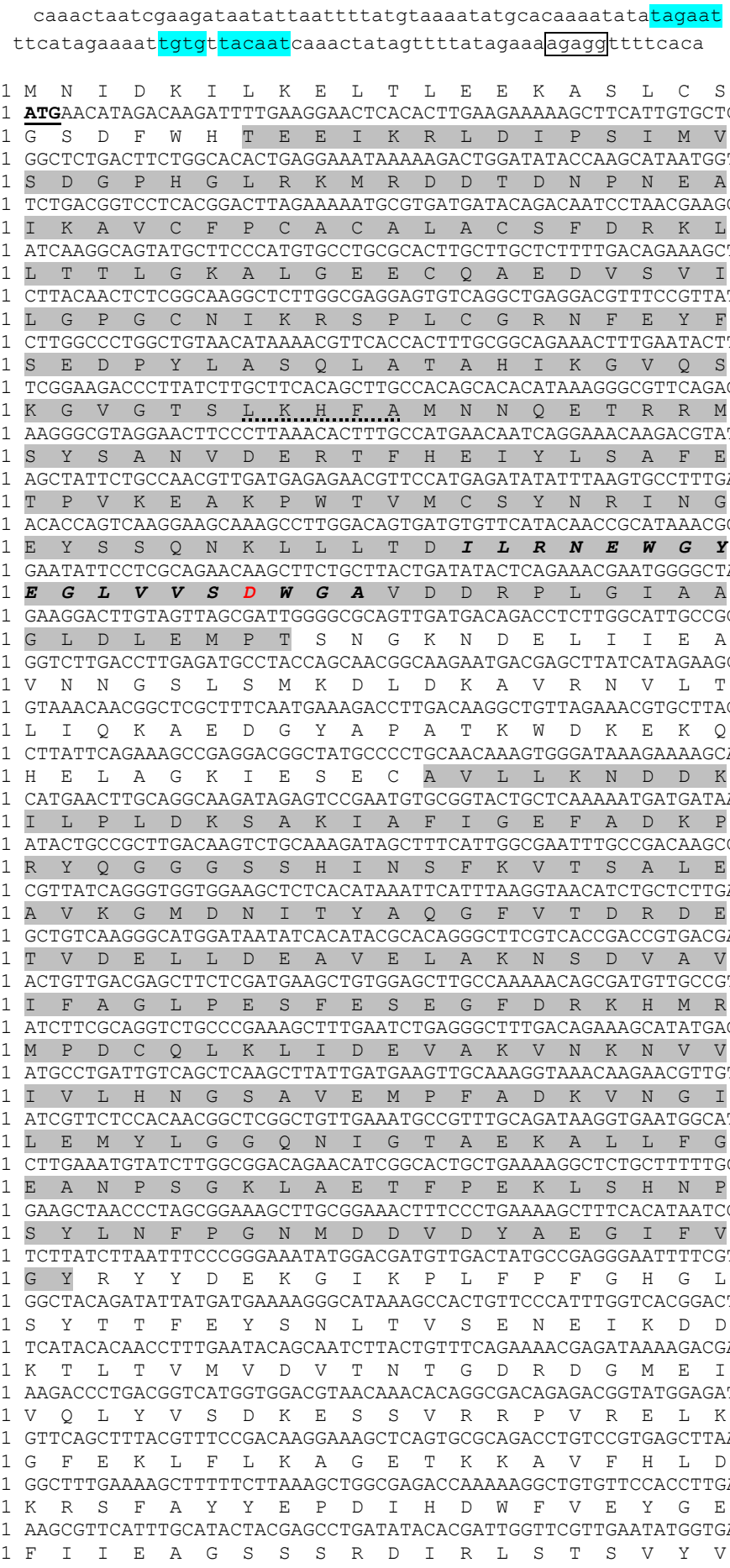
Family 2 catalytic domain: 39-215 – N-terminal sugar binding domain (PF02837), 217-310 – Ig-like β-sandwich domain (PF00703), 312-607 – C-terminal TIM barrel like domain (PF02836), β-galactosidase small chain domain 724-995 (PF02929).


aaatatcaactccatattcatattgttataaattaataatatcattcattgacttgc
ctgtcaatagctgataaaatattacagtaatttataaacttgtaaaggagaacatct
1 M D I L K L A A E N S P R S K M I Y H E
1 ATGGATATTTTAAAGCTTGCGGCGGAAAATTCGCCAAGATCGAAAATGATATATCATGAG
21 D T Q A L H I G T L D K H C Y F V P F A
61 GATACTCAGGCGCTGCACATTGGCACGCTTGATAAGCACTGCTATTTTGTGCCGTTTGCA
41 K G Q D P F E G R K N S Q R M E L L N G
121 AAGGGGCAGGATCCTTTTGAGGGCCGTAAAAATTCTCAGCGAATGGAGCTTTTAAACGGC
61 N W G F R Y Y D S I I D L E D D F I H E
181 AACTGGGGTTTTCGTTATTATGACAGCATAATAGACCTAGAGGACGATTTTATTCACGAA
81 K F E N T I P V P S N W Q L Y G Y D K P
241 AAGTTTGAGAACACTATCCCGGTGCCGTCAAACTGGCAGCTTTACGGCTATGACAAGCCG
101 Q Y T N V C Y P I P F D P P F V P D D I
301 CAGTATACGAACGTGTGCTATCCAATACCATTTGACCCGCCTTTTGTGCCTGACGATATA
121 P V G V Y Q R E Y D Y S P D G M D R V L
361 CCTGTGGGAGTTTATCAGCGTGAATATGATTATTCCCCTGACGGCATGGACAGGGTGCTT
141 V F E G V D S C V Y L Y V N D S F V G Y
421 GTGTTCGAGGGCGTGGATAGCTGTGTTTATTTGTATGTTAACGACAGCTTTGTGGGATAT
161 S Q V S H A T A E F N I T P Y L A E G R
481 TCACAGGTTTCTCACGCTACGGCTGAGTTTAATATCACACCATATCTCGCTGAGGGCAGG
181 N I I T A A V L K W C D G T Y L E D Q D
541 AATATCATTACTGCCGCTGTGCTGAAATGGTGCGACGGAACTTATCTTGAAGATCAGGAC
201 K I R L S G I F R D V Y V L S R P K M R
601 AAGATAAGGCTTTCGGGAATTTTCAGAGATGTTTATGTGCTGTCAAGACCGAAAATGCGT
221 L E N Y I V K T I L G E N G S A R L E F
661 CTTGAAAATTATATTGTGAAAACTATTCTCGGTGAAAATGGCTCTGCAAGGCTGGAGTTT
241 T V F G C D V F A K L C D D D G V T M A
721 ACGGTATTCGGCTGTGATGTTTTTGCAAAGCTTTGCGATGATGATGGTGTTACAATGGCT
261 Q F S A Y D G E S V E I K L E N V K L W
781 CAGTTTTCGGCTTATGACGGTGAGAGCGTTGAGATAAAGTTAGAGAACGTCAAGCTTTGG
281 S A E T P Y L Y R L T I N A G D E V I G
841 AGTGCTGAAACACCTTATCTTTACAGGCTTACTATCAATGCAGGCGACGAGGTTATCGGC
301 E E V G F R D V K V D K G V V K I N G R
901 GAGGAGGTTGGCTTCAGAGATGTAAAGGTGGATAAGGGCGTTGTTAAGATAAACGGCAGG
321 A V K F K G V N R H D S Y P D T G Y Y A
961 GCGGTGAAGTTCAAGGGTGTAAATCGTCATGACAGCTATCCTGACACGGGATATTATGCT
341 S Y E Q M R A D L V L M K K H N I N A V
1021 TCTTATGAGCAGATGAGAGCTGACCTTGTGCTTATGAAAAAGCACAACATAAACGCTGTG
361 R T S H Y P N S P M F Y Q L C D R L G L
1081 AGAACGTCACATTATCCTAACTCCCCTATGTTTTATCAGCTTTGCGACAGGCTGGGGCTT
381 Y V I D E A D L E S H G C V E V Y Y D Y
1141 TATGTTATTGATGAAGCGGATCTGGAGTCACATGGCTGTGTGGAGGTCTATTATGACTAC
401 K W D Y S D Y N G I A M L A S D E R F G
1201 AAGTGGGACTATTCAGATTACAACGGCATTGCGATGCTTGCCTCTGACGAGAGGTTTGGC
421 N A I K D R A E C L V K R D I N R P C V
1261 AATGCCATAAAAGATCGTGCGGAATGTCTTGTAAAGCGTGACATAAACCGCCCTTGCGTG
441 V F W S L G N E S G Y G K N M L D E A E
1321 GTGTTCTGGTCATTGGGCAATGAGAGCGGATATGGCAAAAATATGCTTGATGAAGCTGAG
461 L I K R L D D T R L V H Y E S T H C L D
1381 CTTATAAAGAGGCTTGATGATACAAGGCTTGTGCATTATGAAAGCACACACTGCCTTGAC
481 G T S D S V L D V V S G M Y W D L N G M
1441 GGAACTTCCGACAGCGTGCTTGACGTAGTTTCGGGTATGTACTGGGATCTAAACGGAATG
501 K G Y L E K P E E N R P L V Q C E Y C H
1501 AAAGGCTATCTTGAAAAGCCTGAGGAAAACAGACCTCTTGTGCAGTGTGAATACTGCCAT
521 A M G N G P G D L E D Y H N V F Y S N E
1561 GCCATGGGAAATGGTCCGGGAGATCTGGAGGATTATCACAATGTGTTCTATTCCAATGAG
541 R F C G G F V W E W C D H S V P L G K T
1621 CGTTTCTGCGGTGGCTTTGTGTGGGAGTGGTGCGACCATTCTGTACCGCTTGGTAAGACG
561 A E G K I K Y G Y G G D F G E R H N D G
1681 GCTGAGGGTAAGATAAAATACGGCTACGGCGGAGATTTTGGGGAAAGACACAATGACGGA
581 N F C M D G L V Y P D R T P H T G L L E
1741 AATTTCTGCATGGATGGTCTTGTTTATCCTGACAGAACGCCACACACGGGACTGCTTGAG
601 V K Q V Y R P V R V T K G K S E G N F V
1801 GTAAAGCAGGTATACAGACCTGTGCGAGTAACAAAAGGGAAAAGCGAGGGAAACTTTGTT
621 F S S T L E F V N A G D I L D C R Y E I
1861 TTTTCAAGCACGCTGGAGTTTGTGAACGCTGGGGATATTCTCGATTGTCGCTATGAGATA

```
 641 T  D  K  N  G  I  I  H  I  G  R  V  D  F  D  I  E  P  M  G
1921 ACGGACAAAAACGGTATTATACATATCGGCAGAGTGGATTTTGACATCGAGCCTATGGGA
 661 S  T  V  V  N  V  P  N  D  T  A  E  Y  E  N  E  T  F  I  R
1981 AGCACTGTGGTGAATGTTCCCAATGATACAGCGGAGTATGAAAATGAAACATTCATAAGG
 681 F  I  F  T  A  K  E  D  T  D  Y  C  E  N  G  Y  E  V  C  F
2041 TTTATTTTCACTGCAAAGGAGGACACTGACTATTGTGAAAATGGCTATGAGGTGTGCTTT
 701 D  Q  L  R  I  A  E  G  K  A  C  K  V  A  E  V  K  E  N  V
2101 GATCAGCTTAGGATAGCAGAGGGCAAGGCTTGTAAGGTCGCTGAGGTGAAAGAAAATGTT
 721 E  A  V  E  T  P  L  D  I  T  V  T  V  G  D  V  V  Y  R  F
2161 GAGGCTGTGGAAACGCCTCTTGATATCACTGTGACAGTGGGAGATGTTGTTTACCGCTTT
 741 D  K  R  K  S  A  F  V  S  V  K  F  G  G  R  E  L  L  D  R
2221 GACAAGCGTAAGTCAGCTTTTGTATCTGTGAAATTCGGCGGCAGAGAGCTTCTTGACAGA
 761 P  L  Q  Y  N  F  F  R  A  P  T  D  N  D  V  M  K  C  N  W
2281 CCTTTGCAGTATAACTTCTTCCGCGCGCCTACGGATAATGATGTCATGAAGTGCAATTGG
 781 Y  K  V  H  L  N  D  F  D  V  K  S  Y  G  C  E  L  S  A  G
2341 TACAAGGTGCATCTTAATGACTTTGATGTGAAAAGCTACGGCTGTGAGCTTTCAGCGGGC
 801 E  N  R  A  E  I  S  V  T  Q  S  F  G  W  S  I  Q  Q  P  F
2401 GAGAACAGAGCGGAGATAAGCGTCACACAGTCTTTCGGCTGGTCCATACAACAGCCGTTT
 821 C  R  L  K  A  V  Y  V  I  D  G  S  G  L  D  I  K  C  E  A
2461 TGCAGGCTGAAAGCTGTTTATGTTATTGACGGCAGTGGGCTTGACATAAAATGCGAAGCG
 841 E  F  S  N  K  I  D  M  I  P  R  F  G  I  R  L  F  M  P  K
2521 GAGTTTTCAAATAAGATAGATATGATTCCAAGGTTCGGTATAAGGCTTTTCATGCCAAAA
 861 D  Y  S  R  A  E  Y  F  G  Y  G  P  T  E  S  Y  I  D  K  R
2581 GATTATTCAAGGGCGGAATATTTCGGATATGGTCCTACAGAAAGCTACATTGACAAGCGT
 881 Q  A  C  Y  M  G  R  F  A  A  D  I  G  D  M  H  E  D  Y  I
2641 CAGGCTTGCTATATGGGCAGATTTGCGGCTGATATTGGGGATATGCACGAGGACTATATC
 901 R  P  Q  E  N  S  S  H  Y  G  C  R  Y  L  T  V  S  S  E  D
2701 AGACCGCAGGAAAATTCATCACACTATGGCTGTCGATATCTCACTGTGAGCAGTGAGGAC
 921 T  S  V  M  F  T  A  G  E  E  F  S  F  N  A  S  Q  F  T  Q
2761 ACAAGTGTGATGTTCACTGCGGGAGAGGAGTTTTCATTCAATGCGTCGCAGTTCACGCAG
 941 E  E  L  A  A  K  A  H  N  Y  E  L  E  R  C  E  S  N  V  I
2821 GAGGAGCTTGCGGCAAAGGCACATAACTATGAGCTTGAAAGGTGCGAGAGCAATGTTATC
 961 C  V  D  Y  A  M  A  G  V  G  S  N  S  C  G  P  R  L  A  E
2881 TGCGTTGACTATGCAATGGCAGGGGTAGGTTCTAACTCCTGCGGACCAAGGCTTGCGGAG
 981 K  Y  Q  L  E  K  P  L  I  N  A  H  F  H  I  Q  I  S  K  I
2941 AAATATCAGCTTGAAAAGCCTTTGATAAACGCACATTTTCATATTCAGATAAGTAAAATA
1001 *
3001 TAA
```

Appendix 1.2 Sequence of ORF1, β-galactosidase (3003 bp) from clone pFI2710_4GA. The open reading frame and translation (1000 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 3001 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH2 catalytic domain are in italics and bold. The conserved glutamate residue is in red. Domains are indicated as follows:

Family 2 catalytic domain: 52-217 – N-terminal sugar binding domain (PF02837), 219-306 – Ig-like β-sandwich domain (PF00703), 308-610 – C-terminal TIM barrel like domain (PF02836), β-galactosidase small chain domain 730-996 (PF02929).


caaactaatcgaagataatattaattttatgtaaaatatgcacaaaatatatagaat
ttcatagaaaattgtgttacaatcaaactatagttttatagaaaagaggttttcaca
1 M N I D K I L K E L T L E E K A S L C S
1 ATGAACATAGACAAGATTTTGAAGGAACTCACACTTGAAGAAAAAGCTTCATTGTGCTCA
21 G S D F W H T E E I K R L D I P S I M V
61 GGCTCTGACTTCTGGCACACTGAGGAAATAAAAAGACTGGATATACCAAGCATAATGGTG
41 S D G P H G L R K M R D D T D N P N E A
121 TCTGACGGTCCTCACGGACTTAGAAAAATGCGTGATGATACAGACAATCCTAACGAAGCT
61 I K A V C F P C A C A L A C S F D R K L
181 ATCAAGGCAGTATGCTTCCCATGTGCCTGCGCACTTGCTTGCTCTTTTGACAGAAAGCTG
81 L T T L G K A L G E E C Q A E D V S V I
241 CTTACAACTCTCGGCAAGGCTCTTGGCGAGGAGTGTCAGGCTGAGGACGTTTCCGTTATA
101 L G P G C N I K R S P L C G R N F E Y F
301 CTTGGCCCTGGCTGTAACATAAAACGTTCACCACTTTGCGGCAGAAACTTTGAATACTTC
121 S E D P Y L A S Q L A T A H I K G V Q S
361 TCGGAAGACCCTTATCTTGCTTCACAGCTTGCCACAGCACACATAAAGGGCGTTCAGAGC
141 K G V G T S L K H F A M N N Q E T R R M
421 AAGGGCGTAGGAACTTCCCTTAAACACTTTGCCATGAACAATCAGGAAACAAGACGTATG
161 S Y S A N V D E R T F H E I Y L S A F E
481 AGCTATTCTGCCAACGTTGATGAGAGAACGTTCCATGAGATATATTTAAGTGCCTTTGAA
181 T P V K E A K P W T V M C S Y N R I N G
541 ACACCAGTCAAGGAAGCAAAGCCTTGGACAGTGATGTGTTCATACAACCGCATAAACGGT
201 E Y S S Q N K L L L T D I L R N E W G Y
601 GAATATTCCTCGCAGAACAAGCTTCTGCTTACTGATATACTCAGAAACGAATGGGGCTAT
221 E G L V V S D W G A V D D R P L G I A A
661 GAAGGACTTGTAGTTAGCGATTGGGGCGCAGTTGATGACAGACCTCTTGGCATTGCCGCA
241 G L D L E M P T S N G K N D E L I I E A
721 GGTCTTGACCTTGAGATGCCTACCAGCAACGGCAAGAATGACGAGCTTATCATAGAAGCC
261 V N N G S L S M K D L D K A V R N V L T
781 GTAAACAACGGCTCGCTTTCAATGAAAGACCTTGACAAGGCTGTTAGAAACGTGCTTACG
281 L I Q K A E D G Y A P A T K W D K E K Q
841 CTTATTCAGAAAGCCGAGGACGGCTATGCCCCTGCAACAAAGTGGGATAAAGAAAAGCAG
301 H E L A G K I E S E C A V L L K N D D K
901 CATGAACTTGCAGGCAAGATAGAGTCCGAATGTGCGGTACTGCTCAAAAATGATGATAAG
321 I L P L D K S A K I A F I G E F A D K P
961 ATACTGCCGCTTGACAAGTCTGCAAAGATAGCTTTCATTGGCGAATTTGCCGACAAGCCA
341 R Y Q G G G S S H I N S F K V T S A L E
1021 CGTTATCAGGGTGGTGGAAGCTCTCACATAAATTCATTTAAGGTAACATCTGCTCTTGAA
361 A V K G M D N I T Y A Q G F V T D R D E
1081 GCTGTCAAGGGCATGGATAATATCACATACGCACAGGGCTTCGTCACCGACCGTGACGAA
381 T V D E L L D E A V E L A K N S D V A V
1141 ACTGTTGACGAGCTTCTCGATGAAGCTGTGGAGCTTGCCAAAAACAGCGATGTTGCCGTT
401 I F A G L P E S F E S E G F D R K H M R
1201 ATCTTCGCAGGTCTGCCCGAAAGCTTTGAATCTGAGGGCTTTGACAGAAAGCATATGAGA
421 M P D C Q L K L I D E V A K V N K N V V
1261 ATGCCTGATTGTCAGCTCAAGCTTATTGATGAAGTTGCAAAGGTAAACAAGAACGTTGTT
441 I V L H N G S A V E M P F A D K V N G I
1321 ATCGTTCTCCACAACGGCTCGGCTGTTGAAATGCCGTTTGCAGATAAGGTGAATGGCATA
461 L E M Y L G G Q N I G T A E K A L L F G
1381 CTTGAAATGTATCTTGGCGGACAGAACATCGGCACTGCTGAAAAGGCTCTGCTTTTTGGC
481 E A N P S G K L A E T F P E K L S H N P
1441 GAAGCTAACCCTAGCGGAAAGCTTGCGGAAACTTTCCCTGAAAAGCTTTCACATAATCCG
501 S Y L N F P G N M D D V D Y A E G I F V
1501 TCTTATCTTAATTTCCCGGGAAATATGGACGATGTTGACTATGCCGAGGGAATTTTCGTG
521 G Y R Y Y D E K G I K P L F P F G H G L
1561 GGCTACAGATATTATGATGAAAAGGGCATAAAGCCACTGTTCCCATTTGGTCACGGACTT
541 S Y T T F E Y S N L T V S E N E I K D D
1621 TCATACACAACCTTTGAATACAGCAATCTTACTGTTTCAGAAAACGAGATAAAAGACGAC
561 K T L T V M V D V T N T G D R D G M E I
1681 AAGACCCTGACGGTCATGGTGGACGTAACAAACACAGGCGACAGAGACGGTATGGAGATC
581 V Q L Y V S D K E S S V R R P V R E L K
1741 GTTCAGCTTTACGTTTCCGACAAGGAAAGCTCAGTGCGCAGACCTGTCCGTGAGCTTAAG
601 G F E K L F L K A G E T K K A V F H L D
1801 GGCTTTGAAAAGCTTTTTCTTAAAGCTGGCGAGACCAAAAAGGCTGTGTTCCACCTTGAT
621 K R S F A Y Y E P D I H D W F V E Y G E
1861 AAGCGTTCATTTGCATACTACGAGCCTGATATACACGATTGGTTCGTTGAATATGGTGAG
641 F I I E A G S S S R D I R L S T S V Y V

```
1921 TTTATTATCGAGGCAGGCTCTTCTTCAAGAGATATCAGACTTTCGACTTCTGTTTATGTA
 661 S  S  D  T  K  L  P  V  H  F  T  L  N  T  T  C  G  E  I  N
1981 TCTTCCGATACAAAACTCCCAGTGCATTTCACCTTGAACACCACCTGTGGCGAGATAAAT
 681 S  I  P  E  G  R  A  M  F  E  N  I  L  S  K  I  D  C  D  F
2041 TCTATCCCAGAGGGCAGAGCTATGTTCGAGAACATTTTAAGCAAAATAGACTGCGATTTT
 701 G  D  T  A  S  D  D  L  G  A  S  A  K  E  M  M  E  A  M  I
2101 GGCGACACGGCTTCCGACGATCTTGGCGCTTCTGCAAAGGAAATGATGGAAGCAATGATA
 721 R  D  M  P  L  R  T  L  V  T  F  T  N  V  P  D  I  T  R  A
2161 CGAGATATGCCACTGAGAACCCTTGTTACATTTACAAATGTTCCAGACATCACAAGAGCG
 741 K  M  S  E  M  V  E  N  L  N  E  M  L  A  E  K  *
2221 AAGATGTCAGAAATGGTGGAAAATCTCAACGAAATGCTTGCAGAAAAATAA
```

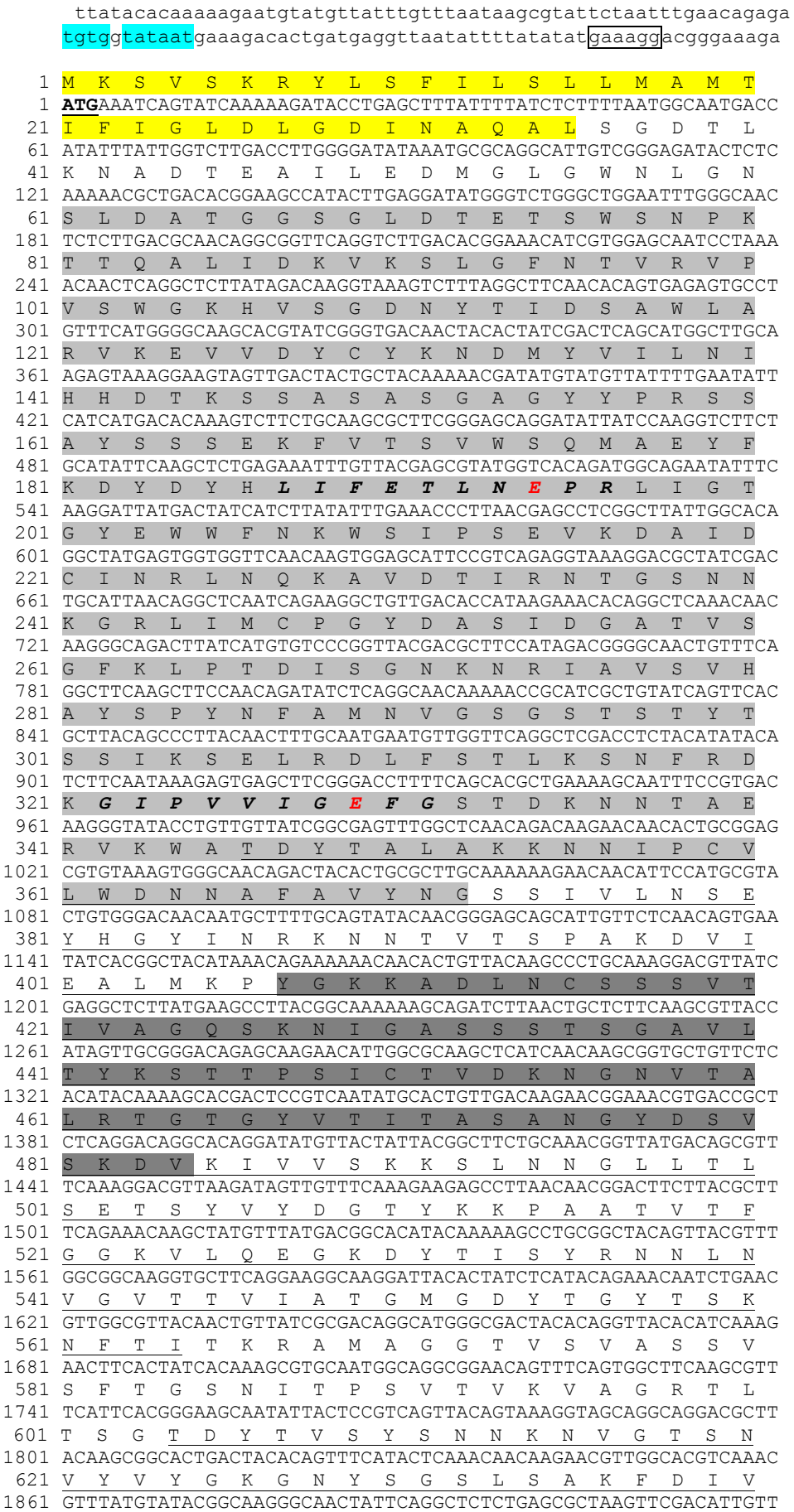
Appendix 1.3 Sequence of ORF3, β-glucosidase (2271 bp) from clones pFI2710_3GL, pFI2170_4GL and pFI2710_7GL. The open reading frame and translation (756 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 2269 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue. Conserved sequence containing a putative carbohydrate binding site is indicated by broken underlining. Signature sequence for GH3 catalytic domain is in italics and bold. Conserved aspartic acid residue is in red. Conserved COOH-terminal antiparallel loop sequence is in yellow. Domains are indicated as follows:

Family 3 catalytic domains: 27-248 – N-terminal domain (PF00933), 312-522 – C-terminal domain (PF01915).

```
      CACCCAGTTTAATAACAATTATCCCACAGTTTCGGCTGTGGGATTTTTTTGTTTACGGCT
      CTTGAAAAATTTGTACATTTGTGCTATAATTAATATACATTTTTTCGGGAGGACGCCAAG

    1 M  N  K  K  E  I  N  E  I  K  K  N  F  N  D  D  C  G  F  F
    1 ATGAACAAAAAGGAAATAAATGAAATAAAGAAAAACTTTAATGACGATTGCGGCTTTTTT
   21 T  V  N  H  V  V  T  A  F  V  D  A  E  K  N  I  K  C  K  T
   61 ACCGTAAACCACGTTGTTACGGCATTTGTGGACGCTGAAAAGAACATAAAGTGCAAGACC
   41 N  Q  L  Y  N  T  I  P  Q  D  E  A  E  L  I  M  I  N  L  K
  121 AATCAGCTTTACAATACTATTCCGCAGGACGAGGCGGAGCTGATAATGATAAACCTGAAA
   61 K  V  L  S  G  S  I  G  K  N  L  L  E  Y  S  F  P  K  D  A
  181 AAGGTGCTCAGCGGCTCTATCGGTAAAAATCTGCTGGAATATTCGTTTCCTAAGGACGCC
   81 Y  L  E  G  G  A  Q  P  F  M  Y  E  T  L  Q  S  K  L  L  D
  241 TACCTTGAGGGCGGTGCTCAGCCTTTCATGTACGAAACACTGCAAAGCAAGCTGCTTGAT
  101 E  E  K  V  D  N  F  L  N  A  I  V  E  K  V  E  Y  V  S  T
  301 GAAGAAAAGGTGGACAACTTCCTGAACGCTATAGTTGAAAAGGTTGAGTATGTGTCAACA
  121 Y  T  I  F  A  A  H  C  T  Y  S  V  L  R  K  N  K  M  D  E
  361 TATACCATTTTTGCGGCACACTGTACATATTCCGTGCTGAGGAAGAACAAAATGGACGAG
  141 F  E  D  E  A  D  T  D  Y  N  F  I  V  T  A  L  C  P  V  N
  421 TTTGAGGACGAAGCTGACACAGATTACAATTTTATAGTGACAGCTCTTTGCCCTGTAAAT
  161 L  R  I  D  G  L  V  Y  D  E  Q  D  N  S  I  A  K  K  E  S
  481 CTGCGTATTGACGGACTTGTGTATGATGAACAGGACAACTCTATTGCAAAGAAAGAGTCA
  181 C  D  R  I  V  E  L  P  S  D  G  F  L  F  P  L  F  N  D  R
  541 TGTGATAGAATTGTTGAACTGCCAAGTGACGGCTTCCTGTTTCCTCTTTTCAATGATCGC
  201 A  P  D  I  N  G  V  L  Y  Y  T  K  N  A  K  K  P  N  T  S
  601 GCACCTGATATCAACGGAGTGCTTTACTACACGAAAAACGCAAAAAAGCCGAACACTTCC
  221 V  V  E  E  L  L  G  C  E  F  S  M  T  C  Q  N  E  K  E  T
  661 GTTGTGGAAGAGCTTCTGGGTTGTGAGTTTTCAATGACCTGTCAGAACGAAAAGGAAACT
  241 F  K  D  I  L  T  S  V  V  G  D  E  L  D  Y  D  L  I  T  T
  721 TTCAAGGATATCCTCACAAGCGTTGTGGGTGATGAGCTTGACTATGATCTTATCACTACT
  261 V  N  D  K  I  S  T  F  V  D  Q  N  A  H  E  T  E  I  P  T
  781 GTGAATGACAAGATTTCCACATTTGTTGACCAAAATGCTCATGAAACTGAGATACCGACA
  281 I  D  E  H  K  L  S  S  I  L  W  E  A  G  V  S  Q  D  K  L
  841 ATTGACGAACATAAACTTTCGTCCATTCTGTGGGAAGCTGGAGTTAGTCAGGATAAGCTT
  301 D  N  L  P  K  V  F  D  A  A  A  G  G  K  P  L  T  A  V  N
  901 GATAATCTTCCTAAAGTTTTTGACGCTGCGGCTGGAGGGAAGCCACTCACAGCGGTAAAC
  321 L  V  E  K  R  T  V  V  S  A  P  S  I  T  V  N  I  G  K  D
  961 CTTGTGGAGAAAAGGACCGTAGTTTCTGCTCCAAGCATAACCGTAAATATTGGCAAGGAC
  341 A  V  D  K  V  K  T  Q  I  V  D  G  R  K  C  L  I  I  N  L
 1021 GCTGTTGACAAGGTAAAAACTCAGATAGTCGATGGCAGAAAGTGCCTTATAATCAATCTT
  361 D  D  P  E  I  E  V  N  G  L  E  T  T  I  K  *
 1081 GATGACCCTGAGATAGAGGTCAACGGACTTGAAACTACAATAAAATAA
```

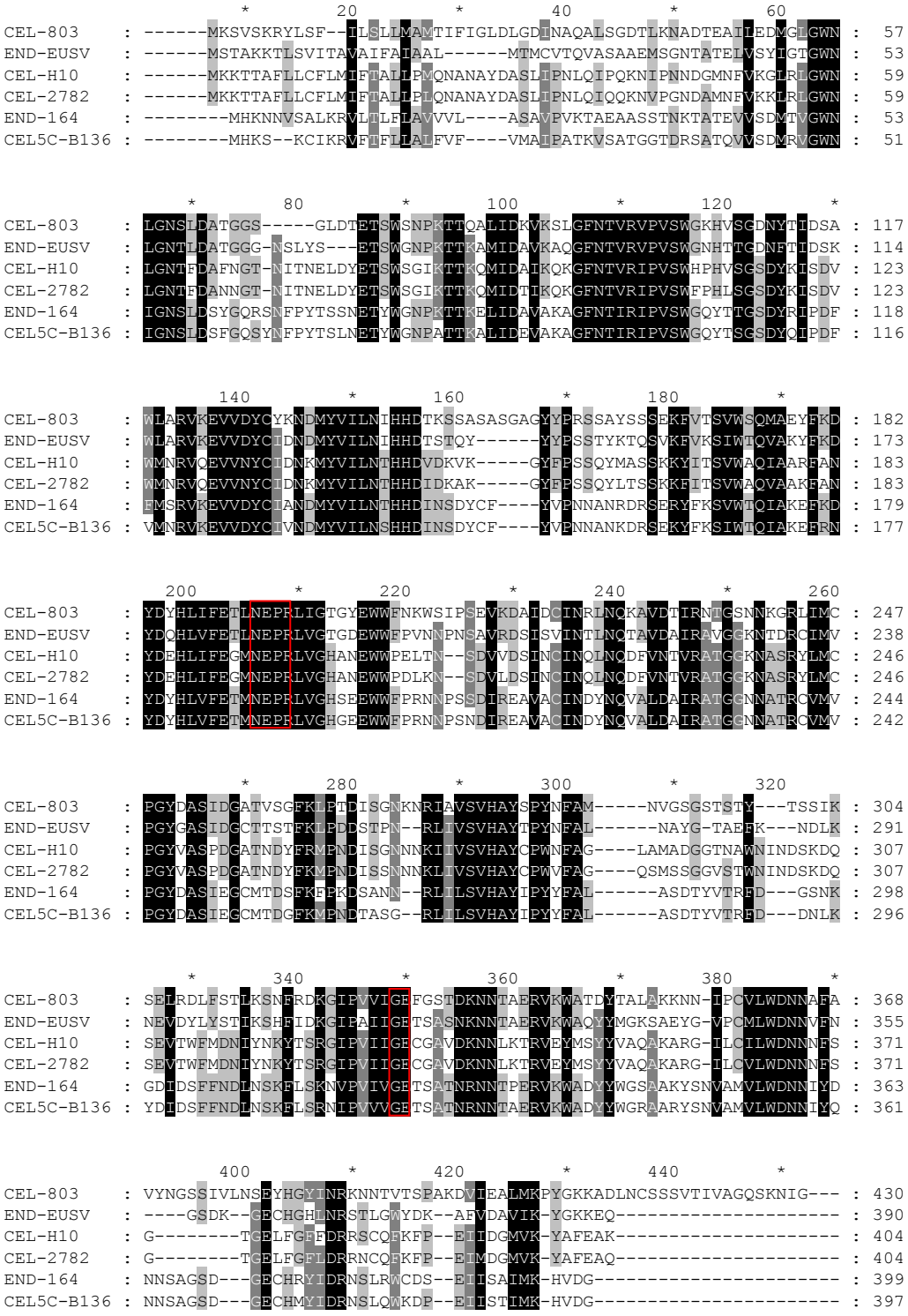
Appendix 1.4 Sequence of ORF4, novel β-glucosidase (1128 bp) from clones pFI2710_1 and pFI2710_12GL. The open reading frame and translation (375 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 1125 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue.


ttatacacaaaaagaatgtatgttatttgtttaataagcgtattctaatttgaacagaga
tgtggtataatgaaagacactgatgaggttaatattttatatatgaaaggacgggaaaga
1 M K S V S K R Y L S F I L S L L M A M T
1 ATGAAATCAGTATCAAAAAGATACCTGAGCTTTATTTTATCTCTTTTAATGGCAATGACC
21 I F I G L D L G D I N A Q A L S G D T L
61 ATATTTATTGGTCTTGACCTTGGGGATATAAATGCGCAGGCATTGTCGGGAGATACTCTC
41 K N A D T E A I L E D M G L G W N L G N
121 AAAAACGCTGACACGGAAGCCATACTTGAGGATATGGGTCTGGGCTGGAATTTGGGCAAC
61 S L D A T G G S G L D T E T S W S N P K
181 TCTCTTGACGCAACAGGCGGTTCAGGTCTTGACACGGAAACATCGTGGAGCAATCCTAAA
81 T T Q A L I D K V K S L G F N T V R V P
241 ACAACTCAGGCTCTTATAGACAAGGTAAAGTCTTTAGGCTTCAACACAGTGAGAGTGCCT
101 V S W G K H V S G D N Y T I D S A W L A
301 GTTTCATGGGGCAAGCACGTATCGGGTGACAACTACACTATCGACTCAGCATGGCTTGCA
121 R V K E V V D Y C Y K N D M Y V I L N I
361 AGAGTAAAGGAAGTAGTTGACTACTGCTACAAAAACGATATGTATGTTATTTTGAATATT
141 H H D T K S S A S A S G A G Y Y P R S S
421 CATCATGACACAAAGTCTTCTGCAAGCGCTTCGGGAGCAGGATATTATCCAAGGTCTTCT
161 A Y S S S E K F V T S V W S Q M A E Y F
481 GCATATTCAAGCTCTGAGAAATTTGTTACGAGCGTATGGTCACAGATGGCAGAATATTTC
181 K D Y D Y H L I F E T L N E P R L I G T
541 AAGGATTATGACTATCATCTTATATTTGAAACCCTTAACGAGCCTCGGCTTATTGGCACA
201 G Y E W W F N K W S I P S E V K D A I D
601 GGCTATGAGTGGTGGTTCAACAAGTGGAGCATTCCGTCAGAGGTAAAGGACGCTATCGAC
221 C I N R L N Q K A V D T I R N T G S N N
661 TGCATTAACAGGCTCAATCAGAAGGCTGTTGACACCATAAGAAACACAGGCTCAAACAAC
241 K G R L I M C P G Y D A S I D G A T V S
721 AAGGGCAGACTTATCATGTGTCCCGGTTACGACGCTTCCATAGACGGGGCAACTGTTTCA
261 G F K L P T D I S G N K N R I A V S V H
781 GGCTTCAAGCTTCCAACAGATATCTCAGGCAACAAAAACCGCATCGCTGTATCAGTTCAC
281 A Y S P Y N F A M N V G S G S T S T Y T
841 GCTTACAGCCCTTACAACTTTGCAATGAATGTTGGTTCAGGCTCGACCTCTACATATACA
301 S S I K S E L R D L F S T L K S N F R D
901 TCTTCAATAAAGAGTGAGCTTCGGGACCTTTTCAGCACGCTGAAAAGCAATTTCCGTGAC
321 K G I P V V I G E F G S T D K N N T A E
961 AAGGGTATACCTGTTGTTATCGGCGAGTTTGGCTCAACAGACAAGAACAACACTGCGGAG
341 R V K W A T D Y T A L A K K N N I P C V
1021 CGTGTAAAGTGGGCAACAGACTACACTGCGCTTGCAAAAAAGAACAACATTCCATGCGTA
361 L W D N N A F A V Y N G S S I V L N S E
1081 CTGTGGGACAACAATGCTTTTGCAGTATACAACGGGAGCAGCATTGTTCTCAACAGTGAA
381 Y H G Y I N R K N N T V T S P A K D V I
1141 TATCACGGCTACATAAACAGAAAAAACAACACTGTTACAAGCCCTGCAAAGGACGTTATC
401 E A L M K P Y G K K A D L N C S S S V T
1201 GAGGCTCTTATGAAGCCTTACGGCAAAAAAGCAGATCTTAACTGCTCTTCAAGCGTTACC
421 I V A G Q S K N I G A S S S T S G A V L
1261 ATAGTTGCGGGACAGAGCAAGAACATTGGCGCAAGCTCATCAACAAGCGGTGCTGTTCTC
441 T Y K S T T P S I C T V D K N G N V T A
1321 ACATACAAAAGCACGACTCCGTCAATATGCACTGTTGACAAGAACGGAAACGTGACCGCT
461 L R T G T G Y V T I T A S A N G Y D S V
1381 CTCAGGACAGGCACAGGATATGTTACTATTACGGCTTCTGCAAACGGTTATGACAGCGTT
481 S K D V K I V V S K K S L N N G L L T L
1441 TCAAAGGACGTTAAGATAGTTGTTTCAAAGAAGAGCCTTAACAACGGACTTCTTACGCTT
501 S E T S Y V Y D G T Y K K P A A T V T F
1501 TCAGAAACAAGCTATGTTTATGACGGCACATACAAAAAGCCTGCGGCTACAGTTACGTTT
521 G G K V L Q E G K D Y T I S Y R N N L N
1561 GGCGGCAAGGTGCTTCAGGAAGGCAAGGATTACACTATCTCATACAGAAACAATCTGAAC
541 V G V T T V I A T G M G D Y T G Y T S K
1621 GTTGGCGTTACAACTGTTATCGCGACAGGCATGGGCGACTACACAGGTTACACATCAAAG
561 N F T I T K R A M A G G T V S V A S S V
1681 AACTTCACTATCACAAAGCGTGCAATGGCAGGCGGAACAGTTTCAGTGGCTTCAAGCGTT
581 S F T G S N I T P S V T V K V A G R T L
1741 TCATTCACGGGAAGCAATATTACTCCGTCAGTTACAGTAAAGGTAGCAGGCAGGACGCTT
601 T S G T D Y T V S Y S N N K N V G T S N
1801 ACAAGCGGCACTGACTACACAGTTTCATACTCAAACAACAAGAACGTTGGCACGTCAAAC
621 V Y V Y G K G N Y S G S L S A K F D I V
1861 GTTTATGTATACGGCAAGGGCAACTATTCAGGCTCTCTGAGCGCTAAGTTCGACATTGTT

```
 641 P  A  K  Q  Q  I  Q  K  L  E  T  K  Y  K  G  F  Y  I  D  W
1921 CCTGCAAAACAGCAGATACAGAAGCTTGAAACAAAATACAAGGGCTTCTATATTGATTGG
 661 A  Q  K  G  S  A  T  G  Y  D  I  E  Y  S  V  N  S  N  M  S
1981 GCGCAGAAAGGCTCTGCAACGGGATATGACATTGAGTATTCTGTAAACTCAAACATGAGC
 681 G  A  V  S  K  H  L  T  A  N  K  P  D  T  L  T  V  S  G  L
2041 GGTGCAGTGTCAAAGCATTTGACAGCAAACAAGCCTGACACGCTGACAGTAAGCGGTCTG
 701 S  G  D  K  T  Y  Y  V  R  V  R  S  Y  T  N  V  N  G  K  V
2101 TCAGGGGACAAGACATATTATGTTCGTGTTCGTTCTTACACAAACGTAAACGGCAAGGTA
 721 Y  Y  G  A  W  S  D  I  K  S  I  K  T  A  N  N  D  I  T  K
2161 TACTACGGCGCATGGTCTGATATAAAGAGCATAAAGACAGCTAACAACGACATCACAAAG
 741 A  T  V  S  G  I  S  T  K  A  F  T  G  K  A  I  T  Q  N  V
2221 GCTACTGTTTCGGGCATTTCCACAAAGGCGTTCACAGGAAAGGCTATTACACAGAACGTA
 761 T  V  K  V  G  N  T  V  L  K  N  G  T  D  Y  T  V  S  Y  S
2281 ACTGTAAAGGTGGGAAACACTGTTCTTAAAAACGGCACTGACTACACAGTTTCATACTCC
 781 N  N  K  K  V  G  K  A  T  V  K  I  T  G  K  G  K  Y  G  G
2341 AACAACAAAAAAGTTGGCAAGGCTACTGTAAAGATAACAGGCAAGGGCAAGTATGGCGGA
 801 V  I  T  K  T  F  K  I  N  P  A  K  Q  E  I  Q  K  L  T  A
2401 GTTATCACAAAGACGTTCAAGATAAATCCTGCAAAGCAGGAGATACAGAAGCTTACGGCT
 821 K  S  K  A  F  F  V  D  W  A  Q  K  G  S  A  T  G  Y  E  I
2461 AAGTCAAAGGCATTCTTTGTGGATTGGGCGCAAAAAGGCTCTGCTACGGGATATGAAATT
 841 Q  Y  A  T  N  S  K  F  T  G  A  K  K  V  A  I  T  N  N  K
2521 CAGTATGCAACCAACTCAAAGTTCACTGGTGCAAAGAAAGTGGCTATAACAAACAACAAG
 861 T  D  K  T  T  V  S  K  L  S  G  N  K  K  Y  Y  V  R  V  R
2581 ACAGACAAGACCACTGTTTCAAAGCTTTCGGGCAACAAGAAATATTACGTTCGTGTGAGA
 881 S  Y  T  T  V  G  G  T  K  Y  Y  G  S  W  S  A  T  K  T  V
2641 TCTTACACAACTGTAGGTGGCACAAAGTATTACGGTTCTTGGTCGGCAACAAAAACTGTC
 901 T  T  K  K  *
2701 ACTACAAAGAAATAA
```

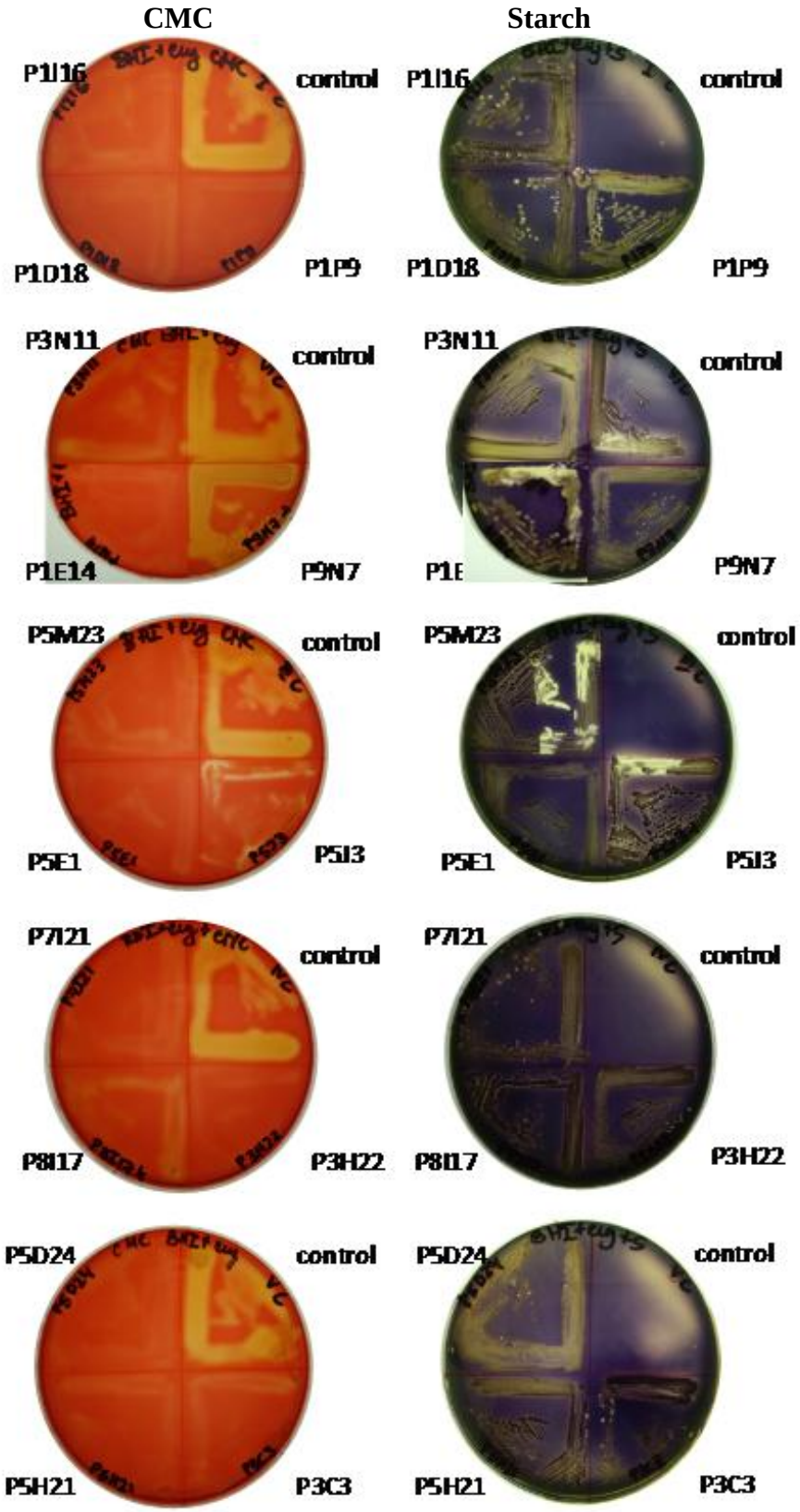
Appendix 1.5 Sequence of ORF4, cellulase (2715 bp) from clones pFI2710_1 and pFI2710_2CMC. The open reading frame and translation (904 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 2713 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequence for GH5 catalytic domain is in italics and bold. Conserved glutamate residues are in red. Internal repeats are underlined.

Domains are indicated as follows:
Signal peptide 1-35, Family 5 catalytic domain 61-372 (PF00150), bacterial Ig-like domain group 2 407-484 (PF02368), fibronectin type III domain 641-720, 810-888 (PF00041).



Appendix 1.6 Multiple alignment of the catalytic region of ORF4 (pFI2710_1CMC and pFI2710_2CMC) with closely related GH5 family enzymes. CEL-803 – ORF4, *Ruminococcus* sp. 80/3, END-EUSV – *Eubacterium siraeum* V10Sc8a (CBL34359.1), CEL-H10 – *Clostridium cellulolyticum* H10 (YP_002505438.1), CEL-2782 - *Clostridium papyrosolvens* DSM 2782 (ZP_08190896.1), END-164 - *Butyrivibrio fibrisolvens* 16/4 (CBK74879.1), CEL5C-B136 - *Butyrivibrio proteoclasticus* B316 (YP_003831029.1). Highly conserved Glu (E) residues, predicted as putative proton donor and nucleophile, are shown in red box.

Appendix 2



Appendix 2.1 Functional secondary screening of metagenomic *E. coli* clones on CMC- and starch-containing plates. Control = CMCase or amylase positive clones

Appendix 3

Sequence of genes encoding glycoside hydrolase family 9 enzymes from *Coprococcus etutactus* ART55/1 and *Coprococcus* sp. L2-50. Domain search was done using on-line database PFAM and SMART accessed on 22 August 2011(Finn *et al.,* 2010, Letunic *et al.,* 2009).

agaaacttcaaaaagaatacttgctaataaagtattggtaaaatagcacagaaaacattggtaattatgcgcataa
ataatgacggattattgggtcgcatgaagatgttttctgtgctatatttttaataatagggtatttttttaggagga
aaaaggaagtacaaggaggaaaaagtaaa

```
   1 M  R  K  K  A  A  S  R  I  L  A  Y  V  L  T  L  C  M  I  I
   1 ATGAGAAAGAAAGCAGCGAGCAGGATCTTGGCATATGTTCTCACCTTATGCATGATCATC
  21 G  S  I  T  W  P  E  I  T  A  K  A  E  S  I  T  K  D  L  K
  61 GGCAGTATAACCTGGCCGGAGATCACAGCTAAGGCGGAGAGCATCACAAAGGATCTGAAA
  41 P  D  T  G  W  K  T  V  T  A  A  T  D  E  W  S  D  Y  G  K
 121 CCGGACACCGGCTGGAAGACGGTGACTGCGGCAACAGATGAGTGGAGTGATTATGGAAAG
  61 A  E  I  R  F  S  P  S  S  D  L  A  S  M  K  A  I  A  D  A
 181 GCGGAGATCAGATTTTCGCCTTCATCAGATCTGGCATCCATGAAGGCTATTGCGGATGCG
  81 G  Y  K  T  L  K  I  T  Y  A  V  D  T  F  T  A  A  S  G  Q
 241 GGATACAAGACGCTCAAGATCACATATGCGGTTGACACATTTACCGCTGCAAGTGGACAG
 101 N  A  G  V  M  P  F  A  S  Y  G  S  S  W  S  N  N  D  K  W
 301 AATGCGGGAGTTATGCCATTTGCATCATATGGATCTTCATGGTCGAATAATGACAAGTGG
 121 I  D  L  S  K  S  G  Q  F  E  T  V  L  D  L  S  S  I  S  T
 361 ATAGATCTGTCCAAGAGTGGTCAGTTTGAGACCGTACTTGATCTGTCATCTATCAGCACA
 141 T  S  T  E  K  V  A  F  G  I  Q  V  A  N  L  Q  E  N  S  T
 421 ACCAGCACGGAGAAGGTGGCATTTGGAATACAGGTTGCCAATCTTCAGGAGAACAGCACG
 161 I  K  F  R  I  V  S  A  V  L  S  G  T  K  S  T  S  G  G  S
 481 ATCAAGTTCAGAATCGTGTCAGCGGTTCTCTCAGGAACAAAGAGCACCTCTGGTGGCTCA
 181 S  G  E  S  G  G  S  G  D  S  G  S  G  S  A  D  L  D  S  I
 541 TCAGGGGAGTCAGGCGGTTCAGGCGACAGCGGATCAGGCAGTGCGGACCTTGATTCAATT
 201 G  N  T  S  S  S  V  T  A  S  L  A  D  G  D  G  T  A  K  G
 601 GGAAACACAAGTTCAAGCGTAACCGCAAGTCTTGCTGATGGAGATGGAACAGCCAAGGGA
 221 D  G  Y  Y  E  T  E  I  T  I  N  N  K  S  N  S  Y  V  A  D
 661 GACGGCTACTACGAGACAGAGATAACGATAAATAACAAGAGCAATTCATACGTAGCAGAC
 241 W  I  A  V  A  D  V  S  G  S  V  T  A  V  K  D  Y  S  S  W
 721 TGGATAGCGGTGGCTGATGTGAGTGGCTCGGTAACTGCAGTTAAGGATTACAGCAGCTGG
 261 S  A  L  K  G  V  F  S  D  G  K  L  Y  I  Y  P  N  T  S  K
 781 TCAGCCCTCAAGGGTGTATTCAGTGATGGAAAGCTGTACATATATCCAAACACATCCAAG
 281 K  S  G  A  V  N  A  G  S  S  V  S  Y  S  K  L  G  Y  T  G
 841 AAGTCAGGTGCAGTAAATGCAGGTTCGAGCGTGAGCTATTCAAAGCTTGGCTACACAGGT
 301 T  A  N  G  V  S  I  T  G  V  K  V  Y  Y  S  S  Q  S  G  A
 901 ACGGCAAACGGTGTGTCTATAACAGGTGTCAAGGTGTACTACAGCTCACAGTCAGGTGCA
 321 F  D  S  F  I  G  S  L  S  S  S  S  G  G  A  G  D  N  T  G
 961 TTTGACTCATTTATCGGTTCACTTTCATCATCATCTGGCGGAGCCGGGGACAATACAGGT
 341 E  I  N  T  D  V  E  Y  N  Y  A  K  L  L  Q  E  S  L  Y  L
1021 GAGATCAATACGGATGTTGAGTACAACTATGCGAAGCTTCTTCAGGAGTCACTGTATCTT
 361 Y  D  A  N  M  C  G  S  D  V  S  A  K  S  E  F  S  W  R  S
1081 TATGATGCGAATATGTGCGGAAGTGATGTATCTGCAAAGAGTGAGTTTTCTTGGAGATCA
 381 N  C  H  T  E  D  A  K  T  T  Y  N  G  K  T  V  D  V  S  G
1141 AACTGCCATACTGAGGATGCAAAAACAACATATAATGGAAAGACTGTGGATGTCAGCGGT
 401 G  Y  H  D  A  G  D  H  A  K  F  G  L  P  Q  A  Y  S  A  T
1201 GGATACCATGATGCGGGAGATCATGCCAAGTTTGGTCTGCCACAGGCATATTCGGCTACA
 421 V  L  G  L  A  H  M  E  F  A  E  A  F  A  D  T  A  T  E  A
1261 GTTCTAGGACTTGCGCACATGGAGTTTGCCGAGGCATTTGCGGATACTGCTACAGAGGCA
 441 H  Y  K  R  I  M  D  R  F  V  A  Y  F  E  R  C  T  V  L  G
1321 CACTATAAGAGGATCATGGACAGATTCGTTGCTTATTTCGAGAGATGCACAGTCCTTGGA
 461 S  D  G  S  V  Q  A  F  C  Y  Q  V  G  D  G  N  V  D  H  G
1381 AGTGATGGCTCGGTTCAGGCTTTCTGTTATCAGGTCGGTGATGGAAATGTCGATCATGGA
 481 Y  W  G  A  P  E  K  Q  S  S  R  S  G  Q  A  T  F  T  S  D
1441 TACTGGGGAGCTCCGGAAAAGCAGTCATCAAGAAGTGGACAGGCAACATTCACATCTGAT
 501 S  D  T  C  T  D  I  V  S  E  T  A  A  A  L  A  A  Y  Y  I
1501 TCGGATACGTGTACAGATATCGTTTCAGAGACAGCGGCGGCACTTGCGGCATATTATATA
 521 N  Y  K  D  K  K  A  L  S  Y  A  E  K  L  F  T  Y  A  D  T
1561 AATTACAAGGATAAGAAAGCGCTTTCCTATGCGGAGAAGCTTTTCACATATGCGGATACT
 541 K  A  K  K  N  S  S  G  P  A  S  S  F  Y  S  S  D  S  W  E
1621 AAGGCTAAGAAGAACAGCTCGGGTCCGGCATCAAGTTTCTATAGCTCAGATTCATGGGAA
 561 D  D  Y  A  L  A  A  A  L  L  Y  K  A  T  G  K  S  A  Y  A
1681 GATGACTATGCCCTTGCGGCGGCTCTTCTGTACAAGGCTACAGGTAAATCGGCATATGCA
 581 T  K  Y  N  N  V  Y  G  G  R  T  N  P  N  W  A  L  C  W  N
1741 ACAAAATACAATAATGTATATGGTGGCAGAACAAATCCTAACTGGGCTCTTTGCTGGAAT
 601 N  V  A  Q  A  A  L  L  Y  S  P  N  S  S  K  K  S  V  F  V
1801 AATGTGGCTCAGGCGGCTCTGCTCTACAGCCCTAACTCATCCAAGAAGAGCGTGTTCGTG
 621 E  N  Q  S  G  L  I  A  S  K  T  Q  S  G  D  N  N  F  C  L
1861 GAGAACCAGTCAGGACTTATAGCAAGCAAGACACAGTCAGGAGATAACAACTTCTGTCTG
```

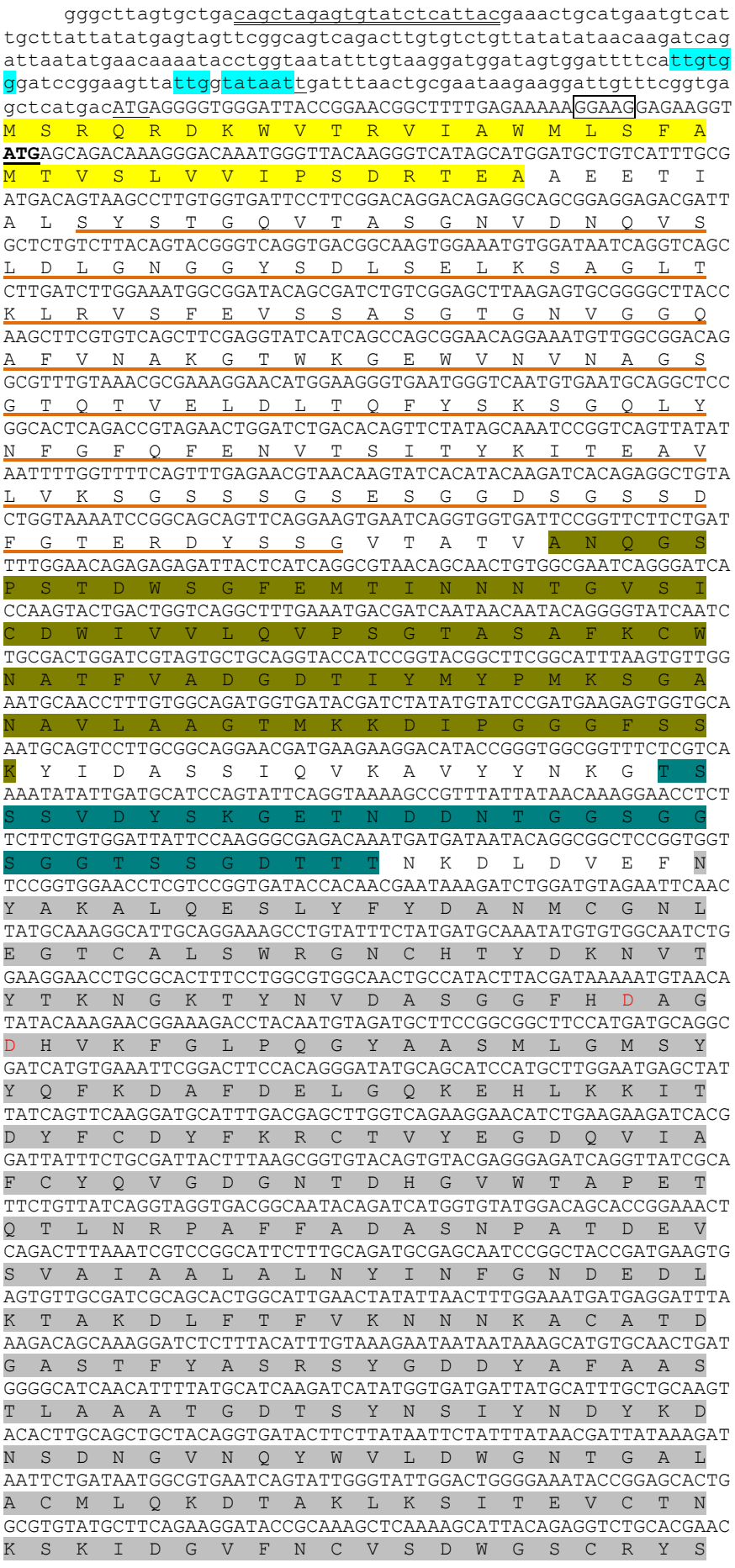
```
 641 I  D  S  W  G  S  A  R  Y  N  T  A  H  Q  M  T  G  L  M  Y
1921 ATAGACAGCTGGGGATCGGCAAGATATAACACCGCTCATCAGATGACAGGTCTCATGTAT
 661 D  T  I  Y  G  K  T  D  Y  S  S  W  A  N  G  Q  M  K  Y  I
1981 GACACGATATATGGAAAGACTGATTATTCATCATGGGCAAATGGTCAGATGAAGTACATA
 681 L  G  N  N  A  G  S  K  C  F  V  V  G  Y  N  K  Y  S  S  K
2041 CTTGGCAACAATGCTGGTTCAAAGTGTTTCGTTGTAGGTTACAACAAGTATTCATCAAAG
 701 Y  P  H  H  R  A  S  S  G  Y  Q  G  S  V  T  G  N  A  Y  T
2101 TACCCTCATCACAGAGCCTCAAGCGGATATCAGGGTTCTGTAACTGGAAATGCATATACA
 721 K  Q  A  H  V  L  V  G  A  L  V  G  G  P  A  G  S  S  T  T
2161 AAGCAGGCACATGTACTAGTCGGTGCTTTGGTTGGCGGACCGGCGGGTTCGAGTACAACC
 741 Y  V  D  S  S  E  D  Y  N  Q  N  E  V  A  L  D  Y  N  A  G
2221 TATGTAGATAGCTCAGAAGATTATAACCAGAACGAGGTTGCACTGGACTATAATGCAGGC
 761 L  V  G  A  A  A  G  L  Y  L  Y  V  K  N  S  G  T  D  E  E
2281 CTTGTGGGTGCAGCAGCAGGACTTTATCTGTATGTGAAGAACAGCGGAACAGATGAGGAG
 781 K  A  V  Q  K  V  V  P  K  S  E  V  S  S  E  L  R  T  I  S
2341 AAGGCAGTACAGAAGGTAGTTCCAAAGAGTGAAGTGTCATCAGAGCTCAGGACAATAAGC
 801 G  E  S  G  G  G  M  T  T  E  V  D  T  K  D  P  S  T  G  S
2401 GGAGAGTCAGGTGGAGGCATGACAACAGAGGTCGACACGAAGGATCCTTCAACTGGCTCA
 821 T  G  S  T  G  S  T  G  S  T  G  S  T  G  S  T  T  G  S  T
2461 ACCGGCTCAACCGGATCGACTGGCTCAACTGGCTCAACTGGATCAACAACGGGTTCAACG
 841 S  E  K  D  T  E  T  P  S  E  P  P  A  V  K  V  T  G  I  S
2521 TCTGAGAAAGACACAGAGACTCCGTCAGAGCCCCCAGCCGTAAAGGTTACAGGAATCTCA
 861 F  D  K  T  Y  I  T  L  N  V  G  D  S  D  E  I  K  A  T  I
2581 TTTGACAAGACATACATAACACTTAATGTGGGCGACAGTGATGAGATCAAGGCAACGATA
 881 T  P  A  D  A  K  D  T  S  L  V  W  S  S  S  D  K  A  K  V
2641 ACACCTGCTGATGCAAAAGATACATCCCTTGTGTGGTCAAGTTCAGACAAGGCTAAGGTA
 901 S  V  Q  N  G  K  I  T  A  L  A  A  G  T  A  T  I  T  A  T
2701 TCCGTACAGAATGGCAAGATAACAGCCCTCGCAGCCGGAACGGCGACTATCACAGCTACG
 921 A  K  D  G  S  E  I  K  S  D  C  V  V  K  V  L  T  P  G  K
2761 GCTAAGGATGGCTCGGAAATAAAATCGGATTGTGTTGTCAAGGTACTCACTCCTGGAAAG
 941 L  S  C  S  S  D  A  K  A  W  T  D  L  V  Y  G  Y  T  N  A
2821 CTGTCATGTAGCAGTGATGCCAAGGCATGGACAGACCTCGTATATGGCTATACAAATGCG
 961 N  H  A  A  I  E  L  S  N  S  G  E  T  T  L  T  D  V  K  A
2881 AATCATGCAGCTATTGAGCTGTCAAACTCCGGGGAGACAACCCTTACAGATGTAAAGGCC
 981 E  L  K  D  G  T  A  F  Q  I  V  S  S  P  A  G  Q  I  S  A
2941 GAGCTTAAGGATGGAACGGCATTCCAGATCGTTTCATCACCAGCAGGGCAGATTTCTGCA
1001 G  N  K  T  T  V  S  V  K  P  V  N  G  L  G  A  G  I  Y  S
3001 GGGAATAAAACAACAGTGTCTGTGAAACCGGTAAACGGTCTTGGCGCCGGAATCTACAGT
1021 D  T  L  V  I  S  T  A  N  G  S  A  N  V  T  L  K  A  T  V
3061 GATACACTGGTTATATCAACAGCAAATGGCTCAGCGAACGTAACACTTAAGGCGACGGTG
1041 A  K  G  E  N  T  A  V  V  T  L  T  K  T  S  V  T  S  D  S
3121 GCAAAGGGAGAAAATACCGCCGTTGTCACTCTGACAAAAACTTCAGTAACATCTGATTCG
1061 V  T  V  S  G  Q  V  S  G  A  A  G  E  I  E  Y  A  I  S  G
3181 GTGACCGTCAGCGGTCAGGTGTCAGGAGCAGCAGGCGAGATTGAGTATGCAATATCAGGC
1081 D  K  N  M  V  E  S  K  L  V  W  Q  K  D  A  Q  F  T  G  L
3241 GATAAGAATATGGTTGAATCAAAGCTTGTATGGCAGAAGGATGCTCAGTTTACCGGGCTT
1101 K  E  F  T  T  Y  Y  V  Y  S  R  V  K  E  T  A  N  V  K  A
3301 AAGGAATTTACTACATATTATGTGTATAGCCGAGTAAAAGAAACTGCAAATGTCAAGGCT
1121 G  Q  I  S  Q  A  L  G  V  T  T  L  L  S  D  P  F  V  I  D
3361 GGTCAGATCAGCCAGGCACTTGGGGTGACTACACTTTTATCAGATCCTTTCGTAATTGAT
1141 I  G  R  L  N  D  S  R  Y  V  G  A  L  V  D  S  E  G  N  P
3421 ATCGGAAGACTGAATGACAGCAGATATGTTGGAGCGCTCGTTGACAGTGAAGGAAATCCG
1161 T  V  K  V  S  Q  S  G  G  I  I  S  V  S  F  T  E  N  K  D
3481 ACTGTAAAGGTATCACAGTCTGGTGGAATCATATCAGTTTCATTTACCGAGAATAAGGAT
1181 Y  T  I  T  G  N  G  E  D  V  I  V  D  T  V  N  A  S  G  I
3541 TACACGATCACAGGCAATGGCGAGGATGTAATTGTCGATACCGTAAATGCAAGTGGAATA
1201 T  V  D  N  A  T  V  K  K  I  T  V  R  P  G  N  S  G  T  F
3601 ACCGTGGACAATGCTACTGTAAAGAAGATCACGGTTCGTCCGGGCAATTCTGGAACATTT
1221 V  I  Q  V  S  G  E  N  K  I  V  D  G  I  T  C  V  S  D  N
3661 GTAATACAGGTAAGCGGAGAAAATAAGATTGTTGATGGTATAACATGTGTAAGCGATAAC
1241 S  N  A  A  D  V  K  V  S  G  K  D  T  S  S  K  I  S  T  A
3721 AGTAACGCAGCGGATGTAAAGGTTTCTGGAAAGGATACATCTTCCAAGATATCAACGGCG
1261 T  P  G  A  A  A  I  K  A  D  G  N  I  E  I  D  N  I  Q  I
3781 ACACCGGGTGCGGCTGCGATCAAGGCAGATGGAAATATCGAGATAGACAATATACAGATA
1281 K  S  D  G  K  G  I  E  S  A  G  T  V  K  I  A  G  G  S  N
3841 AAGTCCGACGGCAAGGGAATTGAATCTGCAGGAACGGTGAAGATAGCTGGAGGAAGCAAC
1301 S  I  D  S  V  S  S  A  I  S  A  S  D  V  E  I  T  G  G  S
3901 AGCATAGATTCCGTATCTTCTGCAATATCGGCATCGGATGTGGAGATCACCGGAGGCTCT
```

```
1321 V  D  A  K  S  S  G  L  G  D  G  E  S  V  I  S  A  D  N  S
3961 GTTGATGCAAAGTCATCAGGACTTGGCGATGGGGAGTCGGTAATCTCAGCAGACAATTCT
1341 I  K  L  V  G  G  S  V  T  A  D  A  S  G  S  T  G  G  N  S
4021 ATCAAACTGGTTGGTGGAAGTGTTACGGCGGATGCATCCGGATCAACCGGTGGAAATTCA
1361 F  G  V  K  S  D  D  G  T  I  I  V  D  G  D  V  S  I  G  G
4081 TTTGGCGTAAAGTCAGATGATGGAACGATCATTGTTGATGGCGATGTGTCAATAGGTGGA
1381 D  P  I  Y  S  K  D  P  V  D  S  K  G  D  S  V  T  M  V  K
4141 GATCCAATTTATTCCAAGGATCCGGTAGACAGCAAGGGCGACAGTGTAACCATGGTAAAG
1401 V  V  F  T  D  E  N  G  N  Q  L  Y  A  S  S  V  N  K  G  S
4201 GTTGTGTTCACAGATGAGAATGGCAACCAGTTGTATGCGTCATCTGTAAACAAGGGATCA
1421 T  L  D  L  S  K  I  N  I  T  M  S  D  G  T  D  Y  K  T  S
4261 ACTCTTGATCTGTCGAAAATAAATATCACCATGTCTGATGGTACTGATTACAAGACATCG
1441 R  A  G  Y  A  L  A  W  Q  D  E  A  G  T  S  Y  N  A  D  S
4321 AGGGCTGGATATGCGCTGGCATGGCAGGATGAAGCCGGAACAAGCTATAATGCTGATTCT
1461 A  Y  G  A  V  S  E  D  I  T  F  K  A  V  W  T  W  I  V  V
4381 GCATACGGCGCTGTATCAGAGGATATAACATTTAAGGCTGTATGGACATGGATAGTTGTG
1481 D  I  S  V  D  A  N  I  T  F  A  A  T  G  N  A  S  Y  K  T
4441 GATATAAGCGTTGATGCAAATATTACATTTGCGGCAACTGGAAATGCTTCGTACAAGACT
1501 T  Y  T  G  A  A  I  R  P  A  V  K  V  V  S  R  G  R  T  L
4501 ACATATACAGGCGCAGCGATAAGACCGGCCGTAAAGGTTGTATCACGTGGAAGAACACTT
1521 D  A  G  T  D  Y  I  V  S  Y  S  S  N  T  N  A  G  T  A  K
4561 GATGCGGGAACTGACTATATAGTTTCATATAGCAGCAATACCAATGCCGGTACAGCAAAG
1541 V  V  V  K  G  Q  G  R  Y  K  G  S  K  T  F  T  F  A  I  N
4621 GTAGTTGTCAAAGGTCAGGGCAGATATAAGGGCTCAAAGACTTTTACATTTGCAATAAAC
1561 K  Q  A  I  S  K  A  S  V  T  I  T  K  S  A  V  Y  T  G  K
4681 AAGCAGGCGATAAGTAAAGCCTCTGTGACAATCACAAAGTCGGCTGTATATACAGGCAAG
1581 A  I  T  P  A  V  K  V  T  S  G  K  R  T  L  T  A  G  K  D
4741 GCTATAACACCAGCGGTGAAGGTGACTAGCGGTAAGCGGACACTCACAGCGGGCAAGGAT
1601 Y  R  V  A  Y  S  S  N  K  D  F  G  K  A  K  V  V  I  M  G
4801 TACAGAGTTGCATATTCGTCAAATAAAGACTTTGGCAAGGCAAAAGTTGTGATAATGGGA
1621 I  G  N  Y  S  G  T  Q  T  K  Y  F  D  I  T  A  A  V  G  K
4861 ATCGGTAATTACAGTGGAACACAGACAAAGTATTTTGACATAACTGCGGCTGTGGGTAAG
1641 I  Y  A  N  G  N  Y  K  Y  K  I  T  N  A  S  L  N  G  K  G
4921 ATCTATGCAAATGGAAATTACAAGTACAAGATCACAAATGCATCCCTGAATGGCAAGGGA
1661 T  V  T  L  V  S  V  V  K  K  T  K  T  V  V  V  P  D  T  I
4981 ACGGTAACTCTCGTATCTGTTGTTAAGAAGACCAAGACAGTTGTGGTTCCTGACACTATA
1681 K  L  G  G  K  T  F  K  V  T  A  I  G  K  A  A  F  K  K  N
5041 AAGCTTGGTGGCAAGACATTTAAGGTCACAGCGATAGGAAAAGCGGCGTTCAAGAAGAAT
1701 V  K  V  T  K  V  T  L  G  K  N  V  K  T  I  G  A  K  A  F
5101 GTAAAGGTTACAAAGGTGACCTTGGGCAAAAACGTCAAGACTATAGGTGCCAAGGCGTTC
1721 Y  G  C  K  K  L  R  T  V  V  I  K  N  T  Q  M  T  G  K  T
5161 TATGGATGTAAGAAGCTCAGGACAGTAGTGATAAAGAATACACAGATGACAGGAAAAACA
1741 V  G  S  G  A  F  T  G  T  Y  A  K  M  T  V  K  V  P  S  K
5221 GTTGGATCTGGTGCATTTACAGGGACATACGCCAAGATGACCGTCAAGGTTCCGTCGAAG
1761 K  L  K  S  Y  K  T  I  L  L  K  R  G  V  S  K  K  A  V  I
5281 AAGCTTAAGAGTTATAAGACAATCCTTCTGAAAAGAGGAGTGTCAAAGAAAGCAGTCATA
1781 K  K  *
5341 AAGAAATAA
```

```
aggctcaataaaataaagaataatttgagatccggggtggttcatatgatcccggatctttttgt
gcgtatggaggtgaccgaagttcgatttgacatgggtaaaaatatgtgtataatatcgttagcatt
ctaata
```

Appendix 3.1 Sequence of ART_GH9/L (putative cellulase, 5349 bp). The open reading frame and translation (1782 amino acids) is shown in one letter code. The start (at position 1) and stop codons (at position 5347 bp) are underlined and bold. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH9 catalytic domain are in italics and bold (Pereira *et al.*, 2009). The conserved aspartate and glutamate residues are in red. Internal repeats are underlined in pink. The GS-rich and K-rich regions are underlined in orange. Primers used in this study are double underlined. Domains are indicated as follows:

Signal peptide : 1- 32, Family 9 catalytic domain: 349-760 (PF00759),TGS linker: 808-840, bacterial Ig-like domain: 856-932 (PF02368).

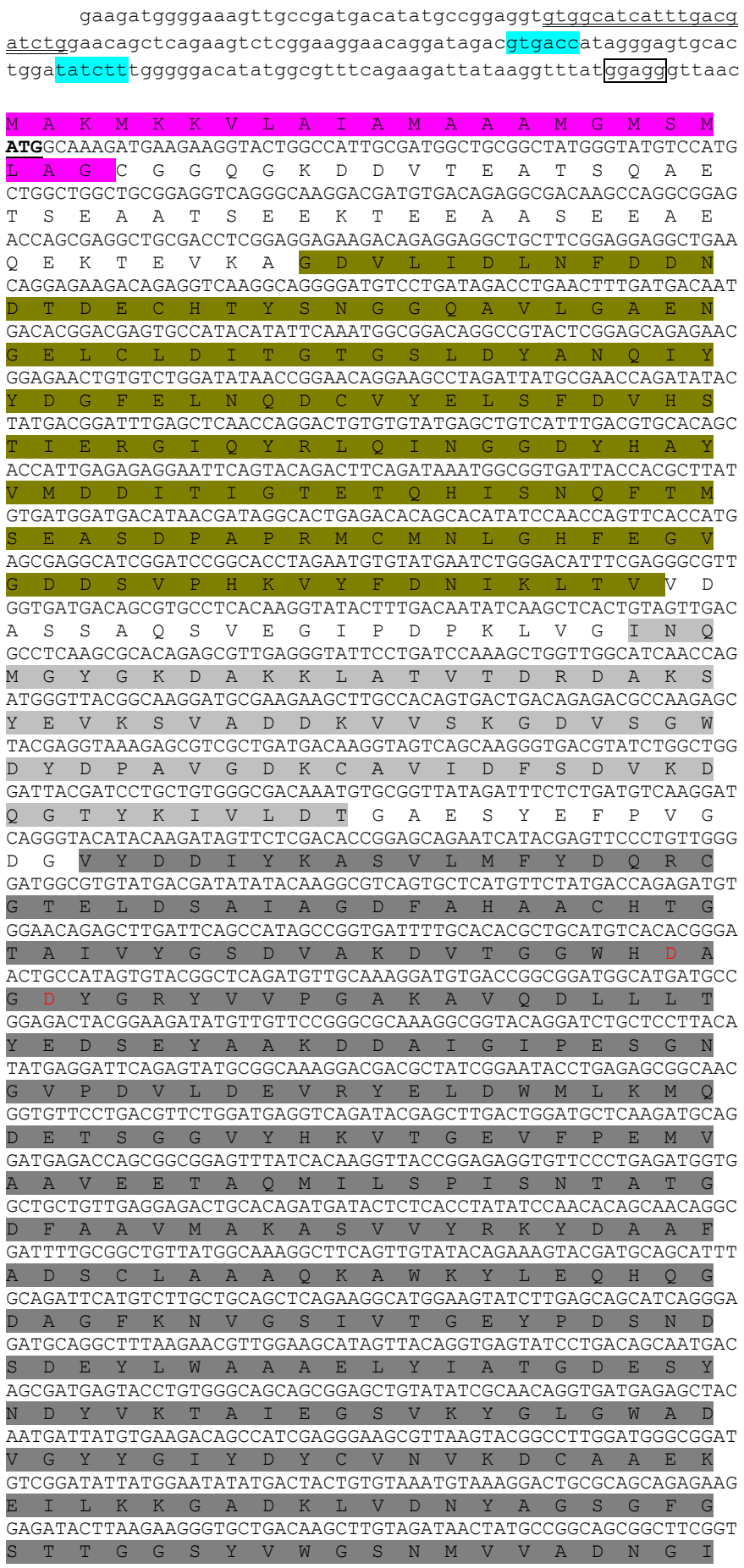

gggcttagtgctgacagctagagtgtatctcattacgaaactgcatgaatgtcat
tgcttattatatgagtagttcggcagtcagacttgtgtctgttatatataacaagatcag
attaatatgaacaaaatacctggtaatatttgtaaggatggatagtggattttcattgtg
ggatccggaagttattggtataattgatttaactgcgaataagaaggattgtttcggtga
gctcatgacATGAGGGGTGGGATTACCGGAACGGCTTTTGAGAAAAAGGAAGGAGAAGGT
1 M S R Q R D K W V T R V I A W M L S F A
1 ATGAGCAGACAAAGGGACAAATGGGTTACAAGGGTCATAGCATGGATGCTGTCATTTGCG
21 M T V S L V V I P S D R T E A A E E T I
61 ATGACAGTAAGCCTTGTGGTGATTCCTTCGGACAGGACAGAGGCAGCGGAGGAGACGATT
41 A L S Y S T G Q V T A S G N V D N Q V S
121 GCTCTGTCTTACAGTACGGGTCAGGTGACGGCAAGTGGAAATGTGGATAATCAGGTCAGC
61 L D L G N G G Y S D L S E L K S A G L T
181 CTTGATCTTGGAAATGGCGGATACAGCGATCTGTCGGAGCTTAAGAGTGCGGGGCTTACC
81 K L R V S F E V S S A S G T G N V G G Q
241 AAGCTTCGTGTCAGCTTCGAGGTATCATCAGCCAGCGGAACAGGAAATGTTGGCGGACAG
101 A F V N A K G T W K G E W V N V N A G S
301 GCGTTTGTAAACGCGAAAGGAACATGGAAGGGTGAATGGGTCAATGTGAATGCAGGCTCC
121 G T Q T V E L D L T Q F Y S K S G Q L Y
361 GGCACTCAGACCGTAGAACTGGATCTGACACAGTTCTATAGCAAATCCGGTCAGTTATAT
141 N F G F Q F E N V T S I T Y K I T E A V
421 AATTTTGGTTTTCAGTTTGAGAACGTAACAAGTATCACATACAAGATCACAGAGGCTGTA
161 L V K S G S S S G S E S G G D S G S S D
481 CTGGTAAAATCCGGCAGCAGTTCAGGAAGTGAATCAGGTGGTGATTCCGGTTCTTCTGAT
181 F G T E R D Y S S G V T A T V A N Q G S
541 TTTGGAACAGAGAGAGATTACTCATCAGGCGTAACAGCAACTGTGGCGAATCAGGGATCA
201 P S T D W S G F E M T I N N N T G V S I
601 CCAAGTACTGACTGGTCAGGCTTTGAAATGACGATCAATAACAATACAGGGGTATCAATC
221 C D W I V V L Q V P S G T A S A F K C W
661 TGCGACTGGATCGTAGTGCTGCAGGTACCATCCGGTACGGCTTCGGCATTTAAGTGTTGG
241 N A T F V A D G D T I Y M Y P M K S G A
721 AATGCAACCTTTGTGGCAGATGGTGATACGATCTATATGTATCCGATGAAGAGTGGTGCA
261 N A V L A A G T M K K D I P G G G F S S
781 AATGCAGTCCTTGCGGCAGGAACGATGAAGAAGGACATACCGGGTGGCGGTTTCTCGTCA
281 K Y I D A S S I Q V K A V Y Y N K G T S
841 AAATATATTGATGCATCCAGTATTCAGGTAAAAGCCGTTTATTATAACAAAGGAACCTCT
301 S S V D Y S K G E T N D D N T G G S G G
901 TCTTCTGTGGATTATTCCAAGGGCGAGACAAATGATGATAATACAGGCGGCTCCGGTGGT
321 S G G T S S G D T T T N K D L D V E F N
961 TCCGGTGGAACCTCGTCCGGTGATACCACAACGAATAAAGATCTGGATGTAGAATTCAAC
341 Y A K A L Q E S L Y F Y D A N M C G N L
1021 TATGCAAAGGCATTGCAGGAAAGCCTGTATTTCTATGATGCAAATATGTGTGGCAATCTG
361 E G T C A L S W R G N C H T Y D K N V T
1081 GAAGGAACCTGCGCACTTTCCTGGCGTGGCAACTGCCATACTTACGATAAAAATGTAACA
381 Y T K N G K T Y N V D A S G G F H D A G
1141 TATACAAAGAACGGAAAGACCTACAATGTAGATGCTTCCGGCGGCTTCCATGATGCAGGC
401 D H V K F G L P Q G Y A A S M L G M S Y
1201 GATCATGTGAAATTCGGACTTCCACAGGGATATGCAGCATCCATGCTTGGAATGAGCTAT
421 Y Q F K D A F D E L G Q K E H L K K I T
1261 TATCAGTTCAAGGATGCATTTGACGAGCTTGGTCAGAAGGAACATCTGAAGAAGATCACG
441 D Y F C D Y F K R C T V Y E G D Q V I A
1321 GATTATTTCTGCGATTACTTTAAGCGGTGTACAGTGTACGAGGGAGATCAGGTTATCGCA
461 F C Y Q V G D G N T D H G V W T A P E T
1381 TTCTGTTATCAGGTAGGTGACGGCAATACAGATCATGGTGTATGGACAGCACCGGAAACT
481 Q T L N R P A F F A D A S N P A T D E V
1441 CAGACTTTAAATCGTCCGGCATTCTTTGCAGATGCGAGCAATCCGGCTACCGATGAAGTG
501 S V A I A A L A L N Y I N F G N D E D L
1501 AGTGTTGCGATCGCAGCACTGGCATTGAACTATATTAACTTTGGAAATGATGAGGATTTA
521 K T A K D L F T F V K N N N K A C A T D
1561 AAGACAGCAAAGGATCTCTTTACATTTGTAAAGAATAATAATAAAGCATGTGCAACTGAT
541 G A S T F Y A S R S Y G D D Y A F A A S
1621 GGGGCATCAACATTTTATGCATCAAGATCATATGGTGATGATTATGCATTTGCTGCAAGT
561 T L A A A T G D T S Y N S I Y N D Y K D
1681 ACACTTGCAGCTGCTACAGGTGATACTTCTTATAATTCTATTTATAACGATTATAAAGAT
581 N S D N G V N Q Y W V L D W G N T G A L
1741 AATTCTGATAATGGCGTGAATCAGTATTGGGTATTGGACTGGGGAAATACCGGAGCACTG
601 A C M L Q K D T A K L K S I T E V C T N
1801 GCGTGTATGCTTCAGAAGGATACCGCAAAGCTCAAAAGCATTACAGAGGTCTGCACGAAC
621 K S K I D G V F N C V S D W G S C R Y S

```
1861 AAGAGTAAGATCGATGGAGTATTTAACTGTGTCAGTGACTGGGGCTCCTGCAGATACAGT
 641 A  A  E  Q  F  T  G  L  V  Y  D  K  L  T  G  T  T  T  Y  A
1921 GCAGCAGAACAGTTTACCGGTCTTGTGTATGATAAGCTGACAGGAACAACGACTTACGCC
 661 K  W  A  T  S  Q  M  N  Y  M  L  G  D  N  P  N  K  R  C  Y
1981 AAATGGGCAACCAGTCAGATGAACTATATGTTAGGCGATAACCCGAATAAACGATGCTAT
 681 I  V  G  Y  N  E  N  S  S  K  Y  P  H  H  R  A  A  S  R  S
2041 ATCGTAGGCTATAATGAGAACTCCAGTAAATATCCGCATCACAGAGCGGCATCCAGATCG
 701 T  D  S  K  I  I  N  S  D  H  Y  T  L  L  G  A  L  V  G  G
2101 ACAGATTCAAAGATCATAAACAGTGATCATTATACGTTGCTTGGAGCATTGGTAGGCGGA
 721 P  G  A  D  G  T  Y  K  D  D  Q  G  D  Y  F  C  N  E  V  A
2161 CCCGGGGCGGATGGAACTTATAAGGATGATCAGGGAGATTATTTCTGTAATGAAGTAGCA
 741 L  D  Y  N  A  G  L  V  G  A  A  A  G  L  Y  L  V  H  K  N
2221 CTTGATTATAATGCCGGACTGGTTGGTGCGGCAGCCGGACTTTATCTGGTACATAAGAAT
 761 D  E  T  V  Y  L  S  Y  A  K  K  N  T  T  N  Y  S  T  K  T
2281 GATGAGACTGTATATTTATCCTATGCTAAGAAGAATACAACGAATTACAGTACAAAGACA
 781 A  S  A  T  E  L  G  A  V  G  V  K  T  Y  Y  G  T  N  S  G
2341 GCATCTGCTACAGAGCTTGGTGCTGTAGGCGTGAAGACTTATTACGGAACCAATTCCGGT
 801 G  D  T  E  E  E  V  K  A  T  A  I  K  L  D  S  L  I  V  E
2401 GGTGATACAGAGGAAGAAGTAAAGGCAACAGCCATCAAACTGGATAGCCTGATCGTTGAA
 821 L  T  E  G  L  D  I  T  I  R  A  T  V  T  P  K  T  A  E  Q
2461 TTAACAGAAGGACTCGACATTACGATCCGCGCAACGGTTACTCCGAAGACTGCGGAACAG
 841 K  V  I  W  K  S  E  N  E  K  V  A  T  V  S  E  T  G  K  V
2521 AAGGTGATATGGAAATCTGAAAACGAGAAGGTAGCGACCGTTTCGGAAACAGGAAAAGTA
 861 T  A  V  G  A  G  E  T  T  V  T  A  T  T  T  D  G  T  N  L
2581 ACTGCGGTCGGAGCAGGTGAGACAACTGTGACAGCTACAACGACAGATGGAACAAATCTC
 881 T  A  S  C  R  V  T  V  K  A  A  P  K  A  E  F  V  T  D  Q
2641 ACAGCAAGCTGCAGGGTGACAGTAAAGGCAGCACCAAAGGCAGAATTTGTTACAGATCAG
 901 T  A  L  T  C  D  P  V  I  Y  G  Y  E  A  G  S  T  A  S  F
2701 ACAGCTTTAACCTGTGATCCTGTTATTTATGGATATGAAGCAGGATCGACAGCATCATTT
 921 V  L  E  N  E  G  N  D  T  G  F  V  T  I  A  L  E  K  G  T
2761 GTATTAGAGAATGAGGGAAATGATACAGGATTTGTAACGATAGCGTTAGAAAAAGGCACA
 941 D  S  P  F  E  I  S  G  D  T  S  I  K  V  A  A  S  D  K  T
2821 GACAGCCCATTTGAAATATCCGGCGATACATCGATAAAGGTTGCAGCGTCAGATAAGACA
 961 T  V  G  I  A  L  K  T  G  K  D  A  G  S  Y  S  D  N  L  L
2881 ACAGTGGGTATCGCATTAAAGACAGGTAAGGATGCAGGTTCTTACAGTGACAATCTGTTG
 981 I  E  A  D  N  G  Q  S  F  Q  I  P  V  S  A  T  V  E  K  C
2941 ATAGAGGCAGATAATGGACAGTCCTTTCAGATCCCGGTATCTGCAACGGTTGAGAAATGT
1001 P  V  T  G  I  T  F  P  T  A  G  M  I  V  T  G  Q  T  L  S
3001 CCGGTAACGGGTATTACATTCCCGACAGCAGGAATGATCGTAACCGGTCAGACACTTTCT
1021 E  S  K  L  S  G  G  D  E  T  Y  G  T  F  A  W  A  D  A  A
3061 GAGTCGAAGTTATCTGGTGGAGATGAGACCTATGGAACATTTGCATGGGCAGATGCAGCA
1041 T  K  P  E  R  G  T  Y  Q  G  Q  V  V  L  T  L  R  S  D  V
3121 ACGAAACCGGAACGAGGAACGTATCAGGGACAGGTGGTGTTGACCCTTCGTTCCGATGTG
1061 K  K  N  Y  A  F  D  R  I  T  G  Y  D  A  A  A  G  T  I  T
3181 AAGAAGAATTATGCATTTGACCGGATTACAGGGTATGATGCGGCAGCAGGCACGATCACA
1081 Q  N  V  T  V  I  V  S  R  A  G  L  P  A  I  T  F  P  E  A
3241 CAGAATGTGACCGTAATAGTCAGCAGAGCAGGACTTCCGGCTATCACGTTCCCGGAGGCT
1101 S  D  L  V  Y  G  Q  T  L  S  D  S  V  L  T  G  G  S  E  E
3301 TCTGATCTGGTATATGGACAGACACTTTCAGATTCTGTGCTGACAGGTGGTTCTGAAGAG
1121 Y  G  T  F  A  W  S  T  P  D  V  K  M  G  E  Q  E  V  A  N
3361 TACGGAACATTTGCGTGGAGCACCCCGGATGTGAAGATGGGTGAACAGGAAGTAGCAAAT
1141 G  T  N  Q  Y  E  V  V  F  T  W  S  D  A  S  K  K  Q  Y  Q
3421 GGAACCAATCAGTATGAAGTTGTATTTACATGGAGTGATGCAAGCAAGAAGCAGTATCAG
1161 I  E  D  T  D  E  D  A  V  Y  K  K  M  V  S  V  K  V  Q  K
3481 ATCGAGGATACAGATGAGGATGCTGTATATAAGAAAATGGTTTCTGTAAAGGTACAGAAG
1181 A  E  Q  T  A  V  P  D  S  V  A  L  L  A  R  T  K  D  T  I
3541 GCAGAACAGACAGCGGTTCCGGATAGTGTGGCACTTCTTGCAAGAACAAAGGATACGATC
1201 R  I  S  G  Q  A  G  V  R  Y  S  V  D  G  T  T  W  K  Q  A
3601 CGGATCTCCGGTCAGGCTGGCGTCCGGTATTCAGTGGATGGAACAACGTGGAAGCAGGCA
1221 A  S  N  G  E  T  I  E  F  T  G  L  R  S  F  T  K  Y  T  V
3661 GCTTCAAATGGCGAAACAATTGAATTTACAGGTTTGCGAAGCTTTACAAAGTATACGGTT
1241 S  G  R  Y  A  E  T  A  T  A  Y  A  G  K  A  V  E  L  L  T
3721 TCAGGCAGATATGCAGAAACTGCAACTGCATATGCAGGAAAAGCAGTGGAATTGCTTACG
1261 V  Y  T  L  V  Q  D  P  Y  T  I  D  I  A  K  I  A  D  K  E
3781 GTATATACACTTGTTCAGGATCCGTATACGATCGATATAGCAAAGATTGCGGACAAAGAA
1281 Y  Q  D  A  L  R  T  E  D  G  R  T  T  V  D  Y  T  E  P  V
3841 TATCAGGATGCATTACGGACAGAGGATGGAAGAACAACTGTCGATTATACAGAACCGGTT
1301 L  R  L  T  E  D  G  K  D  Y  V  I  T  G  K  N  R  E  L  V
```

```
3901 CTTAGACTTACAGAGGATGGCAAAGACTATGTGATCACAGGAAAGAATAGGGAGCTTGTT
1321 I  K  A  G  G  A  T  K  I  T  L  D  Q  A  E  V  G  A  I  E
3961 ATTAAGGCAGGTGGAGCAACAAAGATCACATTGGATCAGGCGGAGGTCGGAGCTATTGAG
1341 I  T  E  N  A  G  G  K  T  E  I  V  R  K  G  T  I  T  I  A
4021 ATCACCGAAAATGCTGGTGGTAAGACCGAGATCGTGAGAAAAGGAACGATAACGATCGCA
1361 G  N  V  E  T  D  G  G  L  V  I  N  G  D  G  T  L  M  V  S
4081 GGTAATGTCGAGACAGACGGCGGACTTGTGATCAATGGTGACGGAACGCTTATGGTTTCC
1381 G  K  I  S  A  T  G  D  I  T  I  K  G  G  K  V  S  V  D  G
4141 GGTAAGATATCTGCGACCGGTGACATAACGATCAAAGGCGGAAAGGTATCGGTTGATGGA
1401 G  V  L  A  G  G  T  V  L  I  R  D  T  E  L  V  A  G  T  D
4201 GGAGTCTTAGCAGGCGGCACAGTTCTGATCCGGGATACAGAGCTTGTCGCAGGTACAGAT
1421 E  T  G  T  A  I  K  A  D  T  V  R  I  E  D  S  K  V  T  V
4261 GAGACAGGAACGGCGATCAAAGCAGATACTGTACGGATCGAGGATTCCAAGGTAACGGTC
1441 G  A  N  Q  D  T  K  N  P  P  I  K  S  D  N  I  I  L  V  G
4321 GGAGCGAATCAGGATACGAAGAATCCACCGATCAAATCAGACAATATTATTCTGGTTGGA
1461 D  N  T  V  A  S  S  S  G  S  K  D  I  F  S  S  K  P  K  D
4381 GACAATACAGTAGCATCCTCTTCAGGTTCAAAGGATATCTTTTCGTCGAAACCAAAGGAT
1481 E  N  G  D  E  I  P  D  T  I  L  I  E  K  I  T  L  N  K  T
4441 GAAAATGGAGATGAGATCCCGGATACGATCCTGATCGAGAAAATCACGTTAAACAAGACC
1501 S  V  V  L  N  I  G  D  T  E  R  I  A  V  A  V  V  I  P  A
4501 AGCGTGGTATTAAATATTGGTGATACAGAACGGATCGCAGTAGCGGTCGTTATACCGGCA
1521 N  A  T  L  K  T  F  D  W  K  S  S  N  E  K  V  A  S  V  S
4561 AATGCAACACTGAAGACATTTGACTGGAAGAGCAGCAACGAGAAAGTGGCAAGTGTATCA
1541 Q  T  G  E  V  K  G  V  S  A  G  T  A  V  I  M  V  T  A  K
4621 CAGACCGGTGAAGTGAAGGGAGTAAGTGCCGGAACTGCAGTTATTATGGTAACAGCGAAG
1561 D  G  S  G  V  T  G  S  C  T  V  T  V  K  K  K  E  T  V  T
4681 GACGGAAGTGGTGTTACAGGCTCCTGCACCGTTACGGTAAAGAAGAAAGAAACTGTAACA
1581 P  T  P  T  P  N  P  E  K  P  T  T  E  K  P  T  I  E  K  P
4741 CCGACACCAACACCGAATCCGGAAAAGCCGACGACCGAGAAACCTACGATCGAAAAGCCG
1601 T  T  E  K  P  T  T  E  E  P  K  K  T  V  K  P  A  K  V  K
4801 ACGACCGAGAAACCAACAACCGAAGAACCAAAGAAGACGGTCAAACCGGCAAAGGTCAAG
1621 I  Q  A  A  T  K  K  I  A  P  G  K  K  L  T  L  K  A  T  V
4861 ATCCAGGCAGCAACCAAAAAGATCGCTCCGGGCAAGAAACTGACCTTAAAGGCAACGGTT
1641 T  P  K  K  A  T  N  K  K  V  K  W  T  I  R  K  S  D  K  K
4921 ACACCTAAGAAAGCAACAAATAAGAAAGTAAAATGGACGATCCGTAAATCGGATAAGAAG
1661 Y  A  S  I  N  S  K  G  V  L  T  V  K  K  A  A  K  G  K  T
4981 TATGCTTCTATCAACAGCAAGGGTGTTCTGACAGTTAAGAAAGCTGCCAAGGGTAAGACG
1681 I  K  V  T  A  T  A  V  G  N  S  K  V  K  A  T  Y  T  V  K
5041 ATCAAGGTAACCGCAACAGCAGTCGGTAACAGTAAGGTGAAAGCAACTTATACTGTTAAA
1701 V  M  K  K  A  V  K  K  V  T  I  T  S  K  K  K  T  V  K  K
5101 GTCATGAAAAAGGCAGTTAAGAAGGTTACGATCACATCAAAGAAAAAGACTGTGAAGAAA
1721 V  T  I  K  K  K  Q  S  V  T  L  K  A  N  L  T  P  T  K  G
5161 GTAACGATCAAGAAGAAACAGTCCGTAACCTTAAAAGCAAATCTGACTCCGACAAAGGGA
1741 I  S  K  E  V  T  W  T  S  G  N  P  K  I  A  K  V  T  A  K
5221 ATCAGTAAAGAGGTGACCTGGACCTCCGGTAATCCGAAGATTGCAAAGGTAACAGCCAAG
1761 G  K  V  T  G  K  K  K  G  T  V  K  I  T  A  K  A  K  D  G
5281 GGTAAGGTAACCGGAAAGAAGAAGGGAACGGTTAAGATCACGGCAAAAGCAAAGGACGGA
1781 S  G  K  K  A  T  I  T  L  K  V  K  *
5341 AGCGGTAAGAAAGCAACCATTACCCTTAAAGTAAAATAA
```

tcttttccatttgtaacaggtggaaacagacaatacgatggcggaagaatgatcagatcctgtaaa
atataaaaaacagatttgatagatgcttccgccgtcttgtgttttgaaaacgggtatgctataatg
tggacgcttgcag

Appendix 3.2 Sequence of L250_GH9/L (putative cellulase, 5379 bp). The open reading frame and translation (1792 amino acids) is shown in one letter code. The start (at position 1) and stop codon (at position 5377 bp) are underlined and bold. An alternative start codon ATG predicted by GeneBank is underline. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH9 catalytic domain are in italics and bold. The conserved aspartate and glutamate residues are in red. Primers used in this study are double underlined. The GS-rich and P-rich regions are underlined in orange. Domains are indicated as follows: Signal peptide: 1-35 CBM_2: 195-281, Linker: 299-331, 763-803, Family 9 catalytic domain: 340-752 (PF00759), bacterial Ig-like domain:808-886, 1494-1573, 1617-1698, 1705-1788 (PF02368), DUF583 domain: 1368-1452 (PF04519).

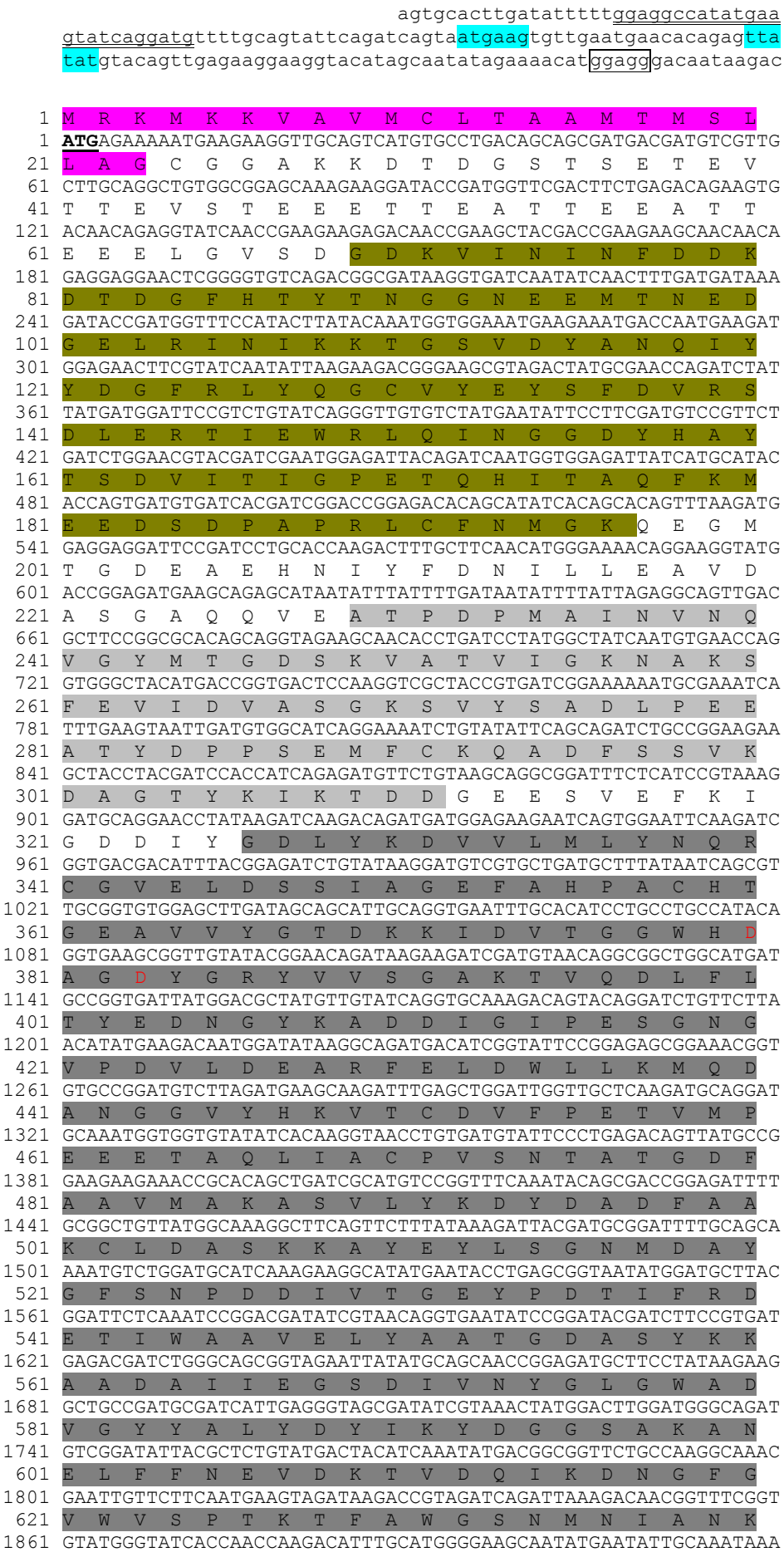

gaagatggggaaagttgccgatgacatatgccggaggtgtggcatcatttgacg
atctggaacagctcagaagtctcggaaggaacaggatagacgtgaccatagggagtgcac
tggatatcttgggggacatatggcgtttcagaagattataaggtttatggaggttaac
1 M A K M K K V L A I A M A A A M G M S M
1 ATGGCAAAGATGAAGAAGGTACTGGCCATTGCGATGGCTGCGGCTATGGGTATGTCCATG
21 L A G C G G Q G K D D V T E A T S Q A E
61 CTGGCTGGCTGCGGAGGTCAGGGCAAGGACGATGTGACAGAGGCGACAAGCCAGGCGGAG
41 T S E A A T S E E K T E E A A S E E A E
121 ACCAGCGAGGCTGCGACCTCGGAGGAGAAGACAGAGGAGGCTGCTTCGGAGGAGGCTGAA
61 Q E K T E V K A G D V L I D L N F D D N
181 CAGGAGAAGACAGAGGTCAAGGCAGGGGATGTCCTGATAGACCTGAACTTTGATGACAAT
81 D T D E C H T Y S N G G Q A V L G A E N
241 GACACGGACGAGTGCCATACATATTCAAATGGCGGACAGGCCGTACTCGGAGCAGAGAAC
101 G E L C L D I T G T G S L D Y A N Q I Y
301 GGAGAACTGTGTCTGGATATAACCGGAACAGGAAGCCTAGATTATGCGAACCAGATATAC
121 Y D G F E L N Q D C V Y E L S F D V H S
361 TATGACGGATTTGAGCTCAACCAGGACTGTGTGTATGAGCTGTCATTTGACGTGCACAGC
141 T I E R G I Q Y R L Q I N G G D Y H A Y
421 ACCATTGAGAGAGGAATTCAGTACAGACTTCAGATAAATGGCGGTGATTACCACGCTTAT
161 V M D D I T I G T E T Q H I S N Q F T M
481 GTGATGGATGACATAACGATAGGCACTGAGACACAGCACATATCCAACCAGTTCACCATG
181 S E A S D P A P R M C M N L G H F E G V
541 AGCGAGGCATCGGATCCGGCACCTAGAATGTGTATGAATCTGGGACATTTCGAGGGCGTT
201 G D D S V P H K V Y F D N I K L T V V D
601 GGTGATGACAGCGTGCCTCACAAGGTATACTTTGACAATATCAAGCTCACTGTAGTTGAC
221 A S S A Q S V E G I P D P K L V G I N Q
661 GCCTCAAGCGCACAGAGCGTTGAGGGTATTCCTGATCCAAAGCTGGTTGGCATCAACCAG
241 M G Y G K D A K K L A T V T D R D A K S
721 ATGGGTTACGGCAAGGATGCGAAGAAGCTTGCCACAGTGACTGACAGAGACGCCAAGAGC
261 Y E V K S V A D D K V V S K G D V S G W
781 TACGAGGTAAAGAGCGTCGCTGATGACAAGGTAGTCAGCAAGGGTGACGTATCTGGCTGG
281 D Y D P A V G D K C A V I D F S D V K D
841 GATTACGATCCTGCTGTGGGCGACAAATGTGCGGTTATAGATTTCTCTGATGTCAAGGAT
301 Q G T Y K I V L D T G A E S Y E F P V G
901 CAGGGTACATACAAGATAGTTCTCGACACCGGAGCAGAATCATACGAGTTCCCTGTTGGG
321 D G V Y D D I Y K A S V L M F Y D Q R C
961 GATGGCGTGTATGACGATATATACAAGGCGTCAGTGCTCATGTTCTATGACCAGAGATGT
341 G T E L D S A I A G D F A H A A C H T G
1021 GGAACAGAGCTTGATTCAGCCATAGCCGGTGATTTTGCACACGCTGCATGTCACACGGGA
361 T A I V Y G S D V A K D V T G G W H D A
1081 ACTGCCATAGTGTACGGCTCAGATGTTGCAAAGGATGTGACCGGCGGATGGCATGATGCC
381 G D Y G R Y V V P G A K A V Q D L L L T
1141 GGAGACTACGGAAGATATGTTGTTCCGGGCGCAAAGGCGGTACAGGATCTGCTCCTTACA
401 Y E D S E Y A A K D D A I G I P E S G N
1201 TATGAGGATTCAGAGTATGCGGCAAAGGACGACGCTATCGGAATACCTGAGAGCGGCAAC
421 G V P D V L D E V R Y E L D W M L K M Q
1261 GGTGTTCCTGACGTTCTGGATGAGGTCAGATACGAGCTTGACTGGATGCTCAAGATGCAG
441 D E T S G G V Y H K V T G E V F P E M V
1321 GATGAGACCAGCGGCGGAGTTTATCACAAGGTTACCGGAGAGGTGTTCCCTGAGATGGTG
461 A A V E E T A Q M I L S P I S N T A T G
1381 GCTGCTGTTGAGGAGACTGCACAGATGATACTCTCACCTATATCCAACACAGCAACAGGC
481 D F A A V M A K A S V V Y R K Y D A A F
1441 GATTTTGCGGCTGTTATGGCAAAGGCTTCAGTTGTATACAGAAAGTACGATGCAGCATTT
501 A D S C L A A A Q K A W K Y L E Q H Q G
1501 GCAGATTCATGTCTTGCTGCAGCTCAGAAGGCATGGAAGTATCTTGAGCAGCATCAGGGA
521 D A G F K N V G S I V T G E Y P D S N D
1561 GATGCAGGCTTTAAGAACGTTGGAAGCATAGTTACAGGTGAGTATCCTGACAGCAATGAC
541 S D E Y L W A A A E L Y I A T G D E S Y
1621 AGCGATGAGTACCTGTGGGCAGCAGCGGAGCTGTATATCGCAACAGGTGATGAGAGCTAC
561 N D Y V K T A I E G S V K Y G L G W A D
1681 AATGATTATGTGAAGACAGCCATCGAGGGAAGCGTTAAGTACGGCCTTGGATGGGCGGAT
581 V G Y Y G I Y D Y C V N V K D C A A E K
1741 GTCGGATATTATGGAATATATGACTACTGTGTGTAAATGTAAAGGACTGCGCAGCAGAGAAG
601 E I L K K G A D K L V D N Y A G S G F G
1801 GAGATACTTAAGAAGGGTGCTGACAAGCTTGTAGATAACTATGCCGGCAGCGGCTTCGGT
621 S T T G G S Y V W G S N M V V A D N G I

```
1861 TCAACAACAGGCGGCTCTTATGTATGGGGAAGCAACATGGTCGTTGCGGATAACGGAATC
 641 L  L  L  M  A  S  K  V  L  G  D  D  S  Y  V  D  Y  A  A  D
1921 CTTCTGCTCATGGCATCAAAGGTTCTGGGCGACGATTCATATGTGGACTATGCGGCAGAC
 661 Q  L  N  Y  I  L  G  R  N  A  V  S  Y  C  Y  V  T  G  Y  G
1981 CAGCTCAACTATATACTTGGAAGAAATGCAGTCAGCTACTGTTATGTGACAGGATATGGT
 681 S  Q  T  P  E  N  P  H  H  R  P  S  E  A  L  G  K  A  M  P
2041 TCACAGACTCCTGAGAATCCACACCACAGACCTTCAGAGGCACTGGGCAAGGCAATGCCT
 701 G  M  L  V  G  G  A  D  G  N  L  E  D  P  Y  A  K  A  V  L
2101 GGAATGCTTGTGGGCGGTGCCGATGGAAACCTTGAGGATCCATATGCGAAGGCTGTTCTT
 721 A  K  I  P  K  E  R  C  Y  V  D  N  A  Q  S  Y  S  C  N  E
2161 GCAAAGATTCCAAAGGAGAGATGCTATGTCGACAACGCACAGAGTTATTCATGTAATGAG
 741 V  T  I  Y  W  N  S  P  L  V  Y  L  L  A  N  F  K  *
2221 GTAACGATCTACTGGAATTCACCTCTTGTATATCTTCTTGCGAACTTCAAGTAG
```

caaatatttaagaatatgtaactataattagtacatcatatttgatataactattgaata
atgtaaggaatcggtatggaaatatgaatgaatatgtagacaaaataaaaaagcttatg
ggagaaagagat

Appendix 3.3 Sequence of ART_GH9/S (putative cellulase, 2274 bp). The open reading frame and translation (757 amino acids) is shown in one letter code. The start (at position 1) and stop codon (at position 2271) are bold and underlined. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH9 catalytic domain are in italics and bold. The conserved aspartate and glutamate residues are in red. Primers used in this study are double underlined.

Domains are indicated as follows:
Signal peptide; 1 - 23 CBM_4_9: 68 – 218 (PF02018), Cel_D Ig-like domain: 238 - 310 (PF02927), Family 9 catalytic domain: 323 – 754 (PF00759).


agtgcacttgatatttttggaggccatatgaa
gtatcaggatgttttgcagtattcagatcagtaatgaagtgttgaatgaacacagagttta
tatgtacagttgagaaggaaggtacatagcaatatagaaaacatggagggacaataagac
1 M R K M K K V A V M C L T A A M T M S L
1 ATGAGAAAAATGAAGAAGGTTGCAGTCATGTGCCTGACAGCAGCGATGACGATGTCGTTG
21 L A G C G G A K K D T D G S T S E T E V
61 CTTGCAGGCTGTGGCGGAGCAAAGAAGGATACCGATGGTTCGACTTCTGAGACAGAAGTG
41 T T E V S T E E E T T E A T T E E A T T
121 ACAACAGAGGTATCAACCGAAGAAGAGACAACCGAAGCTACGACCGAAGAAGCAACAACA
61 E E E L G V S D G D K V I N I N F D D K
181 GAGGAGGAACTCGGGGTGTCAGACGGCGATAAGGTGATCAATATCAACTTTGATGATAAA
81 D T D G F H T Y T N G G N E E M T N E D
241 GATACCGATGGTTTCCATACTTATACAAATGGTGGAAATGAAGAAATGACCAATGAAGAT
101 G E L R I N I K K T G S V D Y A N Q I Y
301 GGAGAACTTCGTATCAATATTAAGAAGACGGGAAGCGTAGACTATGCGAACCAGATCTAT
121 Y D G F R L Y Q G C V Y E Y S F D V R S
361 TATGATGGATTCCGTCTGTATCAGGGTTGTGTCTATGAATATTCCTTCGATGTCCGTTCT
141 D L E R T I E W R L Q I N G G D Y H A Y
421 GATCTGGAACGTACGATCGAATGGAGATTACAGATCAATGGTGGAGATTATCATGCATAC
161 T S D V I T I G P E T Q H I T A Q F K M
481 ACCAGTGATGTGATCACGATCGGACCGGAGACACAGCATATCACAGCACAGTTTAAGATG
181 E E D S D P A P R L C F N M G K Q E G M
541 GAGGAGGATTCCGATCCTGCACCAAGACTTTGCTTCAACATGGGAAAACAGGAAGGTATG
201 T G D E A E H N I Y F D N I L L E A V D
601 ACCGGAGATGAAGCAGAGCATAATATTTATTTTGATAATATTTTATTAGAGGCAGTTGAC
221 A S G A Q Q V E A T P D P M A I N V N Q
661 GCTTCCGGCGCACAGCAGGTAGAAGCAACACCTGATCCTATGGCTATCAATGTGAACCAG
241 V G Y M T G D S K V A T V I G K N A K S
721 GTGGGCTACATGACCGGTGACTCCAAGGTCGCTACCGTGATCGGAAAAAATGCGAAATCA
261 F E V I D V A S G K S V Y S A D L P E E
781 TTTGAAGTAATTGATGTGGCATCAGGAAAATCTGTATATTCAGCAGATCTGCCGGAAGAA
281 A T Y D P P S E M F C K Q A D F S S V K
841 GCTACCTACGATCCACCATCAGAGATGTTCTGTAAGCAGGCGGATTTCTCATCCGTAAAG
301 D A G T Y K I K T D D G E E S V E F K I
901 GATGCAGGAACCTATAAGATCAAGACAGATGATGGAGAAGAATCAGTGGAATTCAAGATC
321 G D D I Y G D L Y K D V V L M L Y N Q R
961 GGTGACGACATTTACGGAGATCTGTATAAGGATGTCGTGCTGATGCTTTATAATCAGCGT
341 C G V E L D S S I A G E F A H P A C H T
1021 TGCGGTGTGGAGCTTGATAGCAGCATTGCAGGTGAATTTGCACATCCTGCCTGCCATACA
361 G E A V V Y G T D K K I D V T G G W H D
1081 GGTGAAGCGGTTGTATACGGAACAGATAAGAAGATCGATGTAACAGGCGGCTGGCATGAT
381 A G D Y G R Y V V S G A K T V Q D L F L
1141 GCCGGTGATTATGGACGCTATGTTGTATCAGGTGCAAAGACAGTACAGGATCTGTTCTTA
401 T Y E D N G Y K A D D I G I P E S G N G
1201 ACATATGAAGACAATGGATATAAGGCAGATGACATCGGTATTCCGGAGAGCGGAAACGGT
421 V P D V L D E A R F E L D W L L K M Q D
1261 GTGCCGGATGTCTTAGATGAAGCAAGATTTGAGCTGGATTGGTTGCTCAAGATGCAGGAT
441 A N G G V Y H K V T C D V F P E T V M P
1321 GCAAATGGTGGTGTATATCACAAGGTAACCTGTGATGTATTCCCTGAGACAGTTATGCCG
461 E E E T A Q L I A C P V S N T A T G D F
1381 GAAGAAGAAACCGCACAGCTGATCGCATGTCCGGTTTCAAATACAGCGACCGGAGATTTT
481 A A V M A K A S V L Y K D Y D A D F A A
1441 GCGGCTGTTATGGCAAAGGCTTCAGTTCTTTATAAAGATTACGATGCGGATTTTGCAGCA
501 K C L D A S K K A Y E Y L S G N M D A Y
1501 AAATGTCTGGATGCATCAAAGAAGGCATATGAATACCTGAGCGGTAATATGGATGCTTAC
521 G F S N P D D I V T G E Y P D T I F R D
1561 GGATTCTCAAATCCGGACGATATCGTAACAGGTGAATATCCGGATACGATCTTCCGTGAT
541 E T I W A A V E L Y A A T G D A S Y K K
1621 GAGACGATCTGGGCAGCGGTAGAATTATATGCAGCAACCGGAGATGCTTCCTATAAGAAG
561 A A D A I I E G S D I V N Y G L G W A D
1681 GCTGCCGATGCGATCATTGAGGGTAGCGATATCGTAAACTATGGACTTGGATGGGCAGAT
581 V G Y Y A L Y D Y I K Y D G G S A K A N
1741 GTCGGATATTACGCTCTGTATGACTACATCAAATATGACGGCGGTTCTGCCAAGGCAAAC
601 E L F F N E V D K T V D Q I K D N G F G
1801 GAATTGTTCTTCAATGAAGTAGATAAGACCGTAGATCAGATTAAAGACAACGGTTTCGGT
621 V W V S P T K T F A W G S N M N I A N K
1861 GTATGGGTATCACCAACCAAGACATTTGCATGGGGAAGCAATATGAATATTGCAAATAAA

```
 641 G  M  L  L  L  M  A  N  K  L  K  P  N  A  D  Y  V  K  Y  A
1921 GGAATGTTACTTCTGATGGCAAATAAGCTTAAGCCAAATGCAGATTATGTAAAATATGCT
 661 S  Y  Q  R  D  Y  L  L  G  R  N  A  V  G  Y  C  Y  V  T  G
1981 TCATACCAAAGAGACTATCTGCTTGGAAGAAATGCAGTTGGCTACTGTTATGTAACAGGC
 681 Y  G  T  R  T  P  L  H  P  H  H  R  P  S  Q  V  L  E  V  A
2041 TATGGTACAAGAACTCCGCTTCATCCGCACCACAGACCATCTCAGGTATTAGAGGTAGCA
 701 M  P  G  M  L  V  G  G  A  D  S  N  L  E  D  P  Y  A  K  A
2101 ATGCCGGGTATGTTAGTAGGTGGCGCAGACAGTAATCTGGAAGATCCATATGCAAAGGCT
 721 V  L  L  G  K  A  P  E  S  R  Y  A  D  N  A  Q  S  F  S  C
2161 GTTCTTCTGGGCAAAGCACCGGAGAGCAGATATGCAGACAATGCGCAGTCATTCTCCTGT
 741 N  E  V  T  I  Y  W  N  S  P  M  I  Y  L  M  S  G  L  G  N
2221 AATGAGGTGACGATCTATTGGAATTCTCCGATGATCTATCTGATGTCAGGACTTGGCAAC
 761 T  K  *
2281 ACGAAATAA

acacacaaaataacacacacataaatagcacagccgatcgggataactgttcggctgtag
ctggtataagagagttttcattatggacaaaataaagaagttattaaaaaaagaaattgt
cagatatgtaatagcaggt
```

Appendix 3.4 Sequence of L250_GH9/S (putative cellulase, 2289 bp). The open reading frame and translation (762 amino acids) is shown in one letter code. The start (at position 1) and stop codon (at position 2287) are bold and underlined. An open box indicates the potential RBS. A likely promoter region is shown in blue. Signature sequences for GH9 catalytic domain are in italics and bold. The conserved aspartate and glutamate residues are in red. Primers used in this study are double underlined.

Domains are indicated as follows:
Signal peptide: 1- 23 CBM_4_9: 69 – 196 (PF02018), Cel_D Ig-like domain: 229 - 311 (PF02927), Family 9 catalytic domain: 326 – 756 (PF00759).

Accession number	Origin	Domain structure	Theme
GUN7_ARATH	*Arabidopsis thaliana*	GH9	A
GUN25_ARATH	*Arabidopsis thaliana*	GH9	A
GUN11 _ARATH	*Arabidopsis thaliana*	SP-GH9	A
O65186 _FRAAN	*Fragana ananassa*	SP-GH9-CBM49	A
GUN15_ARATH	*Arabidopsis thaliana*	SP-GH9	A
Q43105_PHAVU	*Phaseolus vulgaris*	SP-GH9	A
Q43149_SAMNI	*Sambucus nigra*	SP-GH9	A
O4972_SOLLC	*Solanium lycopersicum*	SP-GH9	A
Q96546_CAPAN	*Citrus sinensis*	SP-GH9	A
GUN1_PERAE	*Persea americana*	SP-GH9	A
Q9ZTL0_FRAAN	*Fragana ananassa*	SP-GH9	A
O64402_PINRA	*Pinus radiate*	SP-GH9	A
O64401_PINRA	*Pinus radiate*	SP-GH9	A
GUN17_ARATH	*Arabidopsis thaliana*	GH9	A
Q42875_SOLLC	*Solanium lycopersicum*	SP-GH9	A
GUN4_BACS5	*Bacillus* sp	GH9-CBM3	B
Q9Z4I1_PAEBA	*Paenibacillus barcinonensis*	SP-GH9-CBM3-FN-CBM3	B
GUNB_CELFI	*Cellulomonas fimi*	SP-GH9-CBM3-(Fn)3-CBM2	B
GUNF_CLOTH	*Cl. thermocellum*	SP-GH9-CBM3-D-D	B
GUNZ_CLOSR	*Cl. stercorarium*	SP-GH9-CBM3-(DUF291)2-CBM3	B
O65987_CLOCL	*Cl. cellulolyticum*	SP-GH9-CBM3-D-D	B
GUNI_CLOTH	*Cl. thermocellum*	GH9-CBM3-CBM3	B
GUNA_CALSA	*Caldocellum saccharolyticum*	SP-GH9-(CBM3)3-GH48	B
GUNG_CLOCE	*Cl. cellulovorans*	SP-GH9-CBM3-D-D	B
GUNC-CELFI	*Cellulomonas fimi*	SP-CBM4-CBM4-CelD-GH9-i-SET	C
GUN1_STRRE	*Streptomyces reticuli*	SP-CelD- CBM4_9-GH9	C
GUNA_CELJU	*Cellvibrio japanicus*	Ig-GH9-CBM10_PKD-CBM2	C
GUND_CLOTM	*Cl. thermocellum*	Ig-GH9-D-D	C
Q9Z3X7_9PROT	*Pseudomonas* sp. YD-15	SP-CelD-GH9	C
Q59325_CLOTM	*Cl. thermocellum*	SP-CBM4_9-CelD-GH9-CBM3-D-D	C
P77864_FIBSS	*Fibrobacter succinogenes*	SP-CelD-GH9	C
Q59442_FIBSU	*Fibrobacter succinogenes*	SP-CelD-GH9	C
CBK83282.1	*Coprococcus eutactus*. ART55/1	SP- CBM4_9- CelD-GH9	D
ZP_02206240.1	*C. eutactus* ATCC 27759	SP- CBM4_9- CelD-GH9	D
ZP_02074498.1	*Coprococcus* sp. L2-50	SP- CBM4_9- CelD-GH9	D
ZP_07838541.1	*Eubacterium cellulosolvens* 6	SP- CBM4_9- CelD-GH9	D
YP_003842993.1	*Cl. cellulovorans* 743B	SP- CBM4_9- CelD-GH9	D
YP_004307606.1	*Cl. lentocellum* DSM 5427	SP-CelD-CBM4_9-GH9-CBM3	D
YP_004366651.1	*Treponema succinifaciens* DSM 2489	SP- CBM4_9- CelD-GH9	D
YP_003842992.1	*Cl.cellulovorans* 743B	SP- CBM4_9- CelD-GH9	D
ZP_06141671.1	*Ruminococcus flavefaciens* FD-1	SP-CelD-GH9	D
ACZ98604.1	*Cellulosilyticum ruminicola*	SP- CBM4_9- CelD-GH9	D
CBL16782.1	*Ruminococcus champanellenis* 18P13	SP- CBM4_9- CelD-GH9-D-GH16	D
AEJ61964.1	*Spirochaeta thermophila* DSM 6578	SP- CBM4_9- CelD-GH9	D
YP_003874856.1	*Spirochaeta thermophila* DSM 6192	SP- CBM4_9- CelD-GH9	D
ZP_07837973.1	*Eubacterium cellulosolvens* 6	SP- CBM4_9- CelD-GH9	D
CBK74409.1	*Butyrivibrio fibrisolvens* 16/4	GH9	D
ZP_07897289.1	*Paenibacillus vortex* V453	SP-Ig-GH9	D
YP_003010267.1	*Paenibacillus* sp. JDR-2	SP-Ig-GH9	D
YP_002506548.1	*Cl. cellulolyticum* H10	SP-Ig-GH9	D
CBK83841.1	*Coprococcus eutactus*. ART55/1	SP-GH9-Ig	E
ZP_02205610.1	*C.eutactus* ATCC 27759	SP-GH9-Ig	E
ZP_02073894.1	*Coprococcus* sp. L2-50	SP-CBM2-GH9-Ig-Ig-Ig	E
ZP_06250300.1	*Cl. thermocellum* JW20	SP-GH9-D-D	E
ZP_08194703.1	*Cl. papyrosolvens* DSM 2782	SP-GH9	E
YP_001039204.1	*Cl.thermocellum* ATCC 27405	SP-GH9-D-D	E
ZP_05428039.1	*Cl.thermocellum* DSM 2360	SP-GH9-D-D	E
YP_003844052.1	*Cl. cellulovorans* 743B	SP-GH9-D	E
YP_002505111.1	*Cl. cellulolyticum* H10	SP-GH9	E
ZP_06144437.1	*Ruminococcus flavefaciens* FD-1	SP- GH9	E
CBL16391.1	*Ruminococcus champanellensis*. 18P13	SP-GH9	E
ZP_07327657.1	*Acetivibrio cellulolyticus* CD2	SP-GH9-D	E
ZP_06142866.1	*Ruminococcus flavefaciens* FD-1	SP-GH9	E
YP_004104517.1	*Ruminococcus albus* 7	SP-GH9	E

Appendix 3.5 Modular architecture of enzymes from GH9 family

		ART_GH9/S			ART_GH9/L			L250_GH9/S			L250_GH9/L			*E.coli*		*L. lactis*	
AA	**Code**	No	/1000	F	No	/1000	F	No	/1000	F	No	/1000	F	/1000	F	/1000	F
Ala	**GCA**	33	43.54	0.40	73	40.94	0.41	43	56.36	0.61	107	59.12	0.64	20.56	0.22	21.50	0.35
Ala	**GCC**	12	15.83	0.14	21	11.78	0.12	4	5.24	0.06	9	4.97	0.05	24.97	0.27	9.80	0.16
Ala	**GCG**	18	23.75	0.22	49	27.48	0.27	11	14.42	0.15	28	15.47	0.17	32.20	0.34	6.50	0.11
Ala	**GCT**	20	26.39	0.24	37	20.75	0.21	13	17.04	0.18	24	13.26	0.14	16.22	0.17	23.30	0.38
Arg	**AGA**	11	14.51	1.00	17	9.53	0.65	9	11.80	0.56	12	6.63	0.34	2.72	0.05	10.60	0.29
Arg	**AGG**	0	0.00	0.00	5	2.80	0.19	0	0.00	0.00	6	3.31	0.17	1.63	0.03	2.10	0.06
Arg	**CGA**	0	0.00	0.00	1	0.56	0.04	0	0.00	0.00	3	1.66	0.09	3.66	0.07	6.50	0.18
Arg	**CGC**	0	0.00	0.00	0	0.00	0.00	1	1.31	0.06	1	0.55	0.03	21.15	0.38	3.60	0.10
Arg	**CGG**	0	0.00	0.00	1	0.56	0.04	0	0.00	0.00	8	4.42	0.23	5.67	0.10	2.40	0.07
Arg	**CGT**	0	0.00	0.00	2	1.12	0.08	6	7.86	0.38	5	2.76	0.14	20.70	0.37	11.70	0.32
Asn	**AAC**	16	21.11	0.57	29	16.26	0.33	9	11.80	0.25	21	11.60	0.25	21.58	0.54	14.20	0.25
Asn	**AAT**	12	15.83	0.43	58	32.53	0.67	27	35.39	0.75	64	35.36	0.75	18.68	0.46	42.90	0.75
Asp	**GAC**	34	44.85	0.45	32	17.95	0.28	15	19.66	0.21	23	12.71	0.20	19.32	0.37	14.20	0.26
Asp	**GAT**	41	54.09	0.55	81	45.43	0.72	56	73.39	0.79	93	51.38	0.80	32.32	0.63	40.80	0.74
Cys	**TGC**	4	5.28	0.29	5	2.80	0.38	4	5.24	0.33	9	4.97	0.47	6.33	0.55	1.10	0.22
Cys	**TGT**	10	13.19	0.71	8	4.49	0.62	8	10.48	0.67	10	5.52	0.53	5.17	0.45	4.10	0.78
End	**TAA**	0	0.00	0.00	1	0.56	1.00	1	1.31	1.00	1	0.55	1.00	1.98	0.62	2.20	0.64
End	**TAG**	1	1.32	1.00	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0.26	0.08	0.50	0.14
End	**TGA**	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0.97	0.30	0.80	0.22
Gln	**CAA**	0	0.00	0.00	0	0.00	0.00	1	1.31	0.06	1	0.55	0.02	14.78	0.34	33.30	0.82
Gln	**CAG**	23	30.34	1.00	37	20.75	1.00	17	22.28	0.94	46	25.41	0.98	28.82	0.66	7.20	0.18
Glu	**GAA**	3	3.96	0.06	13	7.29	0.23	36	47.18	0.64	46	25.41	0.50	39.51	0.68	55.00	0.81
Glu	**GAG**	46	60.69	0.94	44	24.68	0.77	20	26.21	0.36	46	25.41	0.50	18.45	0.32	13.10	0.19
Gly	**GGA**	30	39.58	0.41	79	44.31	0.43	28	36.70	0.41	74	40.88	0.43	8.59	0.12	21.80	0.37
Gly	**GGC**	26	34.30	0.35	48	26.92	0.26	13	17.04	0.19	36	19.89	0.21	28.63	0.39	7.60	0.13
Gly	**GGG**	2	2.64	0.03	9	5.05	0.05	1	1.31	0.01	6	3.31	0.03	11.12	0.15	6.90	0.12
Gly	**GGT**	16	21.11	0.22	46	25.80	0.25	26	34.08	0.38	58	32.04	0.33	25.14	0.34	23.30	0.39
His	**CAC**	9	11.87	0.69	3	1.68	0.27	3	3.93	0.27	1	0.55	0.11	9.63	0.43	4.80	0.26
His	**CAT**	4	5.28	0.31	8	4.49	0.73	8	10.48	0.73	8	4.42	0.89	12.68	0.57	13.50	0.74
Ile	**ATA**	19	25.07	0.61	48	26.92	0.50	0	0.00	0.00	14	7.73	0.16	5.34	0.09	15.10	0.20
Ile	**ATC**	7	9.23	0.23	36	20.19	0.38	24	31.45	0.71	51	28.18	0.59	24.63	0.41	14.60	0.19
Ile	**ATT**	5	6.60	0.16	12	6.73	0.13	10	13.11	0.29	22	12.15	0.25	29.92	0.50	45.80	0.61
Leu	**CTA**	1	1.32	0.02	2	1.12	0.02	0	0.00	0.00	0	0.00	0.00	3.98	0.04	9.90	0.10
Leu	**CTC**	10	13.19	0.22	14	7.85	0.17	2	2.62	0.05	4	2.21	0.04	10.60	0.10	6.90	0.07
Leu	**CTG**	16	21.11	0.36	22	12.34	0.27	17	22.28	0.39	28	15.47	0.30	50.94	0.49	5.10	0.05
Leu	**CTT**	18	23.75	0.40	37	20.75	0.46	13	17.04	0.30	33	18.23	0.36	11.33	0.11	21.40	0.22
Leu	**TTA**	0	0.00	0.00	2	1.12	0.02	9	11.80	0.20	17	9.39	0.18	13.83	0.13	34.50	0.36
Leu	**TTG**	0	0.00	0.00	4	2.24	0.05	3	3.93	0.07	10	5.52	0.11	13.28	0.13	18.40	0.19
Lys	**AAA**	1	1.32	0.02	25	14.02	0.16	14	18.35	0.30	57	31.49	0.34	34.38	0.75	67.20	0.79
Lys	**AAG**	43	56.73	0.98	128	71.79	0.84	32	41.94	0.70	111	61.33	0.66	11.38	0.25	17.30	0.21
Met	**ATG**	21	27.70	1.00	17	9.53	1.00	24	31.45	1.00	20	11.05	1.00	27.21	1.00	23.30	1.00

Phe	**TTC**	8	10.55	0.50	16	8.97	0.39	15	19.66	0.63	14	7.73	0.30	16.49	0.43	10.90	0.25
Phe	**TTT**	8	10.55	0.50	25	14.02	0.61	9	11.80	0.38	33	18.23	0.70	22.05	0.57	33.60	0.75
Pro	**CCA**	4	5.28	0.20	8	4.49	0.26	7	9.17	0.28	10	5.52	0.22	8.51	0.20	12.60	0.43
Pro	**CCC**	0	0.00	0.00	1	0.56	0.03	0	0.00	0.00	1	0.55	0.02	5.37	0.12	2.60	0.09
Pro	**CCG**	2	2.64	0.10	13	7.29	0.42	13	17.04	0.52	30	16.57	0.67	22.34	0.51	2.70	0.09
Pro	**CCT**	14	18.47	0.70	9	5.05	0.29	5	6.55	0.20	4	2.21	0.09	7.19	0.17	11.70	0.39
Ser	**AGC**	21	27.70	0.43	40	22.43	0.18	8	10.48	0.19	24	13.26	0.17	15.60	0.26	7.30	0.10
Ser	**AGT**	1	1.32	0.02	28	15.70	0.13	2	2.62	0.05	26	14.36	0.19	9.15	0.15	16.70	0.23
Ser	**TCA**	18	23.75	0.37	87	48.79	0.39	17	22.28	0.40	26	14.36	0.19	7.89	0.13	20.60	0.29
Ser	**TCC**	3	3.96	0.06	14	7.85	0.06	7	9.17	0.17	28	15.47	0.20	8.89	0.15	4.00	0.06
Ser	**TCG**	3	3.96	0.06	24	13.46	0.11	2	2.62	0.05	14	7.73	0.10	8.71	0.15	3.90	0.05
Ser	**TCT**	3	3.96	0.06	28	15.70	0.13	6	7.86	0.14	22	12.15	0.16	9.33	0.16	18.60	0.26
Thr	**ACA**	19	25.07	0.50	93	52.16	0.53	23	30.14	0.43	100	55.25	0.46	7.93	0.15	23.20	0.39
Thr	**ACC**	9	11.87	0.24	27	15.14	0.15	18	23.59	0.34	41	22.65	0.19	22.70	0.42	8.30	0.14
Thr	**ACG**	4	5.28	0.11	24	13.46	0.14	9	11.80	0.17	50	27.62	0.23	14.02	0.26	7.40	0.13
Thr	**ACT**	6	7.92	0.16	33	18.51	0.19	3	3.93	0.06	26	14.36	0.12	9.52	0.18	19.80	0.34
Trp	**TGG**	8	10.55	1.00	20	11.22	1.00	8	10.48	1.00	21	11.60	1.00	14.45	1.00	10.10	1.00
Tyr	**TAC**	21	27.70	0.42	30	16.83	0.35	12	15.73	0.25	14	7.73	0.19	12.30	0.42	10.00	0.25
Tyr	**TAT**	29	38.26	0.58	55	30.85	0.65	36	47.18	0.75	59	32.60	0.81	16.68	0.58	30.60	0.75
Val	**GTA**	15	19.79	0.23	52	29.16	0.34	25	32.77	0.45	61	33.70	0.40	11.23	0.16	14.00	0.23
Val	**GTC**	10	13.19	0.15	17	9.53	0.11	7	9.17	0.13	21	11.60	0.14	14.83	0.21	10.20	0.17
Val	**GTG**	17	22.43	0.26	39	21.87	0.25	13	17.04	0.24	37	20.44	0.24	25.47	0.36	8.70	0.14
Val	**GTT**	23	30.34	0.35	46	25.80	0.30	10	13.11	0.18	35	19.34	0.23	19.06	0.27	27.60	0.46

Appendix 3.6 Codons usage pattern of *E. coli*, *L. lactis* and GH9 enzymes from *Coprococcus eutactus* ART55/1 and *Coprococcus* sp. L2-50. The *E. coli* and *L. lactis* codons usage pattern was obtained from the Codon Usage Database (http://www.kazusa.or.jp/codon/). The GH9 enzyme codon analysis was done by the on-line tool available at: http://www.bioinformatics.org/sms2/index.html. No - total number of occurrences of that codon in the sequence, /1000 –frequency of the codon per 1000 codons in the coding regions of analysed data (coprococcal genes and *E. coli*/ *L. lactis* genomes). F is a frequency of the codon in the set of all codons for the same amino acid. The codons in white boxes have been shown to be problematic in previous gene expression studies.

Bibliographical references

Aakvik, T., Degnes, K.F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., Volker, U., Ellingsen, T.E., Valla, S., (2009). A Plasmid RK2-Based Broad-Host-Range Cloning Vector Useful for Transfer of Metagenomic Libraries to a Variety of Bacterial Species. *FEMS Microbiology Letters,* **296**(2),149-158.

Abell, G.C.J., Cooke, C.M., Bennett, C.N., Conlon, M.A., McOrist, A.L., (2008). Phylotypes Related to *Ruminococcus bromii* are Abundant in the Large Bowel of Humans and Increase in Response to a Diet High in Resistant Starch. *FEMS Microbiology Ecology,* **66**(3),505-515.

Adams, M.M., Allison, G.E., Verma, N.K., (2001). Type IV O Antigen Modification Genes in the Genome of *Shigella flexneri* NCTC 8296. *Microbiology,* **147**(4),851-860.

Agans, R., Rigsbee, L., Kenche, H., Michail, S., Khamis, H.J., Paliy, O., (2011). Distal Gut Microbiota of Adolescent Children is Different from that of Adults. *FEMS Microbiology Ecology,* **77**(2),404-412.

Alberto, F., Bignon, C., Sulzenbacher, G., Henrissat, B., Czjzek, M., (2004). The Three-Dimensional Structure of Invertase (β-Fructosidase) from *Thermotoga maritima* reveals a Bimodular Arrangement and an Evolutionary Relationship between Retaining and Inverting Glycosidases. *Journal of Biological Chemistry,* **279**(18),18903-18910.

Angelov, A., Mientus, M., Liebl, S., Liebl, W., (2009). A Two-Host Fosmid System for Functional Screening of (Meta)Genomic Libraries from Extreme Thermophiles. *Systematic and Applied Microbiology,* **32**(3),177-185.

Araya, T., Ishibashi, N., Shimamura, S., Tanaka, K., Takahashi, H., (1993). Genetic and Molecular Analysis of the rpoD Gene from *Lactococcus lactis*. *Bioscience, Biotechnology, and Biochemistry,* **57**(1),88-92.

Ariza, A., Eklof, J.M., Spadiut, O., Offen, W.A., Roberts, S.M., Besenmatter, W., Friis, E.P., Skjot, M., Wilson, K.S., Brumer, H., Davies, G.J., (2011). Structure and Activity of a *Paenibacillus polymyxa* Xyloglucanase from Glycoside Hydrolase Family 44. *The Journal of biological chemistry,***286**(39),33890-33900.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Dore, J., MetaHIT Consortium, Antolin, M., Artiguenave, F., Blottiere, H.M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K.U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Merieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N.,

Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S.D., Bork, P., (2011). Enterotypes of the Human Gut Microbiome. *Nature,* **473**(7346),174-180.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., Weightman, A.J., (2006). New Screening Software shows that most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras. *Applied and Environmental Microbiology,* **72**(9),5734-5741.

Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., Gordon, J.I., (2005). Host-Bacterial Mutualism in the Human Intestine. *Science (New York, N.Y.),* **307**(5717),1915-1920.

Bailey, M.T., Dowd, S.E., Galley, J.D., Hufnagle, A.R., Allen, R.G., Lyte, M., (2011). Exposure to a Social Stressor Alters the Structure of the Intestinal Microbiota: Implications for Stressor-Induced Immunomodulation. *Brain, Behavior, and Immunity,* **25**(3),397-407.

Bakolitsa, C., Xu, Q., Rife, C.L., Abdubek, P., Astakhova, T., Axelrod, H.L., Carlton, D., Chen, C., Chiu, H.J., Clayton, T., Das, D., Deller, M.C., Duan, L., Ellrott, K., Farr, C.L., Feuerhelm, J., Grant, J.C., Grzechnik, A., Han, G.W., Jaroszewski, L., Jin, K.K., Klock, H.E., Knuth, M.W., Kozbial, P., Krishna, S.S., Kumar, A., Lam, W.W., Marciano, D., McMullan, D., Miller, M.D., Morse, A.T., Nigoghossian, E., Nopakun, A., Okach, L., Puckett, C., Reyes, R., Tien, H.J., Trame, C.B., van den Bedem, H., Weekes, D., Hodgson, K.O., Wooley, J., Elsliger, M.A., Deacon, A.M., Godzik, A., Lesley, S.A., Wilson, I.A., (2010). Structure of BT_3984, a Member of the SusD/RagB Family of Nutrient-Binding Molecules. *Acta crystallographica. Section F, Structural biology and Crystallization Communications,* **66**(Pt 10),1274-1280.

Balderas-Hernandez, V.E., Sabido-Ramos, A., Silva, P., Cabrera-Valladares, N., Hernandez-Chavez, G., Baez-Viveros, J.L., Martinez, A., Bolivar, F., Gosset, G., (2009). Metabolic Engineering for Improving Anthranilate Synthesis from Glucose in *Escherichia coli*. *Microbial Cell Factories,* **8,**19.

Baneyx, F., (1999). Recombinant Protein Expression in *Escherichia coli*. *Current Opinion in Biotechnology,* **10**(5),411-421.

Baneyx, F. and Mujacic, M., (2004). Recombinant Protein Folding and Misfolding in *Escherichia coli*. *Nature Biotechnology,* **22**(11),1399-1408.

Baveye, P.C., (2009). To Sequence Or Not to Sequence the Whole-Soil Metagenome? *Nature Reviews Microbiology,* **7**(10),756.

Bayer, E.A., Lamed, R., White, B.A., Flint, H.J., (2008). From Cellulosomes to Cellulosomics. *The Chemical Record,* **8**(6),364-377.

Bensing, B.A. and Sullam, P.M., (2010). Transport of Preproteins by the Accessory Sec System Requires a Specific Domain Adjacent to the Signal Peptide. *Journal of Bacteriology,* **192**(16),4223-4232.

Berg Miller, M.E., Antonopoulos, D.A., Rincon, M.T., Band, M., Bari, A., Akraiko, T., Hernandez, A., Thimmapuram, J., Henrissat, B., Coutinho, P.M., Borovok, I., Jindou, S., Lamed, R., Flint, H.J., Bayer, E.A., White, B.A., (2009). Diversity and Strain Specificity of Plant Cell Wall Degrading Enzymes Revealed by the Draft Genome of *Ruminococcus flavefaciens* FD-1. *PloS One,* **4**(8),e6650.

Berlemont, R., Pipers, D., Delsaute, M., Angiono, F., Feller, G., Galleni, M., Power, P., (2011). Exploring the Antarctic Soil Metagenome as a Source of Novel Cold-Adapted Enzymes and Genetic Mobile Elements. *Revista Argentina de microbiologia,* **43**(2),94-103.

Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., Nikkila, J., Monti, D., Satokari, R., Franceschi, C., Brigidi, P., De Vos, W., (2010). Through Ageing, and Beyond: Gut Microbiota and Inflammatory Status in Seniors and Centenarians. *PloS One,* **5**(5),e10667.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S.D., Sorokin, A., (2001). The Complete Genome Sequence of the Lactic Acid Bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Research,* **11**(5),731-753.

Booijink, C.C.G.M., Boekhorst, J., Zoetendal, E.G., Smidt, H., Kleerebezem, M., de Vos, W.M., (2010). Metatranscriptome Analysis of the Human Fecal Microbiota Reveals Subject-Specific Expression Profiles, with Genes Encoding Proteins Involved in Carbohydrate Metabolism being Dominantly Expressed. *Applied and Environmental Microbiology,* **76**(16),5533-5540.

Bordignon, E., Grote, M., Schneider, E., (2010). The Maltose ATP-Binding Cassette Transporter in the 21st Century--Towards a Structural Dynamic Perspective on its Mode of Action. *Molecular Microbiology,* **77**(6),1354-1366.

Browning, D.F. and Busby, S.J.W., (2004). The Regulation of Bacterial Transcription Initiation. *Nature Reviews Microbiology,* **2**(1),57-65.

Bruand, C., Le Chatelier, E., Ehrlich, S.D., Janniere, L., (1993). A Fourth Class of Theta-Replicating Plasmids: The pAM Beta 1 Family from Gram-Positive Bacteria. *Proceedings of the National Academy of Sciences of the United States of America,* **90**(24),11668-11672.

Burstein, T., Shulman, M., Jindou, S., Petkun, S., Frolow, F., Shoham, Y., Bayer, E.A., Lamed, R., (2009). Physical Association of the Catalytic and Helper Modules of a Family-9 Glycoside Hydrolase is Essential for Activity. *FEBS Letters,* **583**(5),879-884.

Cani, P.D., Lecourt, E., Dewulf, E.M., Sohet, F.M., Pachikian, B.D., Naslain, D., De Backer, F., Neyrinck, A.M., Delzenne, N.M., (2009). Gut Microbiota Fermentation of Prebiotics Increases Satietogenic and Incretin Gut Peptide Production with Consequences for Appetite Sensation and Glucose Response After a Meal. *The American Journal of Clinical Nutrition,* **90**(5),1236-1243.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B., (2009). The Carbohydrate-Active EnZymes Database (CAZy): An Expert Resource for Glycogenomics. *Nucleic Acids Research,* **37**(Database issue),D233-8.

Casino, P., Rubio, V., Marina, A., (2010). The Mechanism of Signal Transduction by Two-Component Systems. *Current Opinion in Structural Biology,* **20**(6),763-771.

Chambers, S.P., Prior, S.E., Barstow, D.A., Minton, N.P., (1988). The pMTL Nic-Cloning Vectors. I. Improved pUC Polylinker Regions to Facilitate the use of Sonicated DNA for Nucleotide Sequencing. *Gene,* **68**(1),139-149.

Charrier, C., Duncan, G.J., Reid, M.D., Rucklidge, G.J., Henderson, D., Young, P., Russell, V.J., Aminov, R.I., Flint, H.J., Louis, P., (2006). A Novel Class of CoA-Transferase Involved in Short-Chain Fatty Acid Metabolism in Butyrate-Producing Human Colonic Bacteria. *Microbiology,* **152**(1),179-185.

Chassard, C., Delmas, E., Lawson, P.A., Bernalier-Donadille, A., (2008). *Bacteroides xylanisolvens* sp. nov., a Xylan-degrading Bacterium Isolated from Human Faeces. *International Journal of Systematic and Evolutionary Microbiology,* **58**(4),1008-1013.

Chassard, C., Delmas, E., Robert, C., Lawson, P.A., Bernalier-Donadille, A., (2011). *Ruminococcus champanellensis* sp. nov., a Cellulose-Degrading Bacteria from the Human Gut Microbiota. *International Journal of Systematic and Evolutionary Microbiology, .*

Chassard, C., Goumy, V., Leclerc, M., Del'homme, C., Bernalier-Donadille, A., (2007). Characterization of the Xylan-Degrading Microbial Community from Human Faeces. *FEMS Microbiology Ecology,* **61**(1),121-131.

Chen, F., Zhu, Y., Dong, X., Liu, L., Huang, L., Dai, X., (2010). Lignocellulose Degrading Bacteria and their Genes Encoding cellulase/hemicellulase in Rumen--a Review. *Wei sheng wu xue bao = Acta microbiologica Sinica,* **50**(8),981-987.

Chen, G.T. and Inouye, M., (1994). Role of the AGA/AGG Codons, the Rarest Codons in Global Gene Expression in *Escherichia coli*. *Genes & Development,* **8**(21),2641-2652.

Cheng, Y.M., Hsieh, F.C., Meng, M., (2009). Functional Analysis of Conserved Aromatic Amino Acids in the Discoidin Domain of *Paenibacillus* β-1,3-Glucanase. *Microbial Cell Factories,* **8,**62.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., (2003). Multiple Sequence Alignment with the Clustal Series of Programs. *Nucleic Acids Research,* **31**(13),3497-3500.

Chiriac, A.I., Cadena, E.M., Vidal, T., Torres, A.L., Diaz, P., Pastor, F.I., (2010). Engineering a Family 9 Processive Endoglucanase from *Paenibacillus barcinonensis* Displaying a Novel Architecture. *Applied Microbiology and Biotechnology,* **86**(4),1125-1134.

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

Christiansen, C., Abou Hachem, M., Janecek, S., Viksø-Nielsen, A., Blennow, A., Svensson, B., (2009). The Carbohydrate-Binding Module Family 20 - Diversity, Structure, and Function. *FEBS Journal,* **276**(18),5006-5029.

Claesson, M.J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J.R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A.P., Shanahan, F., Twomey, C., Hill, C., Ross, R.P., O'Toole, P.W., (2011). Composition, Variability, and Temporal Stability of the Intestinal Microbiota of the Elderly. *Proceedings of the National Academy of Sciences,* **108**(Supplement 1),4586-4591.

Collins, M.D., Lawson, P.A., Willems, A., Cordoba, J.J., Fernandez-Garayzabal, J., Garcia, P., Cai, J., Hippe, H., Farrow, J.A., (1994). The Phylogeny of the Genus Clostridium: Proposal of Five New Genera and Eleven New Species Combinations. *International Journal of Systematic Bacteriology,* **44**(4),812-826.

Coyne, M.J., Reinap, B., Lee, M.M., Comstock, L.E., (2005). Human Symbionts use a Host-Like Pathway for Surface Fucosylation. *Science (New York, N.Y.),* **307**(5716),1778-1781.

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

Crociani, F., Alessandrini, A., Mucci, M.M., Biavati, B., (1994). Degradation of Complex Carbohydrates by *Bifidobacterium* spp. *International Journal of Food Microbiology,* **24**(1-2),199-210.

Cummings, J.H. and Stephen, A.M., (2007). Carbohydrate Terminology and Classification. *European Journal of Clinical Nutrition,* **61** (Supplement 1) S5-18.

Dabek, M., McCrae, S.I., Stevens, V.J., Duncan, S.H., Louis, P., (2008). Distribution of β-Glucosidase and β-Glucuronidase Activity and of β-Glucuronidase Gene Gus in Human Colonic Bacteria. *FEMS Microbiology Ecology,* **66**(3),487-495.

Dale, J.W. and Park, S.F., 2010. *Molecular Genetics of Bacteria.* 5, illustrated edn. John Wiley and Sons, 2010.

Daniel, R., (2005). The Metagenomics of Soil. *Nature Review Microbiology,* **3**(6),470-478.

Dao, M.L. and Ferretti, J.J., (1985). *Streptococcus-Escherichia coli* Shuttle Vector pSA3 and its use in the Cloning of Streptococcal Genes. *Applied and Environmental Microbiology,* **49**(1),115-119.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., Lionetti, P., (2010). Impact of Diet in Shaping Gut Microbiota Revealed by a Comparative Study in Children from Europe and Rural Africa. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(33),14691-14696.

De Vos, W.M., (1987). Gene Cloning and Expression in Lactic Streptococci. *FEMS Microbiology Letters,* **46**(3),281-295.

Debruyne, L.K., Pinna, K., Noss Whitney, E. and Whitney, E., 2008. *Nutrition and diet therapy: principles and practice.* 7, revised edn. Cengage Learning.

Dellis, S. and Filutowicz, M., (1991). Intergation Host Factor of *Escherichia coli* Reverses the Inhibition of R6K Plasmid Replication by π Initiator Protein. *Journal of Bacteriology,* **173**(3),1279-1286.

Delzenne, N.M., Cani, P.D., Daubioul, C., Neyrinck, A.M., (2005). Impact of Inulin and Oligofructose on Gastrointestinal Peptides. *The British Journal of Nutrition,* **93** (Supplement 1) S 157-61.

Derrien, M., Vaughan, E.E., Plugge, C.M., de Vos, W.M., (2004). *Akkermansia muciniphila* gen. nov., sp. nov., a Human Intestinal Mucin-Degrading Bacterium. *International Journal of Systematic and Evolutionary Microbiology,* **54**(5),1469-1476.

Devillard, E., Goodheart, D.B., Karnati, S.K.R., Bayer, E.A., Lamed, R., Miron, J., Nelson, K.E., Morrison, M., (2004). *Ruminococcus albus* 8 Mutants Defective in Cellulose Degradation are Deficient in Two Processive Endocellulases, Cel48A and Cel9B, both of which Possess a Novel Modular Architecture. *The Journal of Bacteriology,* **186**(1),136-145.

Dimroth, P., Jockel, P., Schmid, M., (2001). Coupling Mechanism of the Oxaloacetate Decarboxylase Na^+ Pump. *Biochimica et Biophysica Acta (BBA) - Bioenergetics,* **1505**(1),1-14.

Ding, S.Y., Bayer, E.A., Steiner, D., Shoham, Y., Lamed, R., (1999). A Novel Cellulosomal Scaffoldin from *Acetivibrio cellulolyticus* that Contains a Family 9 Glycosyl Hydrolase. *Journal of Bacteriology,* **181**(21),6720-6729.

Dodd, D., Mackie, R.I., Cann, I.K.O., (2011). Xylan Degradation, a Metabolic Property Shared by Rumen and Human Colonic Bacteroidetes. *Molecular Microbiology,* **79**(2),292-304.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., Knight, R., (2010). Delivery Mode Shapes the Acquisition and Structure of the Initial Microbiota Across Multiple Body Habitats in Newborns. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(26),11971-11975.

Duan, C.J., Xian, L., Zhao, G.C., Feng, Y., Pang, H., Bai, X.L., Tang, J.L., Ma, Q.S., Feng, J.X., (2009). Isolation and Partial Characterization of Novel Genes Encoding Acidic Cellulases from Metagenomes of Buffalo Rumens. *Journal of Applied Microbiology,* **107**(1),245-256.

Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.P., Belaich, J.P., Davies, G.J., Haser, R., (1995). Crystal Structure of the Catalytic Domain of a Bacterial Cellulase Belonging to Family 5. *Structure (London, England : 1993),* **3**(9),939-949.

Dudley, E. and Steele, J., (2001). *Lactococcus lactis* LM0230 Contains a Single Aminotransferase Involved in Aspartate Biosynthesis, which is Essential for Growth in Milk. *Microbiology,* **147,**215-224.

Duncan, S.H., Lobley, G.E., Holtrop, G., Ince, J., Johnstone, A.M., Louis, P., Flint, H.J., (2008). Human Colonic Microbiota Associated with Diet, Obesity and Weight Loss. *International Journal of Obesity (2005),* **32**(11),1720-1724.

Duncan, S.H., Louis, P., Flint, H.J., (2007). Cultivable Bacterial Diversity from the Human Colon. *Letters in Applied Microbiology,* **44**(4),343-350.

Duncan, S.H., Barcenilla, A., Stewart, C.S., Pryde, S.E., Flint, H.J., (2002). Acetate Utilization and Butyryl Coenzyme A (CoA):Acetate-CoA Transferase in Butyrate-Producing Bacteria from the Human Large Intestine. *Applied and Environmental Microbiology,* **68**(10),5186-5190.

Duncan, S.H., Belenguer, A., Holtrop, G., Johnstone, A.M., Flint, H.J., Lobley, G.E., (2007). Reduced Dietary Intake of Carbohydrates by Obese Subjects Results in Decreased Concentrations of Butyrate and Butyrate-Producing Bacteria in Feces. *Applied and Environmental Microbiology,* **73**(4),1073-1078.

Duncan, S.H., Louis, P., Flint, H.J., (2004). Lactate-Utilizing Bacteria, Isolated from Human Feces, that Produce Butyrate as a Major Fermentation Product. *Applied and Environmental Microbiology,* **70**(10),5810-5817.

Dusko Ehrlich, S. and MetaHIT consortium, (2010). Metagenomics of the Intestinal Microbiota: Potential Applications. *Gastroenterologie Clinique et Biologique,* **34** (Supplement 1) S23-8.

Edberg, S.C. and Kontnick, C.M., (1986). Comparison of β-Glucuronidase-Based Substrate Systems for Identification of *Escherichia coli*. *Journal of Clinical Microbiology,* **24**(3),368-371.

Ekinci, M.S., McCrae, S.I., Flint, H.J., (1997). Isolation and Overexpression of a Gene Encoding an Extracellular Beta-(1,3-1,4)-Glucanase from *Streptococcus bovis* JB1. *Applied and Environmental Microbiology,* **63**(10),3752-3756.

Esbelin, J., Martin, C., Forano, E., Mosoni, P., (2009). Differential Translocation of Green Fluorescent Protein Fused to Signal Sequences of *Ruminococcus albus* Cellulases by the Tat and Sec Pathways of *Escherichia coli*. *FEMS Microbiology Letters,* **294**(2),239-244.

Feng, Y., Duan, C., Pang, H., Mo, X., Wu, C., Yu, Y., Hu, Y., Wei, J., Tang, J., Feng, J., (2007). Cloning and Identification of Novel Cellulase Genes from Uncultured Microorganisms in Rabbit Cecum and Characterization of the Expressed Cellulases. *Applied Microbiology and Biotechnology,* **75**(2),319-328.

Ferrer, M., Golyshina, O.V., Chernikova, T.N., Khachane, A.N., Reyes-Duarte, D., Santos, V.A.P.M.D., Strompl, C., Elborough, K., Jarvis, G., Neef, A., Yakimov, M.M., Timmis, K.N., Golyshin, P.N., (2005). Novel Hydrolase Diversity Retrieved from a Metagenome Library of Bovine Rumen Microflora. *Environmental Microbiology,* **7**(12),1996-2010.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., (2010). The Pfam Protein Families Database. *Nucleic Acids Research,* **38** (Supplement 1),D211-D222.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., White, B.A., (2008). Polysaccharide Utilization by Gut Bacteria: Potential for New Insights from Genomic Analysis. *Nature Reviews Microbiology,* **6**(2),121-131.

Flint, H.J., Whitehead, T.R., Martin, J.C., Gasparic, A., (1997). Interrupted Catalytic Domain Structures in Xylanases from Two Distantly Related Strains of *Prevotella ruminicola*. *Biochimica et Biophysica Acta,* **1337**(2),161-165.

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

Frank, J.A., Reich, C.I., Sharma, S., Weisbaum, J.S., Wilson, B.A., Olsen, G.J., (2008). Critical Evaluation of Two Primers Commonly used for Amplification of Bacterial 16S rRNA Genes. *Applied and Environmental Microbiology,* **74**(8),2461-2470.

Frese, S.A., Benson, A.K., Tannock, G.W., Loach, D.M., Kim, J., Zhang, M., Oh, P.L., Heng, N.C., Patil, P.B., Juge, N., Mackenzie, D.A., Pearson, B.M., Lapidus, A., Dalin, E., Tice, H., Goltsman, E., Land, M., Hauser, L., Ivanova, N., Kyrpides, N.C., Walter, J., (2011). The Evolution of Host Specialization in the Vertebrate Gut Symbiont *Lactobacillus reuteri*. *PLoS Genetics,* **7**(2),e1001314.

Frey, P., (1996). The Leloir Pathway: A Mechanistic Imperative for Three Enzymes to Change the Stereochemical Configuration of a Single Carbon in Galactose. *The FASEB Journal,* **10**(4),461-470.

Fuglsang, A., (2003). Lactic Acid Bacteria as Prime Candidates for Codon Optimization. *Biochemical and Biophysical Research Communications,* **312**(2),285-291.

Gabor, E.M., Alkema, W.B., Janssen, D.B., (2004). Quantifying the Accessibility of the Metagenome by Random Expression Cloning Techniques. *Environmental Microbiology,* **6**(9),879-886.

Gamauf, C., Marchetti, M., Kallio, J., Puranen, T., Vehmaanpera, J., Allmaier, G., Kubicek, C.P., Seiboth, B., (2007). Characterization of the bga1-Encoded Glycoside Hydrolase Family 35 β-Galactosidase of *Hypocrea jecorina* with Galacto-β-D-Galactanase Activity. *FEBS Journal*, **274**(7),1691-1700.

Gasparic, A., Martin, J., Daniel, A.S., Flint, H.J., (1995). A Xylan Hydrolase Gene Cluster in *Prevotella ruminicola* B(1)4: Sequence Relationships, Synergistic Interactions, and Oxygen Sensitivity of a Novel Enzyme with Exoxylanase and β-(1,4)-Xylosidase Activities. *Applied and Environmental Microbiology,* **61**(8),2958-2964.

Gasson, M.J., (1983). Plasmid Complements of *Streptococcus lactis* NCDO 712 and Other Lactic Streptococci After Protoplast-Induced Curing. *Journal of Bacteriology,* **154**(1),1-9.

Gasson, M.J., Swindell, S., Maeda, S., Dodd, H.M., (1992). Molecular Rearrangement of Lactose Plasmid DNA Associated with High-Frequency Transfer and Cell Aggregation in *Lactococcus lactis* 712. *Molecular Microbiology,* **6**(21),3213-3223.

Gaudin, C., Belaich, A., Champ, S., Belaich, J., (2000). CelE, a Multidomain Cellulase from *Clostridium cellulolyticum*: A Key Enzyme in the Cellulosome? *The Journal of Bacteriology,* **182**(7),1910-1915.

Gebler, J.C., Aebersold, R., Withers, S.G., (1992). Glu-537, Not Glu-461, is the Nucleophile in the Active Site of (Lac Z) β-Galactosidase from *Escherichia coli*. *The Journal of Biological Chemistry,* **267**(16),11126-11130.

Geertsma, E.R. and Poolman, B., (2007). High-Throughput Cloning and Expression in Recalcitrant Bacteria. *Nature Methods,* **4**(9),705-707.

Gerber S.D., Solioz M., (2007). Efficient Transformation of *Lactococcus lactis* IL1403 and Generation of Knock-Out Mutants by Homologous Recombination. *Journal of Basic Microbiology,* **47**(3),281-286.

Gilad, R., Rabinovich, L., Yaron, S., Bayer, E.A., Lamed, R., Gilbert, H.J., Shoham, Y., (2003). CelI, a Noncellulosomal Family 9 Enzyme from *Clostridium thermocellum,* is a Processive Endoglucanase that Degrades Crystalline Cellulose. *Journal of Bacteriology,* **185**(2),391-398.

Gilbert, H.J., Stålbrand, H., Brumer, H., (2008). How the Walls Come Crumbling Down: Recent Structural Biochemistry of Plant Polysaccharide Degradation. *Current Opinion in Plant Biology,* **11**(3),338-348.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., Nelson, K.E., (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. *Science,* **312**(5778),1355-1359.

Glogauer, A., Martini, V.P., Faoro, H., Couto, G.H., Muller-Santos, M., Monteiro, R.A., Mitchell, D.A., de Souza, E.M., Pedrosa, F.O., Krieger, N., (2011). Identification and Characterization of a New True Lipase Isolated through Metagenomic Approach. *Microbial Cell Factories*, **10,**54.

Gloster, T.M., Turkenburg, J.P., Potts, J.R., Henrissat, B., Davies, G.J., (2008). Divergence of Catalytic Mechanism within a Glycosidase Family Provides Insight into Evolution of Carbohydrate Metabolism by Human Gut Flora. *Chemistry & Biology*, **15**(10),1058-1067.

Gloux, K., Berteau, O., El oumami, H., Béguet, F., Leclerc, M., Doré, J., (2011). A Metagenomic β-Glucuronidase Uncovers a Core Adaptive Function of the Human Intestinal Microbiome. *Proceedings of the National Academy of Sciences,* **108** (Supplement 1),4539-4546.

Goodman, A.L., Kallstrom, G., Faith, J.J., Reyes, A., Moore, A., Dantas, G., Gordon, J.I., (2011). Extensive Personal Human Gut Microbiota Culture Collections Characterized and Manipulated in Gnotobiotic Mice. *Proceedings of the National Academy of Sciences of the United States of America,* **108**(15),6252-6257.

Gosalbes, M.J., Durban, A., Pignatelli, M., Abellan, J.J., Jimenez-Hernandez, N., Perez-Cobas, A.E., Latorre, A., Moya, A., (2011). Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota. *PloS One*, **6**(3),e17447.

Goulas, T.K., Goulas, A.K., Tzortzis, G., Gibson, G.R., (2007). Molecular Cloning and Comparative Analysis of Four β-Galactosidase Genes from *Bifidobacterium bifidum* NCIMB41171. *Applied Microbiology and Biotechnology,* **76**(6),1365-1372.

Gruber, T.M. and Gross, C.A., (2003). Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. *Annual Review of Microbiology,* **57**(1),441-466.

Grynberg, M., Erlandsen, H., Godzik, A., (2003). HEPN: A Common Domain in Bacterial Drug Resistance and Human Neurodegenerative Proteins. *Trends in Biochemical Sciences*, **28**(5),224-226.

Guarner, F. and Malagelada, J.R., (2003). Gut Flora in Health and Disease. *The Lancet*, **361**(9356),512-519.

Guillen, D., Sanchez, S., Rodriguez-Sanoja, R., (2010). Carbohydrate-Binding Domains: Multiplicity of Biological Roles. *Applied Microbiology and Biotechnology*, **85**(5),1241-1249.

Han, S.O., Yukawa, H., Inui, M., Doi, R.H., (2005). Molecular Cloning and Transcriptional and Expression Analysis of engO, Encoding a New Noncellulosomal Family 9 Enzyme, from *Clostridium Cellulovorans*. *Journal of Bacteriology*, **187**(14),4884-4889.

Han, S.O., Yukawa, H., Inui, M., Doi, R.H., (2004). Isolation and Expression of the xynB Gene and its Product, XynB, a Consistent Component of the *Clostridium cellulovorans* Cellulosome. *Journal of Bacteriology,* **186**(24),8347-8355.

Hancock, K.R., Rockman, E., Young, C.A., Pearce, L., Maddox, I.S., Scott, D.B., (1991). Expression and Nucleotide Sequence of the *Clostridium acetobutylicum* β-Galactosidase Gene Cloned in *Escherichia coli*. *Journal of Bacteriology,* **173**(10),3084-3095.

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

Hardeman, F. and Sjoling, S., (2007). Metagenomic Approach for the Isolation of a Novel Low-Temperature-Active Lipase from Uncultured Bacteria of Marine Sediment. *FEMS Microbiology Ecology,* **59**(2),524-534.

Harvey, A.J., Hrmova, M., De Gori, R., Varghese, J.N., Fincher, G.B., (2000). Comparative Modeling of the Three-Dimensional Structures of Family 3 Glycoside Hydrolases. *Proteins: Structure, Function, and Bioinformatics,* **41**(2),257-269.

Harwood, C.R. and Cranenburgh, R., (2008). *Bacillus* Protein Secretion: An Unfolding Story. *Trends in Microbiology,* **16**(2),73-79.

Henrissat, B., (1991). A Classification of Glycosyl Hydrolases Based on Amino Acid Sequence Similarities. *The Biochemical Journal,* **280** (Pt 2),309-316.

Herve, C., Rogowski, A., Blake, A.W., Marcus, S.E., Gilbert, H.J., Knox, J.P., (2010). Carbohydrate-Binding Modules Promote the Enzymatic Deconstruction of Intact Plant Cell Walls by Targeting and Proximity Effects. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(34),15293-15298.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M., (2011). Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science,* **331**(6016),463-467.

Holdeman, L.V. and Moore, W.E., (1974). New Genus, *Coprococcus*, Twelve New Species, and Emended Descriptions for Four Previously Described Species of Bacteria from Human Feces. *International Journal of Systematic Bacteriology,* **24**(2),260-277.

Holt, R.A., Warren, R., Flibotte, S., Missirlis, P.I., Smailus, D.E., (2007). Rebuilding Microbial Genomes. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology,* **29**(6),580-590.

Hooper, L.V. (2004). Bacterial contributions to mammalian gut development. *Trends in Microbiology,* **12**(3),129-134.

Hooper, L.V. and Macpherson, A.J., (2010). Immune Adaptations that Maintain Homeostasis with the Intestinal Microbiota. *Nature Reviews Immunology,* **10**(3),159-169.

Hooper, L.V. and Gordon, J.I., (2001). Commensal Host-Bacterial Relationships in the Gut. *Science,* **292**(5519),1115-1118.

Hoover, R., (2010). The Impact of Heat-Moisture Treatment on Molecular Structures and Properties of Starches Isolated from Different Botanical Sources. *Critical Reviews in Food Science and Nutrition,* **50**(9),835-847.

Hosseini, E., Grootaert, C., Verstraete, W., Van de Wiele, T., (2011). Propionate as a Health-Promoting Microbial Metabolite in the Human Gut. *Nutrition Reviews,* **69**(5),245-258.

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., Schuster, S.C., (2011). Integrative Analysis of Environmental Sequences using MEGAN4. *Genome Research,* **21**(9),1552-1560.

Hutchings, M.I., Palmer, T., Harrington, D.J., Sutcliffe, I.C., (2009). Lipoprotein Biogenesis in Gram-Positive Bacteria: Knowing when to Hold 'Em, Knowing when to Fold 'Em. *Trends in Microbiology,* **17**(1),13-21.

Inan, M.S., Rasoulpour, R.J., Yin, L., Hubbard, A.K., Rosenberg, D.W., Giardina, C., (2000). The Luminal Short-Chain Fatty Acid Butyrate Modulates NF-kappaB Activity in a Human Colonic Epithelial Cell Line. *Gastroenterology,* **118**(4),724-734.

Jana, S. and Deb, J.K., (2005). Strategies for Efficient Production of Heterologous Proteins in *Escherichia coli*. *Applied Microbiology and Biotechnology,* **67**(3),289-298.

Jayasankar, N.P. and Graham, P.H., (1970). An Agar Plate Method for Screening and Enumerating Pectinolytic Microorganisms. *Canadian Journal of Microbiology,* **16**(10),1023.

Jensen, P.R. and Hammer, K., (1998). The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters. *Applied and Environmental Microbiology,* **64**(1),82-87.

Jeong, D., Chul Choi, Y., Min Lee, J., Hwan Kim, J., Lee, J., Heon Kim, K., Lee, H.J., (2006). Isolation and Characterization of Promoters from *Lactococcus lactis* ssp. *cremoris* LM0230. *Food Microbiology,* **23**(1),82-89.

Jiang, C., Ma, G., Li, S., Hu, T., Che, Z., Shen, P., Yan, B., Wu, B., (2009). Characterization of a Novel β-Glucosidase-Like Activity from a Soil Metagenome. *The Journal of Microbiology,* **47**(5),542-548.

Johansson, M.E., Larsson, J.M., Hansson, G.C., (2011). The Two Mucus Layers of Colon are Organized by the MUC2 Mucin, Whereas the Outer Layer is a Legislator of Host-Microbial Interactions. *Proceedings of the National Academy of Sciences of the United States of America,* **108** (Supplement 1)**,**4659-4665.

Johansson, M.E., Phillipson, M., Petersson, J., Velcich, A., Holm, L., Hansson, G.C., (2008). The Inner of the Two Muc2 Mucin-Dependent Mucus Layers in Colon is Devoid of Bacteria. *Proceedings of the National Academy of Sciences of the United States of America,* **105**(39),15064-15069.

Johnson, P.E., Joshi, M.D., Tomme, P., Kilburn, D.G., McIntosh, L.P., (1996). Structure of the N-Terminal Cellulose-Binding Domain of *Cellulomonas fimi* CenC Determined by Nuclear Magnetic Resonance Spectroscopy. *Biochemistry,* **35**(45),14381-14394.

Jones, B.V., Begley, M., Hill, C., Gahan, C.G.M., Marchesi, J.R., (2008). Functional and Comparative Metagenomic Analysis of Bile Salt Hydrolase Activity in the Human Gut Microbiome. *Proceedings of the National Academy of Sciences of the United States of America,* **105**(36),13580-13585.

Jones, B.V., Sun, F., Marchesi, J.R., (2007). Using Skimmed Milk Agar to Functionally Screen a Gut Metagenomic Library for Proteases may Lead to False Positives. *Letters in Applied Microbiology,* **45**(4),418-420.

Jun, H., Qi, M., Gong, J., Egbosimba, E.E., Forsberg, C.W., (2007). Outer Membrane Proteins of Fibrobacter Succinogenes with Potential Roles in Adhesion to Cellulose and in Cellulose Digestion. *The Journal of Bacteriology,* **189**(19),6806-6815.

Kageyama, A. and Benno, Y., (2000). Emendation of Genus Collinsella and Proposal of *Collinsella stercoris* sp. nov. and *Collinsella intestinalis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology,* **50**(5),1767-1774.

Kakirde, K.S., Parsley, L.C., Liles, M.R., (2010). Size does Matter: Application-Driven Approaches for Soil Metagenomics. *Soil Biology and Biochemistry,* **42**(11),1911-1923.

Karlsson, F.H., (2011). A Closer Look at Bacteroides: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut. *Microbial Ecology,* **61**(3),473-485.

Kassinen, A., Krogius-Kurikka, L., Makivuokko, H., Rinttila, T., Paulin, L., Corander, J., Malinen, E., Apajalahti, J., Palva, A., (2007). The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs significantly from that of Healthy Subjects. *Gastroenterology,* **133**(1),24-33.

Kataeva, I., Li, X.L., Chen, H., Choi, S.K., Ljungdahl, L.G., (1999). Cloning and Sequence Analysis of a New Cellulase Gene Encoding CelK, a Major Cellulosome Component of *Clostridium thermocellum*: Evidence for Gene Duplication and Recombination. *Journal of Bacteriology,* **181**(17),5288-5295.

Kataeva, I.A., Seidel, R.D.,3rd, Li, X.L., Ljungdahl, L.G., (2001). Properties and Mutation Analysis of the CelK Cellulose-Binding Domain from the *Clostridium thermocellum* Cellulosome. *Journal of Bacteriology,* **183**(5),1552-1559.

Kataeva, I.A., Seidel, R.D.,3rd, Shah, A., West, L.T., Li, X.L., Ljungdahl, L.G., (2002). The Fibronectin Type 3-Like Repeat from the *Clostridium thermocellum* Cellobiohydrolase CbhA Promotes Hydrolysis of Cellulose by Modifying its Surface. *Applied and Environmental Microbiology,* **68**(9),4292-4300.

Kataeva, I.A., Uversky, V.N., Brewer, J.M., Schubot, F., Rose, J.P., Wang, B.C., Ljungdahl, L.G., (2004). Interactions between Immunoglobulin-Like and Catalytic Modules in *Clostridium thermocellum* Cellulosomal Cellobiohydrolase CbhA. *Protein Engineering, Design & Selection : PEDS,* **17**(11),759-769.

Kay, B.K., Williamson, M.P., Sudol, M., (2000). The Importance of being Proline: The Interaction of Proline-Rich Motifs in Signaling Proteins with their Cognate Domains. *The FASEB Journal,* **14**(2),231-241.

Kelly, D., Conway, S., Aminov, R., (2005). Commensal Gut Bacteria: Mechanisms of Immune Modulation. *Trends in Immunology,* **26**(6),326-333.

Kelly, D., King, T., Aminov, R., (2007). Importance of Microbial Colonization of the Gut in Early Life to the Development of Immunity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* **622**(1-2),58-69.

Kerkhof, L.J. and Goodman, R.M., (2009). Ocean Microbial Metagenomics. *Deep-Sea Research Part II: Topical Studies in Oceanography,* **56**(19-20),1824-1829.

Kiewiet, R., Kok, J., Seegers, J.F., Venema, G., Bron, S., (1993). The Mode of Replication is a Major Factor in Segregational Plasmid Instability in *Lactococcus lactis*. *Applied and Environmental Microbiology,* **59**(2),358-364.

Kim, D., Kim, S.N., Baik, K.S., Park, S.C., Lim, C.H., Kim, J.O., Shin, T.S., Oh, M.J., Seong, C.N., (2011). Screening and Characterization of a Cellulase Gene from the Gut Microflora of Abalone using Metagenomic Library. *Journal of Microbiology (Seoul, Korea),* **49**(1),141-145.

Kosugi, A., Murashima, K., Doi, R.H., (2002). Characterization of Two Noncellulosomal Subunits, ArfA and BgaA, from *Clostridium cellulovorans* that Cooperate with the Cellulosome in Plant Cell Wall Degradation. *The Journal of Bacteriology,* **184**(24),6859-6865.

Kuhn, J., Briegel, A., Morschel, E., Kahnt, J., Leser, K., Wick, S., Jensen, G.J., Thanbichler, M., (2010). Bactofilins, a Ubiquitous Class of Cytoskeletal Proteins Mediating Polar Localization of a Cell Wall Synthase in Caulobacter Crescentus. *The EMBO Journal,* **29**(2),327-339.

Kunji, E.R.S., Slotboom, D., Poolman, B., (2003). *Lactococcus lactis* as Host for Overproduction of Functional Membrane Proteins. *Biochimica et Biophysica Acta (BBA) - Biomembranes,* **1610**(1),97-108.

Kurland, C.G., (1991). Codon Bias and Gene Expression. *FEBS Letters,* **285**(2),165-169.

Kurland, C. and Gallant, J., (1996). Errors of Heterologous Protein Expression. *Current Opinion in Biotechnology,* **7**(5),489-493.

Kurokawa, J., Hemjinda, E., Arai, T., Kimura, T., Sakka, K., Ohmiya, K., (2002). *Clostridium thermocellum* Cellulase CelT, a Family 9 Endoglucanase without an Ig-Like Domain Or Family 3c Carbohydrate-Binding Module. *Applied Microbiology and Biotechnology,* **59**(4-5),455-461.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P., Taylor, T.D., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D.S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T., Hattori, M., (2007). Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Research,* **14**(4),169-181.

Laparra, J.M. and Sanz, Y., (2010). Interactions of Gut Microbiota with Functional Food Components and Nutraceuticals. *Pharmacological Research,* **61**(3),219-225.

Le Loir, Y., Azevedo, V., Oliveira, S.C., Freitas, D.A., Miyoshi, A., Bermúdez-Humarán, L.G., Nouaille, S., Ribeiro, L.A., Leclercq, S., Gabriel, J.E., Guimaraes, V.D., Oliveira, M.N., Charlier, C., Gautier, M., Langella, P., (2005). Protein Secretion in *Lactococcus lactis*: An Efficient Way to Increase the overall Heterologous Protein Production. *Microbial Cell Factories,* **4**(1),2.

Leitch, E.C.M., Alan, W.W., Sylvia, H.D., Grietje, H., Harry, J.F., (2007). Selective Colonization of Insoluble Substrates by Human Faecal Bacteria. *Environmental Microbiology,* **9**(3),667-679.

Letunic, I., Doerks, T., Bork, P., (2009). SMART 6: Recent Updates and New Developments. *Nucleic Acids Research,* **37**(Supplement 1),D229-D232.

Li, L.L., McCorkle, S.R., Monchy, S., Taghavi, S., van der Lelie, D., (2009). Bioprospecting Metagenomes: Glycosyl Hydrolases for Converting Biomass. *Biotechnology for Biofuels,* **2,**10.

Liles, M.R., Williamson, L.L., Rodbumrer, J., Torsvik, V., Parsley, L.C., Goodman, R.M., Handelsman, J., (2009). Isolation and Cloning of High-Molecular-Weight Metagenomic DNA from Soil Microorganisms. *Cold Spring Harbor Protocols,* **2009**(8),pdb.prot5271.

Linares, D.M., Geertsma, E.R., Poolman, B., (2010). Evolved *Lactococcus lactis* Strains for Enhanced Expression of Recombinant Membrane Proteins. *Journal of Molecular Biology,* **401**(1),45-55.

Litthauer, D., Abbai, N.S., Piater, L.A., van Heerden, E., (2010). Pitfalls using Tributyrin Agar Screening to Detect Liplytic Activity in Metagenomic Studies. *African Journal of Biotechnology,* **9**(27),4282-4285.

Liu, H., Pereira, J.H., Adams, P.D., Sapra, R., Simmons, B.A., Sale, K.L., (2010). Molecular Simulations Provide New Insights into the Role of the Accessory Immunoglobulin-Like Domain of Cel9A. *FEBS letters,* **584**(15),3431-3435.

Liu, J., Liu, W.D., Zhao, X.L., Shen, W.J., Cao, H., Cui, Z.L., (2011). Cloning and Functional Characterization of a Novel Endo-β-1,4-Glucanase Gene from a Soil-Derived Metagenomic Library. *Applied Microbiology and Biotechnology,* **89**(4),1083-1092.

Liu, J., Yu, B., Zhao, X., Cheng, K., (2007). Coexpression of Rumen Microbial β-Glucanase and Xylanase Genes in *Lactobacillus reuteri*. *Applied Microbiology and Biotechnology,* **77**(1),117-124.

Lo, A.C., MacKay, R.M., Seligy, V.L., Willick, G.E., (1988). Bacillus Subtilis β-1,4-Endoglucanase Products from Intact and Truncated Genes are Secreted into the Extracellular Medium by *Escherichia coli*. *Applied and Environmental Microbiology,* **54**(9),2287-2292.

Louis, P. and Flint, H.J., (2009). Diversity, Metabolism and Microbial Ecology of Butyrate-Producing Bacteria from the Human Large Intestine. *FEMS Microbiology Letters,* **294**(1),1-8.

Louis, P., Duncan, S.H., McCrae, S.I., Millar, J., Jackson, M.S., Flint, H.J., (2004). Restricted Distribution of the Butyrate Kinase Pathway among Butyrate-Producing Bacteria from the Human Colon. *Journal of Bacteriology,* **186**(7),2099-2106.

Lozupone, C.A., Hamady, M., Cantarel, B.L., Coutinho, P.M., Henrissat, B., Gordon, J.I., Knight, R., (2008). The Convergence of Carbohydrate Active Gene Repertoires in Human Gut Microbes. *Proceedings of the National Academy of Sciences,* **105**(39),15076-15081.

Lynd, L.R., Weimer, P.J., van Zyl, W.H., Pretorius, I.S., (2002). Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiology and Molecular Biology Reviews,* **66**(3),506-77.

Macfarlane, G.T., Cummings, J.H., Allison, C., (1986). Protein Degradation by Human Intestinal Bacteria. *Journal of General Microbiology,* **132**(6),1647-1656.

MacGregor, E.A., Janeček, Š., Svensson, B., (2001). Relationship of Sequence and Structure to Specificity in the α-Amylase Family of Enzymes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology,* **1546**(1),1-20.

Machovic, M. and Janecek, S., (2006). Starch-Binding Domains in the Post-Genome Era. *Cellular and Molecular Life Sciences : CMLS,* **63**(23),2710-2724.

MacLean, D., Jones, J.D., Studholme, D.J., (2009). Application of 'Next-Generation' Sequencing Technologies to Microbial Genetics. *Nature Reviews.Microbiology,* **7**(4),287-296.

Maguin, E., Prevost, H., Ehrlich, S.D., Gruss, A., (1996). Efficient Insertional Mutagenesis in Lactococci and Other Gram-Positive Bacteria. *Journal of Bacteriology,* **178**(3),931-935.

Malinen, E., Krogius-Kurikka, L., Lyra, A., Nikkila, J., Jaaskelainen, A., Rinttila, T., Vilpponen-Salmela, T., von Wright, A.J., Palva, A., (2010). Association of Symptoms with Gastrointestinal Microbiota in Irritable Bowel Syndrome. *World Journal of Gastroenterology : WJG,* **16**(36),4532-4540.

Margolles, A. and de los Reyes-Gavilan, C.G., (2003). Purification and Functional Characterization of a Novel α-L-Arabinofuranosidase from *Bifidobacterium longum* B667. *Applied and Environmental Microbiology,* **69**(9),5096-5103.

Mariat, D., Firmesse, O., Levenez, F., Guimaraes, V., Sokol, H., Dore, J., Corthier, G., Furet, J.P., (2009). The Firmicutes/Bacteroidetes Ratio of the Human Microbiota Changes with Age. *BMC microbiology,* **9,**123.

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

Martinez, A., Kolvek, S.J., Yip, C.L.T., Hopke, J., Brown, K.A., MacNeil, I.A., Osburne, M.S., (2004). Genetically Modified Bacterial Strains and Novel Bacterial Artificial Chromosome Shuttle Vectors for Constructing Environmental Libraries and Detecting Heterologous Natural Products in Multiple Expression Hosts. *Applied and Environmental Microbiology,* **70**(4),2452-2463.

Martinez, I., Kim, J., Duffy, P.R., Schlegel, V.L., Walter, J., (2010). Resistant Starches Types 2 and 4 have Differential Effects on the Composition of the Fecal Microbiota in Human Subjects. *PloS One,* **5**(11),e15046.

Matthews, B.W., (2005). The Structure of *E. coli* β-Galactosidase. *Comptes Rendus Biologies,* **328**(6),549-556.

Matthews, P.R., Schindler, M., Howles, P., Arioli, T., Williamson, R.E., (2010). A CESA from *Griffithsia monilis* (Rhodophyta, Florideophyceae) has a Family 48 Carbohydrate-Binding Module. *Journal of Experimental Botany,* **61**(15),4461-4468.

Matuschek, M., Burchhardt, G., Sahm, K., Bahl, H., (1994). Pullulanase of *Thermoanaerobacterium thermosulfurigenes* EM1 (*Clostridium thermosulfurogenes*): Molecular Analysis of the Gene, Composite Structure of the Enzyme, and a Common Model for its Attachment to the Cell Surface. *Journal of Bacteriology,* **176**(11),3295-3302.

McBain, A.J. and Macfarlane, G.T., (1998). Ecological and Physiological Studies on Large Intestinal Bacteria in Relation to Production of Hydrolytic and Reductive Enzymes Involved in Formation of Genotoxic Metabolites. *Journal of Medical Microbiology,* **47**(5),407-416.

McIntosh, F.M., Shingfield, K.J., Devillard, E., Russell, W.R., Wallace, R.J., (2009). Mechanism of Conjugated Linoleic Acid and Vaccenic Acid Formation in Human Faecal Suspensions and Pure Cultures of Intestinal Bacteria. *Microbiology,* **155**(1),285-294.

Meilleur, C., Hupé, J., Juteau, P., Shareck, F., (2009). Isolation and Characterization of a New Alkali-Thermostable Lipase Cloned from a Metagenomic Library. *Journal of Industrial Microbiology and Biotechnology*, **36**(6),853-861.

Meinke, A., Gilkes, N.R., Kilburn, D.G., Miller, R.C.,Jr, Warren, R.A., (1993). Cellulose-Binding Polypeptides from *Cellulomonas fimi*: Endoglucanase D (CenD), a Family A β-1,4-Glucanase. *Journal of Bacteriology,* **175**(7),1910-1918.

Mello, L.V., Chen, X., Rigden, D.J., (2010). Mining Metagenomic Data for Novel Domains: BACON, a New Carbohydrate-Binding Module. *FEBS Letters*, **584**(11),2421-2426.

Mergulhão, F.J.M., Summers, D.K., Monteiro, G.A., (2005). Recombinant Protein Secretion in *Escherichia coli*. *Biotechnology Advances,* **23**(3),177-202.

Meyer, D. and Stasse-Wolthuis, M., (2009). The Bifidogenic Effect of Inulin and Oligofructose and its Consequences for Gut Health. *European Journal of Clinical Nutrition,* **63**(11),1277-1289.

Mikami, B., Iwamoto, H., Malle, D., Yoon, H.J., Demirkan-Sarikaya, E., Mezaki, Y., Katsuya, Y., (2006). Crystal Structure of Pullulanase: Evidence for Parallel Binding of Oligosaccharides in the Active Site. *Journal of Molecular Biology,* **359**(3),690-707.

Miller, T.L. and Wolin, M.J., (1996). Pathways of Acetate, Propionate, and Butyrate Formation by the Human Fecal Microbial Flora. *Applied and Environmental Microbiology,* **62**(5),1589-1592.

Mills, S., McAuliffe, O.E., Coffey, A., Fitzgerald, G.F., Ross, R.P., (2006). Plasmids of Lactococci - Genetic Accessories Or Genetic Necessities? *FEMS Microbiology Reviews,* **30**(2),243-273.

Millward-Sadler, S.J., Poole, D.M., Henrissat, B., Hazlewood, G.P., Clarke, J.H., Gilbert, H.J., (1994). Evidence for a General Role for High-Affinity Non-Catalytic Cellulose Binding Domains in Microbial Plant Ceil Wall Hydroiases. *Molecular Microbiology,* **11**(2),375-382.

Mingardon, F., Bagert, J.D., Maisonnier, C., Trudeau, D.L., Arnold, F.H., (2011). Comparison of Family 9 Cellulases from Mesophilic and Thermophilic Bacteria. *Applied and Environmental Microbiology*, **77**(4),1436-1442.

Mirande, C., Kadlecikova, E., Matulova, M., Capek, P., Bernalier-Donadille, A., Forano, E., Béra-Maillet, C., (2010). Dietary Fibre Degradation and Fermentation by Two Xylanolytic Bacteria *Bacteroides xylanisolvens* XB1AT and *Roseburia intestinalis* XB6B4 from the Human Intestine. *Journal of Applied Microbiology,* **109**(2),451-460.

Moore, W.E. and Moore, L.H., (1995). Intestinal Floras of Populations that have a High Risk of Colon Cancer. *Applied and Environmental Microbiology,* **61**(9),3202-3207.

Morello, E., Bermúdez-Humarán, L.G., Llull, D., Solé, V., Miraglio, N., Langella, P., Poquet, I., (2008). *Lactococcus lactis*, an Efficient Cell Factory for Recombinant Protein Production and Secretion. *Journal of Molecular Microbiology and Biotechnology*, **14**(1-3),48-58.

Morrison, M., Pope, P.B., Denman, S.E., McSweeney, C.S., (2009). Plant Biomass Degradation by Gut Microbiomes: More of the Same Or Something New? *Current Opinion in Biotechnology*, **20**(3),358-363.

Mueller, S., Saunier, K., Hanisch, C., Norin, E., Alm, L., Midtvedt, T., Cresci, A., Silvi, S., Orpianesi, C., Verdenelli, M.C., Clavel, T., Koebnick, C., Zunft, H.F., Dore, J., Blaut, M., (2006). Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: A Cross-Sectional Study. *Applied and Environmental Microbiology*, **72**(2),1027-1033.

Murooka, Y., Doi, N., Harada, T., (1979). Distribution of Membrane-Bound Monoamine Oxidase in Bacteria. *Applied and Environmental Microbiology*, **38**(4),565-569.

Nadkarni, M.A., Martin, F.E., Jacques, N.A., Hunter, N., (2002). Determination of Bacterial Load by Real-Time PCR using a Broad-Range (Universal) Probe and Primers Set. *Microbiology (Reading, England)*, **148**(Pt 1),257-266.

Nakamura, N., Gaskins, H.R., Collier, C.T., Nava, G.M., Rai, D., Petschow, B., Russell, W.M., Harris, C., Mackie, R.I., Wampler, J.L., Walker, D.C., (2009). Molecular Ecological Analysis of Fecal Bacterial Populations from Term Infants Fed Formula Supplemented with Selected Blends of Prebiotics. *Applied and Environmental Microbiology*, **75**(4),1121-1128.

Narushima, S., Itoha, K., Miyamoto, Y., Park, S.H., Nagata, K., Kuruma, K., Uchida, K., (2006). Deoxycholic Acid Formation in Gnotobiotic Mice Associated with Human Intestinal Bacteria. *Lipids*, **41**(9),835-843.

Natale, P., Brüser, T., Driessen, A.J.M., (2008). Sec- and Tat-Mediated Protein Secretion Across the Bacterial Cytoplasmic membrane—Distinct Translocases and Mechanisms. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1778**(9),1735-1756.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J.K., Rubin, E.M., (2006). Sequencing and Analysis of Neanderthal Genomic DNA. *Science*, **314**(5802),1113-1118.

Noonan, J.P., Hofreiter, M., Smith, D., Priest, J.R., Rohland, N., Rabeder, G., Krause, J., Detter, J.C., Paabo, S., Rubin, E.M., (2005). Genomic Sequencing of Pleistocene Cave Bears. *Science*, **309**(5734),597-599.

Nougayrede, J.P., Fernandes, P.J., Donnenberg, M.S., (2003). Adhesion of Enteropathogenic *Escherichia coli* to Host Cells. *Cellular Microbiology*, **5**(6),359-372.

O'Hara, A.M. and Shanahan, F., (2006). The Gut Flora as a Forgotten Organ. *EMBO,* **7**(7),688-692; 688.

Ohara, H., Noguchi, J., Karita, S., Kimura, T., Sakka, K., Ohmiya, K., (2000). Sequence of egV and Properties of EgV, a *Ruminococcus albus* Endoglucanase Containing a Dockerin Domain. *Bioscience, Biotechnology, and Biochemistry,* **64**(1),80-88.

Oliva, G., Fontes, M.R., Garratt, R.C., Altamirano, M.M., Calcagno, M.L., Horjales, E., (1995). Structure and Catalytic Mechanism of Glucosamine 6-Phosphate Deaminase from *Escherichia coli* at 2.1 å Resolution. *Structure,* **3**(12),1323-1332.

Oslancova, A. and Janecek, S., (2002). Oligo-1,6-Glucosidase and Neopullulanase Enzyme Subfamilies from the α-Amylase Family Defined by the Fifth Conserved Sequence Region. *Cellular and Molecular Life Sciences : CMLS,* **59**(11),1945-1959.

O'Sullivan, D.J. and Klaenhammer, T.R., (1993). High- and Low-Copy-Number Lactococcus Shuttle Cloning Vectors with Features for Clone Screening. *Gene,* **137**(2),227-231.

Ouyang, Y., Dai, S., Xie, L., Ravi Kumar, M.S., Sun, W., Sun, H., Tang, D., Li, X., (2010). Isolation of High Molecular Weight DNA from Marine Sponge Bacteria for BAC Library Construction. *Marine Biotechnology (New York, N.Y.),* **12**(3),318-325.

Oxley, A.P.A., Lanfranconi, M.P., Würdemann, D., Ott, S., Schreiber, S., McGenity, T.J., Timmis, K.N., Nogales, B., (2010). Halophilic Archaea in the Human Intestinal Mucosa. *Environmental Microbiology,* **12**(9),2398-2410.

Ozkose, E., Akyol, I., Kar, B., Comlekcioglu, U., Ekinci, M.S., (2009). Expression of Fungal Cellulase Gene in *Lactococcus Lactis* to Construct Novel Recombinant Silage Inoculants. *Folia Microbiologica,* **54**(4),335-342.

Pages, S., Belaich, A., Belaich, J.P., Morag, E., Lamed, R., Shoham, Y., Bayer, E.A., (1997). Species-Specificity of the Cohesin-Dockerin Interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of Specificity Determinants of the Dockerin Domain. *Proteins,* **29**(4),517-527.

Palackal, N., Lyon, C., Zaidi, S., Luginbühl, P., Dupree, P., Goubet, F., Macomber, J., Short, J., Hazlewood, G., Robertson, D., Steer, B., (2007). A Multifunctional Hybrid Glycosyl Hydrolase Discovered in an Uncultured Microbial Consortium from Ruminant Gut. *Applied Microbiology and Biotechnology,* **74**(1),113-124.

Papagianni, M., Avramidis, N., Filioussis, G., (2007). High Efficiency Electrotransformation of *Lactococcus lactis* spp. *lactis* Cells Pretreated with Lithium Acetate and Dithiothreitol. *BMC Biotechnology,* **7**(1),15.

Papamichail, D. and Delihas, N., (2006). Outer Membrane Protein Genes and their Small Non-Coding RNA Regulator Genes in *Photorhabdus luminescens*. *Biology Direct,* **1**(1),12.

Pereira, J.H., Sapra, R., Volponi, J.V., Kozina, C.L., Simmons, B., Adams, P.D., (2009). Structure of Endoglucanase Cel9A from the Thermoacidophilic *Alicyclobacillus acidocaldarius*. *Acta crystallographica. Section D, Biological crystallography,* **65**(Pt 8),744-750.

Pereira, J.H., Chen, Z., McAndrew, R.P., Sapra, R., Chhabra, S.R., Sale, K.L., Simmons, B.A., Adams, P.D., (2010). Biochemical Characterization and Crystal Structure of Endoglucanase Cel5A from the Hyperthermophilic *Thermotoga maritima*. *Journal of Structural Biology,* **172**(3),372-379.

Peris-Bondia, F., Latorre, A., Artacho, A., Moya, A., D'Auria, G., (2011). The Active Human Gut Microbiota Differs from the Total Microbiota. *PloS One,* **6**(7),e22448.

Perry, J.D., Morris, K.A., James, A.L., Oliver, M., Gould, F.K., (2007). Evaluation of Novel Chromogenic Substrates for the Detection of Bacterial β-Glucosidase. *Journal of Applied Microbiology,* **102**(2),410-415.

Pfeiler, E.A. and Klaenhammer, T.R., (2007). The Genomics of Lactic Acid Bacteria. *Trends in Microbiology,* **15**(12),546-553.

Pope, P.B., Denman, S.E., Jones, M., Tringe, S.G., Barry, K., Malfatti, S.A., McHardy, A.C., Cheng, J.F., Hugenholtz, P., McSweeney, C.S., Morrison, M., (2010). Adaptation to Herbivory by the Tammar Wallaby Includes Bacterial and Glycoside Hydrolase Profiles Different from Other Herbivores. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(33),14793-14798.

Posta, K., Béki, E., Wilson, D.B., Kukolya, J., Hornok, L., (2004). Cloning, Characterization and Phylogenetic Relationships of cel5B, a New Endoglucanase Encoding Gene from *Thermobifida fusca*. *Journal of Basic Microbiology,* **44**(5),383-399.

Price, L.B., Liu, C.M., Melendez, J.H., Frankel, Y.M., Engelthaler, D., Aziz, M., Bowers, J., Rattray, R., Ravel, J., Kingsley, C., Keim, P.S., Lazarus, G.S., Zenilman, J.M., (2009). Community Analysis of Chronic Wound Bacteria using 16S rRNA Gene-Based Pyrosequencing: Impact of Diabetes and Antibiotics on Chronic Wound Microbiota. *PloS One,* **4**(7),e6462.

Pryde, S.E., Duncan, S.H., Hold, G.L., Stewart, C.S., Flint, H.J., (2002). The Microbiology of Butyrate Formation in the Human Colon. *FEMS Microbiology Letters,* **217**(2),133-139.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium,

Bork, P., Ehrlich, S.D., Wang, J., (2010). A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature*, **464**(7285),59-65.

Rajilic-Stojanovic, M., Biagi, E., Heilig, H.G., Kajander, K., Kekkonen, R.A., Tims, S., de Vos, W.M., (2011). Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples from Patients with Irritable Bowel Syndrome. *Gastroenterology*, **5**.

Ramirez-Farias, C., Slezak, K., Fuller, Z., Duncan, A., Holtrop, G., Louis, P., (2009). Effect of Inulin on the Human Gut Microbiota: Stimulation of *Bifidobacterium adolescentis* and *Faecalibacterium prausnitzii*. *The British Journal of Nutrition*, **101**(4),541-550.

Ramsay, A.G., Scott, K.P., Martin, J.C., Rincon, M.T., Flint, H.J., (2006). Cell-Associated Alpha-Amylases of Butyrate-Producing Firmicute Bacteria from the Human Colon. *Microbiology (Reading, England)*, **152**(Pt 11),3281-3290.

Reid, G., Younes, J.A., Van der Mei, H.C., Gloor, G.B., Knight, R., Busscher, H.J., (2011). Microbiota Restoration: Natural and Supplemented Recovery of Human Microbial Communities. *Nature Reviews.Microbiology*, **9**(1),27-38.

Reigstad, L.J., Bartossek, R., Schleper, C., (2011). Preparation of High-Molecular Weight DNA and Metagenomic Libraries from Soils and Hot Springs. *Methods in Enzymology*, **496**319-344.

Rhee, J., Ahn, D., Kim, Y., Oh, J., (2005). New Thermophilic and Thermostable Esterase with Sequence Similarity to the Hormone-Sensitive Lipase Family, Cloned from a Metagenomic Library. *Applied and Environmental Microbiology*, **71**(2),817-825.

Robert, C. and Bernalier-Donadille, A., (2003). The Cellulolytic Microflora of the Human Colon: Evidence of Microcrystalline Cellulose-Degrading Bacteria in Methane-Excreting Subjects. *FEMS Microbiology Ecology*, **46**(1),81-89.

Robert, C., Chassard, C., Lawson, P.A., Bernalier-Donadille, A., (2007). *Bacteroides cellulosilyticus* sp. nov., a Cellulolytic Bacterium from the Human Gut Microbial Community. *International Journal of Systematic and Evolutionary Microbiology*, **57**(Pt 7),1516-1520.

Rosewarne, C.P., Pope, P.B., Denman, S.E., McSweeney, C.S., O'Cuiv, P., Morrison, M., (2011). High-Yield and Phylogenetically Robust Methods of DNA Recovery for Analysis of Microbial Biofilms Adherent to Plant Biomass in the Herbivore Gut. *Microbial Ecology*, **61**(2),448-454.

Ruiz, R. and Rubio, L.A., (2009). Lyophilisation Improves the Extraction of PCR-Quality Community DNA from Pig Faecal Samples. *Journal of the Science of Food and Agriculture*, **89**(4),723-727.

Russell, D.A., Ross, R.P., Fitzgerald, G.F., Stanton, C., (2011). Metabolic Activities and Probiotic Potential of Bifidobacteria. *International Journal of Food Microbiology,* **149**(1),88-105.

Ryan, S.M., Fitzgerald, G.F., van Sinderen, D., (2006). Screening for and Identification of Starch-, Amylopectin-, and Pullulan-Degrading Activities in Bifidobacterial Strains. *Applied and Environmental Microbiology,* **72**(8),5289-5296.

Rye, C.S. and Withers, S.G., (2000). Glycoside Mechanisms. *Current Opinion in Chemical Biology*, **4**(5),573-580

Sakka, K., Yoshikawa, K., Kojima, Y., Karita, S., Ohmiya, K., Shimada, K., (1993). Nucleotide Sequence of the *Clostridium stercorarium* xylA Gene Encoding a Bifunctional Protein with β-D-Xylosidase and α-L-Arabinofuranosidase Activities, and Properties of the Translated Product. *Bioscience, Biotechnology, and Biochemistry,* **57**(2),268-272.

Salyers, A.A., Vercellotti, J.R., West, S.E., Wilkins, T.D., (1977). Fermentation of Mucin and Plant Polysaccharides by Strains of Bacteroides from the Human Colon. *Applied and Environmental Microbiology,* **33**(2),319-322.

Sambrook, J. and Russell, D.W., 2001. *Molecular Cloning. A laboratory manual.* Third edn. New York: Cold Spring Harbor Laboratory Press.

Satokari, R.M., Vaughan, E.E., Akkermans, A.D., Saarela, M., de Vos, W.M., (2001). Bifidobacterial Diversity in Human Feces Detected by Genus-Specific PCR and Denaturing Gradient Gel Electrophoresis. *Applied and Environmental Microbiology,* **67**(2),504-513.

Scanlan, P.D., Shanahan, F., Marchesi, J.R., (2009). Culture-Independent Analysis of Desulfovibrios in the Human Distal Colon of Healthy, Colorectal Cancer and Polypectomized Individuals. *FEMS Microbiology Ecology,* **69**(2),213-221.

Schallmey, M., Ly, A., Wang, C., Meglei, G., Voget, S., Streit, W.R., Driscoll, B.T., Charles, T.C., (2011). Harvesting of Novel Polyhydroxyalkanaote (PHA) Synthase Encoding Genes from a Soil Metagenome Library using Phenotypic Screening. *FEMS Microbiology Letters,* **321**(2),150-156.

Scharlau, D., Borowicki, A., Habermann, N., Hofmann, T., Klenow, S., Miene, C., Munjal, U., Stein, K., Glei, M., (2009). Mechanisms of Primary Cancer Prevention by Butyrate and Other Products Formed during Gut Flora-Mediated Fermentation of Dietary Fibre. *Mutation Research - Reviews in Mutation Research,* **682**(1),39-53.

Schell, M.A., Karmirantzou, M., Snel, B., Vilanova, D., Berger, B., Pessi, G., Zwahlen, M., Desiere, F., Bork, P., Delley, M., Pridmore, R.D., Arigoni, F., (2002). The Genome Sequence of *Bifidobacterium longum* Reflects its Adaptation to the Human Gastrointestinal Tract. *Proceedings of the National Academy of Sciences,* **99**(22),14422-14427.

Scheppach, W., (1994). Effects of Short Chain Fatty Acids on Gut Morphology and Function. *Gut,* **35**(Supplement 1),S35-38.

Schloss, P.D. and Handelsman, J., (2005). Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology,* **71**(3),1501-1506.

Schumacher, M.A., (2007). Structural Biology of Plasmid Segregation Proteins. *Current Opinion in Structural Biology*, **17**(1),103-109.

Scott, K.P., Martin, J.C., Chassard, C., Clerget, M., Potrykus, J., Campbell, G., Mayer, C.D., Young, P., Rucklidge, G., Ramsay, A.G., Flint, H.J., (2011). Substrate-Driven Gene Expression in Roseburia Inulinivorans: Importance of Inducible Enzymes in the Utilization of Inulin and Starch. *Proceedings of the National Academy of Sciences of the United States of America,* **108** (Supplement 1),4672-4679.

Segain, J.P., Raingeard de la Bletiere, D., Bourreille, A., Leray, V., Gervois, N., Rosales, C., Ferrier, L., Bonnet, C., Blottiere, H.M., Galmiche, J.P., (2000). Butyrate Inhibits Inflammatory Responses through NFkappaB Inhibition: Implications for Crohn's Disease. *Gut,* **47**(3),397-403.

Selvaraj, T., Kim, S.K., Kim, Y.H., Jeong, Y.S., Kim, Y.J., Phuong, N.D., Jung, K.H., Kim, J., Yun, H.D., Kim, H., (2010). The Role of Carbohydrate-Binding Module (CBM) Repeat of a Multimodular Xylanase (XynX) from *Clostridium thermocellum* in Cellulose and Xylan Binding . *The Journal of Microbiology,* **48**(6),856.

Servin, A.L., (2004). Antagonistic Activities of Lactobacilli and Bifidobacteria Against Microbial Pathogens. *FEMS Microbiology Reviews,* **28**(4),405-440.

Seveno, M., Voxeur, A., Rihouey, C., Wu, A.M., Ishii, T., Chevalier, C., Ralet, M.C., Driouich, A., Marchant, A., Lerouge, P., (2009). Structural Characterisation of the Pectic Polysaccharide Rhamnogalacturonan II using an Acidic Fingerprinting Methodology. *Planta,* **230**(5),947-957.

Sharma, S., Khan, F.G., Qazi, G.N., (2010). Molecular Cloning and Characterization of Amylase from Soil Metagenomic Library Derived from Northwestern Himalayas. *Applied Microbiology and Biotechnology,* **86**(6),1821-1828.

Shen, H., Gilkes, N.R., Kilburn, D.G., Miller, R.C.,Jr, Warren, R.A., (1995). Cellobiohydrolase B, a Second Exo-Cellobiohydrolase from the Cellulolytic Bacterium *Cellulomonas fimi*. *The Biochemical journal,* **311**,67-74.

Sheng, Y.L., Mancino, V., Birren, B., (1995). Transformation of *Escherichia coli* with large DNA molecules by electrporation. *Nucleic Acids Research*, **23**(11), 1990-1996.

Shine, J. and Dalgarno, L., (1974). The 3'-Terminal Sequence of Escherichia Coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding

Sites. *Proceedings of the National Academy of Sciences of the United States of America,* **71**(4),1342-1346.

Simon, C. and Daniel, R., (2011). Metagenomic Analyses: Past and Future Trends. *Applied and Environmental Microbiology,* **77**(4),1153-1161.

Simpson, P.J., Xie, H., Bolam, D.N., Gilbert, H.J., Williamson, M.P., (2000). The Structural Basis for the Ligand Specificity of Family 2 Carbohydrate-Binding Modules. *The Journal of Biological Chemistry,* **275**(52),41137-41142.

Smith, E.A. and Macfarlane, G.T., (1997). Dissimilatory Amino Acid Metabolism in Human Colonic Bacteria. *Anaerobe,* **3**(5),327-337.

Sobhani, I., Tap, J., Roudot-Thoraval, F., Roperch, J.P., Letulle, S., Langella, P., Corthier, G., Tran Van Nhieu, J., Furet, J.P., (2011). Microbial Dysbiosis in Colorectal Cancer (CRC) Patients. *PloS One,* **6**(1),e16393.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L.G., Gratadoux, J.J., Blugeon, S., Bridonneau, C., Furet, J.P., Corthier, G., Grangette, C., Vasquez, N., Pochart, P., Trugnan, G., Thomas, G., Blottiere, H.M., Dore, J., Marteau, P., Seksik, P., Langella, P., (2008). *Faecalibacterium prausnitzii* is an Anti-Inflammatory Commensal Bacterium Identified by Gut Microbiota Analysis of Crohn Disease Patients. *Proceedings of the National Academy of Sciences of the United States of America,* **105**(43),16731-16736.

Songsiriritthigul, C., Buranabanyat, B., Haltrich, D., Yamabhai, M., (2010). Efficient Recombinant Expression and Secretion of a Thermostable GH26 Mannan Endo-1,4-β-Mannosidase from *Bacillus licheniformis* in *Escherichia coli*. *Microbial Cell Factories,* **9,**20.

Sorensen, K.I. and Hove-Jensen, B., (1996). Ribose Catabolism of *Escherichia coli*: Characterization of the rpiB Gene Encoding Ribose Phosphate Isomerase B and of the rpiR Gene, which is Involved in Regulation of rpiB Expression. *Journal of Bacteriology,* **178**(4),1003-1011.

Spor, A., Koren, O., Ley, R., (2011). Unravelling the Effects of the Environment and Host Genotype on the Gut Microbiome. *Nature Reviews.Microbiology,* **9**(4),279-290.

Sriraman, K. and Jayaraman, G., (2008). HtrA is Essential for Efficient Secretion of Recombinant Proteins by *Lactococcus lactis*. *Applied and Environmental Microbiology,* **74**(23),7442-7446.

Stephenson, K. and Harwood, C.R., (1998). Influence of a Cell-Wall-Associated Protease on Production of α-Amylase by *Bacillus subtilis*. *Applied and Environmental Microbiology,* **64**(8),2875-2881.

Strickler, M.A., Hall, J.A., Gaiko, O., Pajor, A.M., (2009). Functional Characterization of a Na+-Coupled Dicarboxylate Transporter from *Bacillus licheniformis*. *Biochimica et Biophysica Acta (BBA) - Biomembranes,* **1788**(12),2489-2496.

Surade, S., Klein, M., Stolt-Bergner, P.C., Muenke, C., Roy, A., Michel, H., (2006). Comparative Analysis and "Expression Space" Coverage of the Production of Prokaryotic Membrane Proteins for Structural Genomics. *Protein Science : a publication of the Protein Society,* **15**(9),2178-2189.

Suryani, Kimura, T., Sakka, K., Ohmiya, K., (2004). Sequencing and Expression of the Gene Encoding the *Clostridium stercorarium* β-Xylosidase Xyl43B in *Escherichia coli*. *Bioscience, Biotechnology, and Biochemistry,* **68**(3),609-614.

Tamaru, Y., Miyake, H., Kuroda, K., Nakanishi, A., Kawade, Y., Yamamoto, K., Uemura, M., Fujita, Y., Doi, R.H., Ueda, M., (2010). Genome Sequence of the Cellulosome-Producing Mesophilic Organism *Clostridium cellulovorans* 743B. *Journal of Bacteriology,* **192**(3),901-902.

Tanous, C., Chambellon, E., Yvon, M., (2007). Sequence Analysis of the Mobilizable Lactococcal Plasmid pGdh442 Encoding Glutamate Dehydrogenase Activity. *Microbiology*, **153**(5),1664-1675.

Tap, J., Mondot, S., Levenez, F., Pelletier, E., Caron, C., Furet, J.P., Ugarte, E., Munoz-Tamayo, R., Paslier, D.L., Nalin, R., Dore, J., Leclerc, M., (2009). Towards the Human Intestinal Microbiota Phylogenetic Core. *Environmental Microbiology,* **11**(10),2574-2584.

Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., Cantarel, B.L., Coutinho, P.M., Henrissat, B., Leclerc, M., Dore, J., Monsan, P., Remaud-Simeon, M., Potocki-Veronese, G., (2010). Functional Metagenomics to Mine the Human Gut Microbiome for Dietary Fiber Catabolic Enzymes. *Genome Research,* **20**(11),1605-1612.

Taupp, M., Mewis, K., Hallam, S.J., (2011). The Art and Design of Functional Metagenomic Screens. *Current Opinion in Biotechnology,* **22**(3),465-472.

Taylor, E.J., Smith, N.L., Turkenburg, J.P., D'Souza, S., Gilbert, H.J., Davies, G.J., (2006). Structural Insight into the Ligand Specificity of a Thermostable Family 51 Arabinofuranosidase, Araf51, from *Clostridium thermocellum*. *The Biochemical Journal,* **395**(1),31-37.

Teather, R.M. and Wood, P.J., (1982). Use of Congo Red-Polysaccharide Interactions in Enumeration and Characterization of Cellulolytic Bacteria from the Bovine Rumen. *Applied and Environmental Microbiology.,* **43**(4),777-780.

Tull, D., Withers, S.G., Gilkes, N.R., Kilburn, D.G., Warren, R.A., Aebersold, R., (1991). Glutamic Acid 274 is the Nucleophile in the Active Site of a "Retaining" Exoglucanase from *Cellulomonas fimi*. *Journal of Biological Chemistry,* **266**(24),15621-15625.

Turnbaugh, P.J. and Gordon, J.I., (2009). The Core Gut Microbiome, Energy Balance and Obesity. *Journal of Physiology*, **587**(17),4153-4158.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., Egholm, M., Henrissat, B., Heath, A.C., Knight, R., Gordon, J.I., (2009). A Core Gut Microbiome in Obese and Lean Twins. *Nature,* **457**(7228),480-484.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F., (2004). Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. *Nature,* **428**(6978),37-43.

Uchiyama, T. and Watanabe, K., (2008). Substrate-Induced Gene Expression (SIGEX) Screening of Metagenome Libraries. *Nature Protocols,* **3**(7),1202-1212.

Vaillancourt, K., Moineau, S., Frenette, M., Lessard, C., Vadeboncoeur, C., (2002). Galactose and Lactose Genes from the Galactose-Positive Bacterium *Streptococcus salivarius* and the Phylogenetically Related Galactose-Negative Bacterium *Streptococcus thermophilus*: Organization, Sequence, Transcription, and Activity of the Gal Gene Products. *Journal of Bacteriology,* **184**(3),785-793.

Vaishampayan, P.A., Kuehl, J.V., Froula, J.L., Morgan, J.L., Ochman, H., Francino, M.P., (2010). Comparative Metagenomics and Population Dynamics of the Gut Microbiota in Mother and Infant. *Genome Biology and Evolution,* **2,**53-66.

Van De Guchte, M., Kok, J., Venema, G., (1992). Gene Expression in *Lactococcus lactis*. *FEMS Microbiology Reviews,* **88**(2),73-92.

Van den Abbeele, P., Gerard, P., Rabot, S., Bruneau, A., El Aidy, S., Derrien, M., Kleerebezem, M., Zoetendal, E.G., Smidt, H., Verstraete, W., Van de Wiele, T., Possemiers, S., (2011). Arabinoxylans and Inulin Differentially Modulate the Mucosal and Luminal Gut Microbiota and Mucin-Degradation in Humanized Rats. *Environmental Microbiology,* **13**(10),2667-2680.

Van Den Broek, L.A.M. and Voragen, A.G.J., (2008). Bifidobacterium Glycoside Hydrolases and (Potential) Prebiotics. *Innovative Food Science and Emerging Technologies,* **9**(4),401-407.

van Wely, K.H., Swaving, J., Freudl, R., Driessen, A.J., (2001). Translocation of Proteins Across the Cell Envelope of Gram-Positive Bacteria. *FEMS Microbiology Reviews,* **25**(4),437-454.

Varghese, J.N., Hrmova, M., Fincher, G.B., (1999). Three-Dimensional Structure of a Barley β-D-Glucan Exohydrolase, a Family 3 Glycosyl Hydrolase. *Structure,* **7**(2),179-190.

Várnai, A., Huikko, L., Pere, J., Siika-aho, M., Viikari, L., (2011). Synergistic Action of Xylanase and Mannanase Improves the Total Hydrolysis of Softwood. *Bioresource Technology,* **102**(19),9096-9104.

Vaughan, E.E., de Vries, M., Zoetendal, E.G., Ben-Amor, K., Akkermans, A.D., de Vos, W., (2002). The Intestinal LABs. *Antonie van Leeuwenhoek,* **82**(1-4),341-352.

Vazana, Y., Morais, S., Barak, Y., Lamed, R., Bayer, E.A., (2010). Interplay between *Clostridium thermocellum* Family 48 and Family 9 Cellulases in Cellulosomal Versus Noncellulosomal States. *Applied and Environmental Microbiology,* **76**(10),3236-3243.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y., Smith, H.O., (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science,* **304**(5667),66-74.

Ventura, M., O'Flaherty, S., Claesson, M.J., Turroni, F., Klaenhammer, T.R., van Sinderen, D., O'Toole, P.W., (2009). Genome-Scale Analyses of Health-Promoting Bacteria: Probiogenomics. *Nature Reviews. Microbiology,* **7**(1),61-71.

Voskuil, M.I. and Chambliss, G.H., (1998). The -16 Region of *Bacillus subtilis* and Other Gram-Positive Bacterial Promoters. *Nucleic Acids Research,* **26**(15),3584-3590.

Walker, A.W., Duncan, S.H., Harmsen, H.J.M., Holtrop, G., Welling, G.W., Flint, H.J., (2008). The Species Composition of the Human Intestinal Microbiota Differs between Particle-Associated and Liquid Phase Communities. *Environmental Microbiology,* **10**(12),3275-3283.

Walker, A.W., Ince, J., Duncan, S.H., Webster, L.M., Holtrop, G., Ze, X., Brown, D., Stares, M.D., Scott, P., Bergerat, A., Louis, P., McIntosh, F., Johnstone, A.M., Lobley, G.E., Parkhill, J., Flint, H.J., (2011). Dominant and Diet-Responsive Groups of Bacteria within the Human Colonic Microbiota. *The ISME Journal,* **5**(2),220-230.

Walker, A.W., Duncan, S.H., McWilliam Leitch, E.C., Child, M.W., Flint, H.J., (2005). PH and Peptide Supply can Radically Alter Bacterial Populations and Short-Chain Fatty Acid Ratios within Microbial Communities from the Human Colon. *Appl. Environ. Microbiol.,* **71**(7),3692-3700.

Walter, J. and Ley, R.E., (2010). The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. *Annual Review of Microbiology,* **65,**411-429.

Walter, J., Mangold, M., Tannock, G.W., (2005). Construction, Analysis, and β-Glucanase Screening of a Bacterial Artificial Chromosome Library from the Large-Bowel Microbiota of Mice. *Applied and Environmental Microbiology,* **71**(5),2347-2354.

Walter, S. and Schrempf, H., (2008). Characteristics of the Surface-Located Carbohydrate-Binding Protein CbpC from *Streptomyces coelicolor* A32. *Archives of Microbiology,* **190**(2),119-127.

Wang, G.Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., Meurer, G., Saxena, G., Andersen, R.J., Davies, J., (2000). Novel Natural Products from Soil DNA Libraries in a Streptomycete Host. *Organic Letters,* **2**(16),2401-2404.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.M., Wu, Y., Schauer, P., Smith, J.D., Allayee, H., Tang, W.H., DiDonato, J.A., Lusis, A.J., Hazen, S.L., (2011). Gut Flora Metabolism of Phosphatidylcholine Promotes Cardiovascular Disease. *Nature,* **472**(7341),57-63.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S.G., Podar, M., Martin, H.G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N.C., Matson, E.G., Ottesen, E.A., Zhang, X., Hernandez, M., Murillo, C., Acosta, L.G., Rigoutsos, I., Tamayo, G., Green, B.D., Chang, C., Rubin, E.M., Mathur, E.J., Robertson, D.E., Hugenholtz, P., Leadbetter, J.R., (2007). Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood-Feeding Higher Termite. *Nature,* **450**(7169),560-565.

Warren, R.L., Freeman, J.D., Levesque, R.C., Smailus, D.E., Flibotte, S., Holt, R.A., (2008). Transcription of Foreign DNA in *Escherichia coli*. *Genome Research,* **18**(11),1798-1805.

Wegmann, U., Klein, J.R., Drumm, I., Kuipers, O.P., Henrich, B., (1999). Introduction of Peptidase Genes from *Lactobacillus delbrueckii* subsp. *lactis* into *Lactococcus lactis* and Controlled Expression. *Applied Environmental Microbiology,* **65**(11),4729-4733.

Wegmann, U., O'Connell-Motherway, M., Zomer, A., Buist, G., Shearman, C., Canchaya, C., Ventura, M., Goesmann, A., Gasson, M.J., Kuipers, O.P., (2007). Complete Genome Sequence of the Prototype Lactic Acid Bacterium *Lactococcus lactis* subsp. *cremoris* MG1363. *Journal of Bacteriology,* **189**(8),3256-3270.

Wells, J.M., Wilson, P.W., Le Page, R.W., (1993). Improved Cloning Vectors and Transformation Procedure for *Lactococcus lactis*. *The Journal of Applied Bacteriology,* **74**(6),629-636.

Whitehead, T.R., (1995). Nucleotide Sequences of Xylan-Inducible Xylanase and xylosidase/arabinosidase Genes from *Bacteroides ovatus* V975. *Biochimica et Biophysica Acta (BBA) - General Subjects,* **1244**(1),239-241.

Wommack, K.E., Bhavsar, J., Ravel, J., (2008). Metagenomics: Read Length Matters. *Applied Environmental Microbiology,* **74**(5),1453-1463.

Wong, D., (2008). Enzymatic Deconstruction of Backbone Structures of the Ramified Regions in Pectins. *The Protein Journal,* **27**(1),30-42.

Wong, J.M.W. and Jenkins, D.J.A., (2007). Carbohydrate Digestibility and Metabolic Effects. *Journal of Nutrition,* **137**(11),2539-2546.

Xie, G., Bruce, D.C., Challacombe, J.F., Chertkov, O., Detter, J.C., Gilna, P., Han, C.S., Lucas, S., Misra, M., Myers, G.L., Richardson, P., Tapia, R., Thayer, N., Thompson, L.S., Brettin, T.S., Henrissat, B., Wilson, D.B., McBride, M.J., (2007).

Genome Sequence of the Cellulolytic Gliding Bacterium *Cytophaga hutchinsonii*. *Applied and Environmental Microbiology,* **73**(11),3536-3546.

Xu, J., Mahowald, M.A., Ley, R.E., Lozupone, C.A., Hamady, M., Martens, E.C., Henrissat, B., Coutinho, P.M., Minx, P., Latreille, P., Cordum, H., Van Brunt, A., Kim, K., Fulton, R.S., Fulton, L.A., Clifton, S.W., Wilson, R.K., Knight, R.D., Gordon, J.I., (2007). Evolution of Symbiotic Bacteria in the Distal Human Intestine. *PLoS Biology,* **5**(7),e156.

Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V., Gordon, J.I., (2003). A Genomic View of the Human-*Bacteroides thetaiotaomicron* Symbiosis. *Science,* **299**(5615),2074-2076.

Xu, Q., Morrison, M., Nelson, K.E., Bayer, E.A., Atamna, N., Lamed, R., (2004). A Novel Family of Carbohydrate-Binding Modules Identified with *Ruminococcus albus* Proteins. *FEBS Letters,* **566**(1-3),11-16.

Yamabhai, M., Emrat, S., Sukasem, S., Pesatcha, P., Jaruseranee, N., Buranabanyat, B., (2008). Secretion of Recombinant Bacillus Hydrolytic Enzymes using *Escherichia coli* Expression Systems. *Journal of Biotechnology,* **133**(1),50-57.

Yoda, K., Toyoda, A., Mukoyama, Y., Nakamura, Y., Minato, H., (2005). Cloning, Sequencing, and Expression of a Eubacterium Cellulosolvens 5 Gene Encoding an Endoglucanase (Cel5A) with Novel Carbohydrate-Binding Modules, and Properties of Cel5A. *Applied and Environmental Microbiology,* **71**(10),5787-5793.

Yun, J., Kang, S., Park, S., Yoon, H., Kim, M.J., Heu, S., Ryu, S., (2004). Characterization of a Novel Amylolytic Enzyme Encoded by a Gene from a Soil-Derived Metagenomic Library. *Applied and Environmental Microbiology,* **70**(12),7229-7235.

Zoetendal, E.G., Rajilic-Stojanovic, M., de Vos, W.M., (2008). High-Throughput Diversity and Functionality Analysis of the Gastrointestinal Tract Microbiota. *Gut,* **57**(11),1605-1615.

Zweers, J.C., Barák, I., Becher, D., Driessen, A.J., Hecker, M., Kontinen, V.P., Saller, M.J., Vavrová, L., van Dijl, J.M., (2008). Towards the development of *Bacillus subtilis* as a cell factory for membrane proteins and protein complexes. *Microbial Cell Factories* **7,**10.