# University of Salford
## MANCHESTER

# Analysis of the gut microbiome of the common black slug *Arion ater*: In search of novel lignocellulose degrading enzymes

Ryan Edward Joynson

School of Environment & Life Sciences

University of Salford, UK

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy, 2015

# Table of Contents

# List of figures

# List of tables

# Acknowledgements

Firstly I would like to thank my supervisor Dr Natalie Ferry for her great supervision and support that has made this project possible and a special thanks to my co-supervisor Dr Rhod Elder for all of his help and advice.

Thank to the technicians at Salford University who are always willing to bend over backwards to help us especially, Mark Parlby, Sally Shepherd, Tony Bodell and Manisha Patel.

Thanks also to Dr Leighton Pritchard for his advice on microbial ecology and for providing the *Dickeya sp*. Reference genomes and thanks to Dominic Wood for helping me take my first steps into the field of bioinformatics

A special mention for Patrick Killoran with whom I've shared countless scientific discussions, beers and hours watching critically acclaimed classics such as Predator and Die Hard. Thanks also to Tom Roedl, Soran Baban and Andrew McGown for their support and friendship.

I also worked with a number of great students and I'd like to thank them for all of their hard work, including Lucie De Longprez, Ambre Chapuis, Ekenakema Osemwekha and Paz Aranega Bou.

I would also like to thank Martin Mckee for his belief in me and invaluable advice over the years.

Thanks to my parents whos help and constanst selfless support have got me to where I am today.

Finally, and most of all I thank my partner Becky McKee, for all of the countless times she has motivated me and supported me in every way possible.

## Abbreviations

| | |
|---|---|
| AA | Auxiliary activity |
| AFEX | Ammonia fibre explosion |
| AG | Apiogalacturonan |
| APS | Ammonium persulphate |
| BAM | Binary AlignMent (file format) |
| BLAST | Basic Local Alignment Search Tool |
| BWA | Burrow-Wheeler Aligner |
| CAZyme | Carbohydrate active enzyme |
| CBM | Carbohydrate binding module |
| CBP | Consolidated bioprocessing |
| CGR | Centre for genomic research |
| CMC | Carboxymethylcellulose |
| DGGE | Denaturing gradient gel electrophoresis |
| EDTA | Ethylenediaminetetraacetic acid |
| FAME | Fatty acid methyl ester |
| GH | Glycoside hydrolase |
| HG | Homogalacturonan |
| HMF | 5-hyroxymethyl-2-furaldehyde |
| HRP | Horseradish peroxidase |
| KEGG | Kyoto encyclopaedia of genes and genomes |
| LLC | Lignin carbohydrate complex |
| MDA | Multiple displacement amplification |
| MG-RAST | Metagenomic rapid annotations using subsystem technologies |

| | |
|---|---|
| NCBI | National centre for biotechnology information |
| NGS | Next generation sequencing |
| NR database | Non-redundant protein database |
| PAGE | Polyacrylamide gel electrophoresis |
| Pfam | Protein family |
| PTS | Phosphotransferase system |
| R1 | Forward sequencing read of paired-end read data |
| R2 | Reverse sequencing read of paired-end read data |
| RGI | Rhamnogalacturonan |
| RG-II | Rhamnogalacturonan II |
| SAM | Simple AlignMent (file format) |
| SDS | Sodium dodecyl sulphate |
| SHF | Separate hydrolysis and fermentation |
| SOC | Super optimal broth with catabolite repression |
| SSF | Simultaneous saccharification and fermentation |
| TAE | Tris-Acetic acid-EDTA buffer |
| TBST | Tris buffered saline-tween 20 |
| UF solution | Urea-formamide solution |
| WGA | Whole genome amplification/amplified |
| XGA | Xylogalacturonan |

## Abstract

Some eukaryotes are able to gain access to otherwise well-protected carbon sources in plant biomass by exploiting microorganisms in the environment, or harboured in their digestive system. One such organism is the European black slug, *Arion ater,* which takes advantage of a gut microbial consortium that can break down plant tissues, including the widely available, but difficult to digest, carbohydrate polymers in lignocellulose. This ability is considered to be one of the major factors that have enabled *A. ater* to become one of the most widespread plant pest species in Western Europe and North America. Here we have identified the *A. ater* gut environment as a target for metagenomic study through identification of cellulolytic activity of bacterial origin. Next generation sequencing technology was used to characterize the bacterial diversity and functional capability of the gut microbiome of this notorious agricultural pest. Over 6 Gbp of gut metagenomic community sequences were analysed to reveal populations of known lignocellulose-degrading bacteria, along with abundant well-characterized bacterial plant pathogens. This study also revealed a repertoire of more than 3,000 carbohydrate active enzymes (CAZymes), indicating a microbial consortium capable of degradation of all components of lignocellulose, including cellulose, hemicellulose, pectin and lignin. Together, these functions would allow *A. ater* to make extensive use of plant biomass as a source of nutrients. This thesis demonstrates the importance of studying microbial communities in understudied groups such as the gastropods, firstly with respect to understanding links between feeding and evolutionary success and, secondly, as sources of novel enzymes with biotechnological potential, such as CAZYmes that could be used in the production of biofuel.

# Chapter 1. Literature Review: Lignocellulose and Its Role In Biofuel Production

## 1.1 Introduction

This thesis addresses one part of perhaps the greatest contemporary challenges facing our planet. How can a society that has prospered and grown for over two centuries, powered by a finite and diminishing stock of fossil fuels, secure a supply of energy long into the future? The answer is likely to lie in a transition to more sustainable energy sources, including solar and wind power and, more controversially, nuclear power. However, these sources may also include biofuels, derived from carbon trapped by photosynthetic activity in plants that are alive today and are therefore still trapping carbon. Yet none of these solutions is without challenges and, as this chapter will show, some of the initial promise of the early biofuels seems unlikely to be realised with current technology.

One of the major issues with biofuel production is the cost associated with use of inadequate carbohydrate active enzymes (CAZymes) during the degradation of lignocellulose, which leads to either large amounts of enzymes needed or lengthy incubation periods being required or both. CAZymes are a collection of proteins that contain domains/modules that bind to, create, break and modify glycosidic bonds, a large number of which are found in the carbohydrate polymers found in plant cell walls (Lombard *et al.*, 2014). It is widely thought that great savings could be made in this area if more active and stable CAZymes, believed to be present in nature, were discovered. To that end we have carried out a series of studies in search of novel CAZymes in nature, which are described in this thesis.

## 1.2 First Generation Biofuels

The development of biofuels began with the use of high energy crops that could be easily converted to a conventional liquid or gas fuel source. These are now called first generation biofuels. The three main types of first generation biofuels are bioethanol, fermented from sugar cane or corn sugars, fatty acid methyl ester (FAME or biodiesel) from transestrification of fatty acids from seed/vegetable oils, and biogas which is made by anaerobic fermentation of manure (Naik *et al.*, 2010). All three of these fuels have been developed to the extent that they can be considered as 'established technology', with bioethanol and biodiesel being produced on a scale of billions of litres a year globally (Naik *et al.*, 2010). Bioethanol derived from sugar cane and corn dominates biofuel production in North America and Brazil while Biodiesel dominates the market in Western Europe (Havlík *et al.*, 2010), mainly due to the climate requirements for optimal growth of feed stocks. Large scale production of bioethanol from high energy crops boomed between 2000 and 2004, when production doubled, and the upward trend has continued, with an increase from 66.77 billion litres in 2008 to 88.69 billion litres in 2013 (Gupta and Verma, 2015). Similarly the worldwide production of biodiesel has been steadily increasing year on year from 11 million litres in 1991 to over 1.77 billion litres by 2003 (Srinivasan, 2009) and to 30 billion litres by 2014 (Sarma *et al.*, 2014). These increases reflect the relatively low cost and ease of production when using high energy food crops as a substrate for biofuels. The simplicity of using sugar cane as a feedstock has allowed Brazil to fulfil >40% of its liquid fuel demands with bioethanol (Basso *et al.*, 2008). With bioethanol and biodiesel being seemingly viable replacements for petrol and diesel respectively, the future of biofuels looked bright until heightened demands for first generation biofuel

sparked the food versus fuel debate which challenges both the feasibility and morality of using food crops for energy production in a world that must feed an ever increasing population. One of the most controversial examples is in the USA where, in 2007, 92.9 million acres of corn crops were planted but one third of the corn produced was used to make bioethanol. Recent food price spikes, in 2007/2008, have been attributed to the increased use of food crops for biofuel, with food importing countries in the developing world being worst affected (Mohr and Raman, 2013). It has recently been estimated that an extra 200 million tonnes per year of corn will be required to fulfil demand for biofuels by 2017, putting great strains on production capacity for both fuel and food (Edgerton, 2009). This pressure has been linked to the spike in the price of corn, by 73%, in 2010 (Graham-Rowe, 2011) with some predicting that this could lead to a sustained price rise over the next decade (Mohr and Raman, 2013), leading to some staple foods becoming unaffordable in poorer countries. This has created many moral and ethical barriers that have yet to be resolved. Proposed solutions include either increasing productivity of existing farmland or expanding the areas in which food crops are grown. In theory, expanding the area that is farmed should be the easiest option but much of the world's cultivable land is already in use and what remains is unsuitable for almost all types of cropping (e.g. permafrost, inadequate water supply) or is pristine environment (e.g. rainforest) which, if used, would raise further ethical issues, for example through an adverse impact on biodiversity. The more desirable outcome would be to improve productivity of the land currently being farmed as it would avoid mass disruption of ecosystems and not require more space. Research to improve yields of corn has been undertaken for over 100 years and, since the early 20th century, yields per harvest have increased from 1.6

tonnes to 9.6 tonnes (Edgerton, 2009), attributed to modern agricultural methods such as hybrid corn, synthetic fertilizers and more recently, utilisation of modern biotechnology to enhance breeding methodology and the use of DNA-based markers to improve productivity (Edgerton, 2009). But even with these increases, production is still projected to fall short of demand.

There are also other bones of contention regarding first generation biofuels. These include the requirement for government subsidies if they are to be sold at prices that can compete with petroleum based products, a problem accentuated by recent falls in global oil prices, and also that some estimates of the production of greenhouse gases from biofuels suggest that they exceed those associated with fossil fuel production (Reijnders and Huijbregts, 2007). With all of these constraints, stemming from population growth, land shortage, and the need for subsidies, it is clear that the use of food crops as a sole means of producing biofuel is not a viable solution.

## 1.3  Second Generation Biofuels

To circumvent some of the political, ethical and practical issues that emerged following the initial production boom of first generations biofuels, new methods were sought out which do not utilise food crops as the feedstock for fuel production. These methods focussed on using non-edible biomass, including agricultural waste and dedicated non-food energy crops. To qualify as a second generation biofuel, the source of energy must not be suitable for human consumption and, in the case of dedicated energy crops, must grow on marginal land that could not support the cultivation of arable crops. The vast majority of second generation methods utilize plant biomass as a feedstock, using sugars locked in a number of carbohydrate polymers in plant cell walls as a source of

carbon. The plant cell wall superstructure, called lignocellulose after the predominant components, cellulose and the recalcitrant non-carbohydrate polymer lignin, contains multiple sugar based polymers that can be hydrolysed and fermented into bioethanol (Sun and Cheng, 2002). Lignocellulose is indigestible by humans and many other animals, so its main sources are cheap relative to crude oil and food crops. However the high cost of producing fuel from this feedstock has prevented lignocellulosic bioethanol from competing with first generation biofuels until very recently (Figure 1-1).
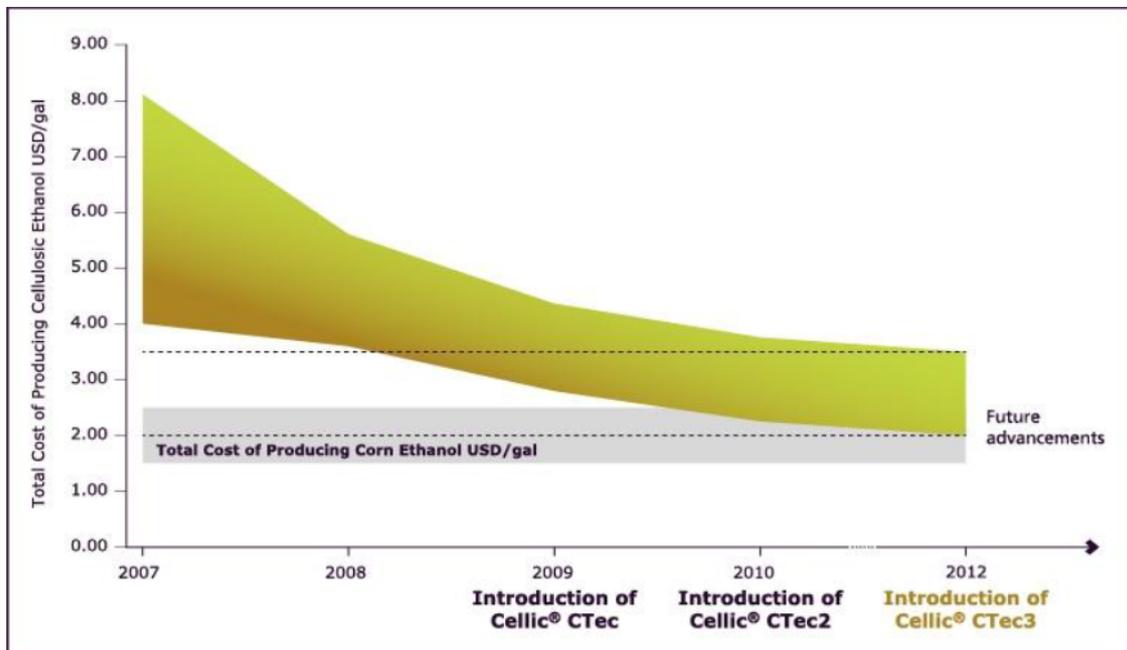


**Figure 1-1 The cost of bioethanol production using lignocellulose as a feedstock between 2007 and 2012 compared with the cost of using first generation feedstocks**

Source: http://www.bioenergy.novozymes.com/en/learn-more [accessed 20/01/15]

This situation is, however, changing, with the introduction of enzymatic methods to break down lignin. As Figure 1-1 shows, the cost of production of lignocellulosic bioethanol using Novozyme Cellic CTec enzyme cocktails is falling rapidly and

when used with some feedstocks is now comparable to that of producing ethanol from corn. Secondary biofuels have the great advantage of using cellulose that, along with accompanying plant cell wall components, is amongst the most abundant sources of organic material on the planet (Orfao *et al.*, 1999) and one that is truly renewable through photosynthetically fixing carbon from carbon dioxide (Figure 1-2).
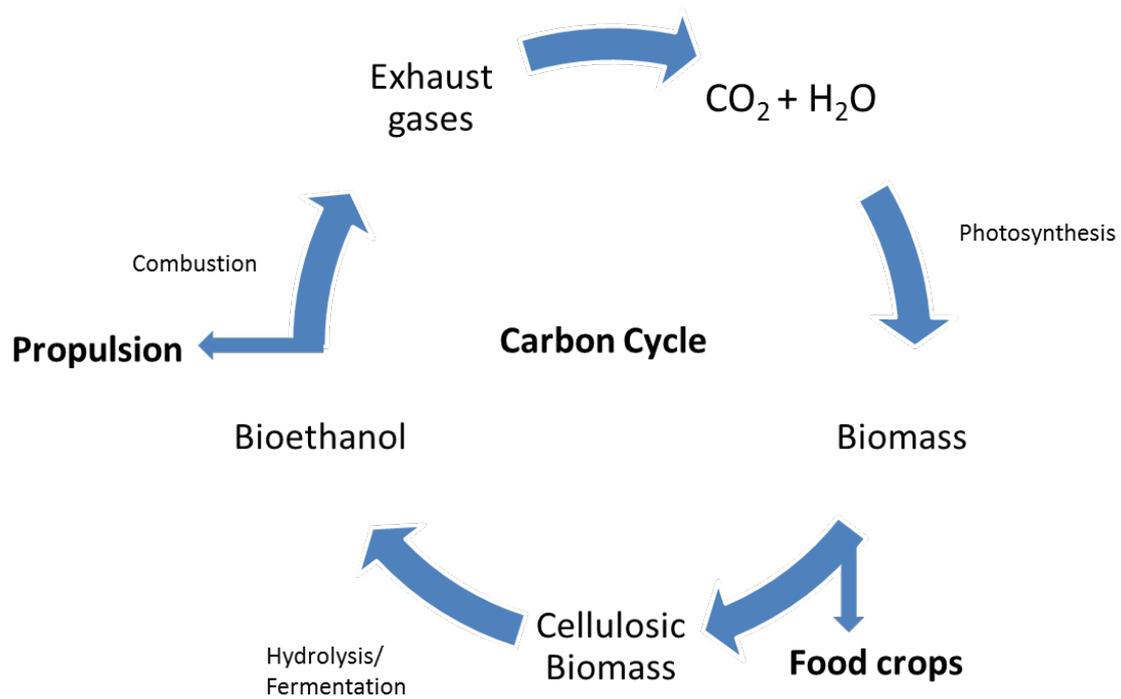


**Figure 1-2** **The carbon cycle, showing the possible journey of carbon from fixation to biofuel production to release as CO$_2$**

This gives lignocellulosic ethanol the potential to contribute substantially to the replacement of liquid fossil fuels globally.

## 1.4 The Structure and Components of Lignocellulose

Photosynthetically fixed carbon in the superstructures of plant cell walls is the most abundant source of organic carbon in the world (Sticklen, 2008). This carbon is found in polysaccharide chains that make up various support structures designed to provide both strength and protection to plant cells. The three main polysaccharide constituents of the plant cells wall are cellulose, hemicelluloses and pectin and plant cell walls are also impregnated to various degrees with the heterogeneous aromatic polymer lignin. The structures of these components, along with the interactions between them, are responsible for the strength and flexibility of the plant cell wall (Harholt *et al.*, 2010).

### 1.4.1 Cellulose

Native Cellulose is a highly crystalline polysaccharide composed of $\beta$-[1,4] linked D-glucose subunits and, unlike pectin and hemicellulose, shows no branching. It was first described in 1838 by Anselme Payen as a 'resistant fibrous solid' when it remained intact after acids and ammonia were used to treat plant tissue (Brown and Saxena, 2007). Payen then went on to determine the molecular formula of cellulose, $C_6H_{10}O_5$, using elemental analysis and observed the isomerism that it shares with starch. More recent studies shows that native cellulose has a microcrystalline structure comprised of individual chains with a twofold screw symmetry forming in parallel (Baker *et al.*, 2000) in one of two phases, I$\alpha$ or I$\beta$. These phases were discovered using analysis of spectral line splitting in solid-state $^{13}C$ magic angle spinning Nuclear Magnetic Resonance (NMR) spectroscopy, which provides dramatically increased resolution in solid state NMR by spinning the sample at the 'magic angle' $\theta_m$ relative to the magnetic field (VanderHart and Atalla, 1984). The *I$\alpha$ or I*$\beta$ phases differ in the orientation of the

hydroxyl side chains of the glucose monomers within the chains. The importance of these orientations arises from the formation of inter chain hydrogen bonds (Figure *1-3*), the number of which and their combined individual strength are responsible for the tensile strength and flexibility of cellulose and many of the ensuing properties of the plant cell wall.
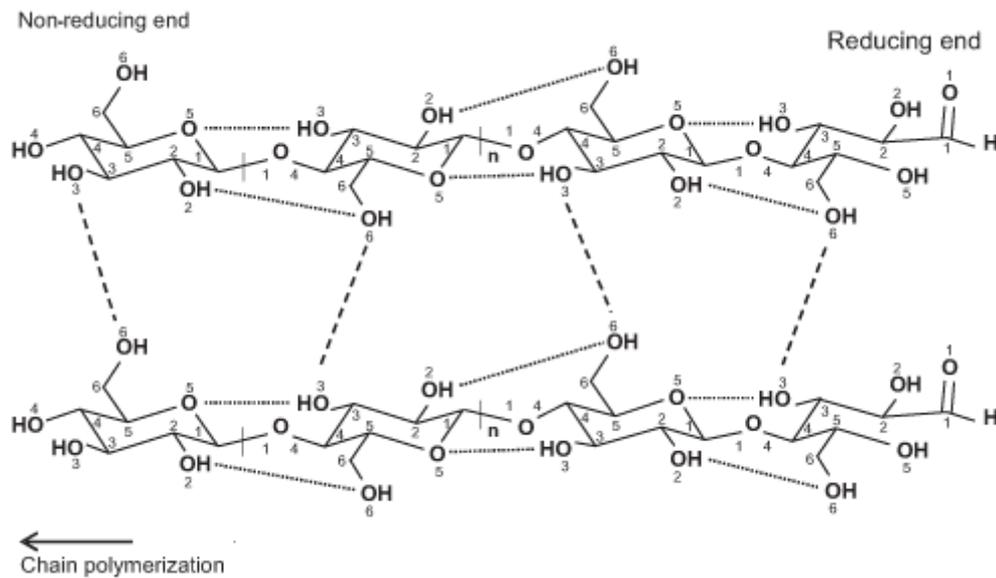


**Figure 1-3 The chemical structure of two cellulose oligomers showing inter and inter strand hydrogen bond interactions**

**Source***: (Festucci-Buselli *et al.*, 2007)

Figure 1-3 demonstrates the structure of the cellulose polysaccharide and the inter strand hydrogen bonds it can form between chains. The amount of the two native phases of cellulose can vary between species and even within species, with levels of stress during cell wall biosynthesis being implicated as a cause of the differences (Horii *et al.*, 1997, Sugiyama *et al.*, 1990). Although the more intricate details of cellulose structure have been determined quite recently, cellulose has been used for thousands of years in the form of cotton and paper and its potential is now being rediscovered as a biofuel feedstock to fulfil ever increasing energy

requirements. These four plant cell wall components described combine to form a complex superstructure with many different parts interacting with each other to hold the structure together.

## 1.4.2 Pectin

Pectin is the most structurally complicated of the three polysaccharide components and makes up between 2-10% (weight for weight [w/w]) of cell walls in grasses and up to 5% in woody tissue (Ridley *et al.*, 2001, Mohnen, 2008). Pectin comprises a complex mixture of polysaccharides, including: homogalacturonan (HG), xylogalacturonan (XGA), apiogalacturonan, rhamnogalacturonan I (RGI), and rhamnogalacturonan II (RG-II). Of these HG is usually the most abundant, making up as much as 65% of the pectin polymer, followed by RG1, contributing between 25 and 35% of the total (Mohnen, 2008) and RGII, the most complex polysaccharide, contributing less than 10% (O'Neill *et al.*, 2004). Most pectin polysaccharides share the same D-Galacturonic acid backbone so chains are named according to their levels of branching and by the components that make up the branches (Figure 1-4). The side chains of pectin polysaccharides are thought to be integral to the overall super-structure as they enable the formation of borate mediated inter strand cross links that add to the strength of plant cell walls. The structure of these side chains is also known to be highly conserved throughout evolution in many plants and very few RGHII mutants have been identified, suggesting they are essential to the plant cell wall structure (O'Neill *et al.*, 2004). In all, there are 12 monosaccharides incorporated into pectin polysaccharides, which are linked with as many as 22 different glycosidic bonds (Harholt *et al.*, 2010). This makes it a very complex molecule both to synthesise and to degrade enzymatically.

Legend for figure:

- ⬡ = D-Galacturonic acid
- ⬡ ⬡ = L-Arabinose
- ⬡ = D−Apiose
- ⬤ = O-Acetyl
- ◯ = L-Rhamnose
- ⬡ = D-Galactose
- ⬡ = L-Fucose
- ⬤ = O-Methyl
- ⬡ = D-Glucuronic acid
- ⬡ = L-Aceric acid
- ⬡ = D-Xylose
- Ⓑ = Borate
- ⬡ = Kdo
- ⬡ = D-Dha
- ⬡ = L-Galactose
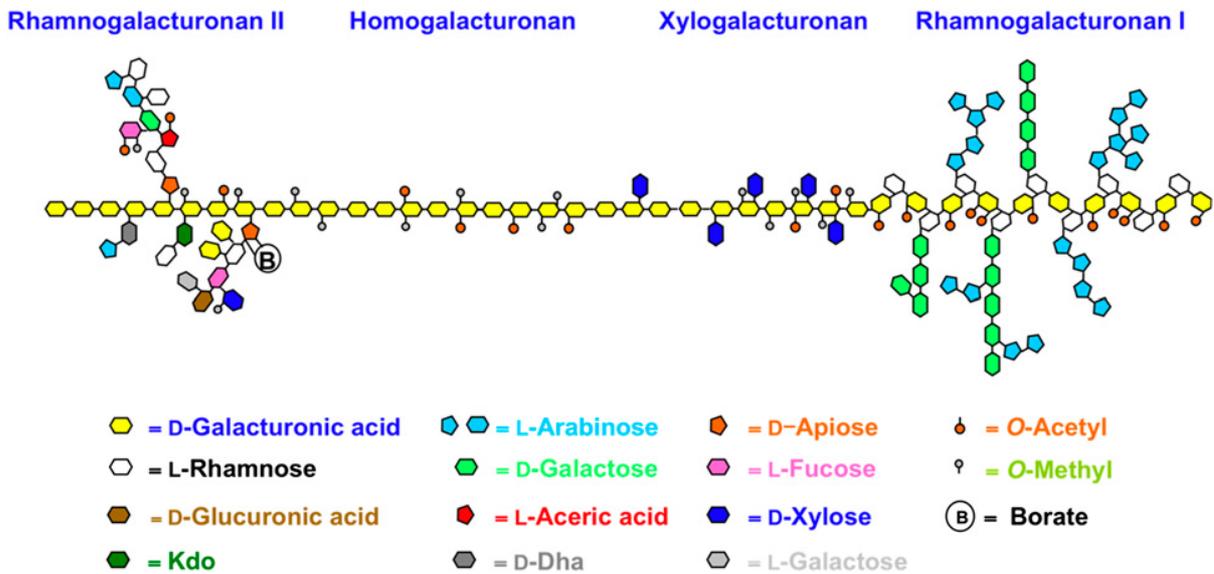
**Figure 1-4 Structure of pectin polysaccharides**

Source: {Harholt, 2010 #15}

Figure 1-4 shows the different structures of pectin glycan chains as well as some of the possible side chain components and conformations associated with each designation.

### 1.4.3 Hemicellulose

Hemicellulose is a low weight polysaccharide (relative to cellulose) composed of several types of monosaccharide, most commonly, but not limited to: D-mannose, D-glucose, L-arabinose, 4-O-methyl-D-glucuronic acid, D-galactose, D-galacturonic acid and D-xylose. As with pectin, the amounts and arrangement of these sugars can substantially affect the properties of hemicelluloses and can vary greatly between species and wood types, such as soft wood and hard woods. The backbone of hemicellulose is composed of beta-1,4 linked sugars, most commonly xylose sugars. These backbone sugars can contain OH subsitions at the O-2 and O-3 positions with sugars such as arabinose and glucironic acid (Scheller and Ulvskov, 2010). Hemicellulose is thought to contribute significantly to the

mechanical strength of plant cell walls, especially in woods (Sweet and Winandy, 1999) by acting as an adhesive holding together cellulose fibres in the three dimensional network (Dahlman *et al.*, 2003). Structurally, it is highly branched, with some regular and some irregular side chain formation. These side chains are mostly one monosaccharide in length and prevent excessive inter hemicellulose strand interactions, which is thought to be the cause of the non-crystaline structure of hemicelullose. An example of this is the common side chain addition of xylose as a substitute to the $CH_2OH$ of glucose molecules. This prevents hydrogen bond formation between chains (Smith, 1977). However hemicelluloses do form inter chain hydrogen bonds with the surface of cellulose microfirbrils where the $CH_2OH$ of glucose molecules in cellulose act as hydrogen bond donators and the oxygen, in the glycosidic bonds in the hemicellulose back bone, act as hydrogen bond acceptors. Large numbers of these interactions can take place, leading to the creation of a monolayer of hemicelluloses around cellulose microfibrils, which also allows hemicellulose to link different microfibrils of cellulose together, adding strength and stability to the cell wall super structure (Bauer *et al.*, 1973). This is also thought to provide protection from enzymatic degradation, to which hemicelluloses are more resistant than the more homologous cellulose (Zheng *et al.*, 2009).

### 1.4.4 Lignin

Lignin is an aromatic heteropolymer that consists of different phenylpropane monomers also known as monolignols. These monomers include but are not limited to: para-coumaryl alcohol, coni feryl alcohol, and sinapyl alcohol (Figure 1-5).
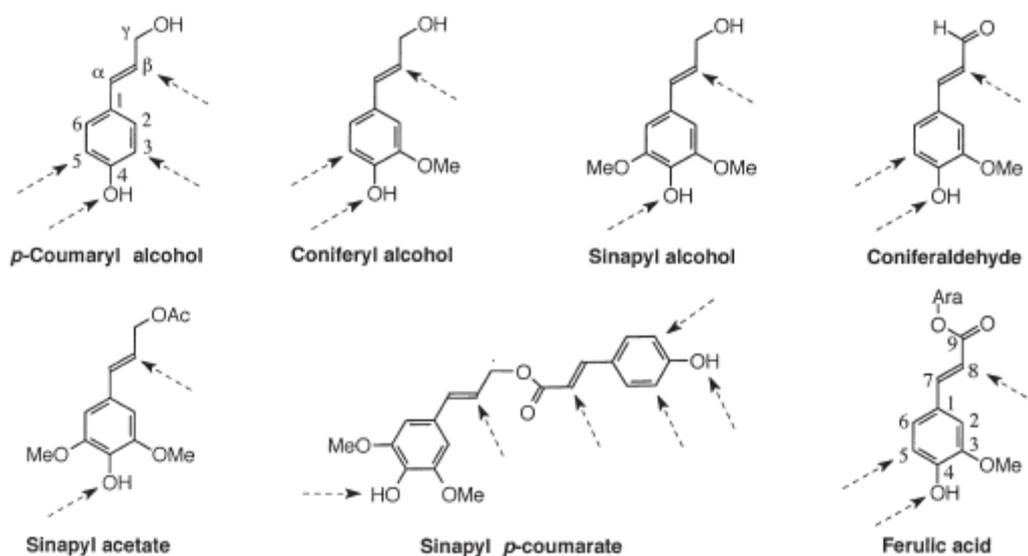
**Figure 1-5 A selection of the possible lignin precursors (monolignols)**

**Source**: (Lebo *et al.,* 2000)

Lignin makes up an integral part of the cell wall of plant cells (Lebo *et al.*, 2000). It is the second most abundant of the organic polymers after cellulose, making up as much as 30% of organic carbon outside of fossil fuels (Boerjan *et al.*, 2003). It has a structural role, providing support in the more fibrous tissues, physiological properties, facilitating transport of water through the xylem, and also a role in protecting plants from infection by limiting the spread of plant pathogens within the host (Iiyama *et al.*, 1994). Lignin is also known to bond with cellulose and hemicellulose (Grabber, 2005). Lignin is highly resistant to biological and chemical degradation. This also means that it protects other cell wall components, thereby reinforcing the plant's structural integrity. The consequence of this protection is that lignin is one of the major barriers to converting cell wall polysaccharides to liquid fuels (Stinson and Ham, 1995).

The protection that lignin offers to cell wall polysaccharides has stimulated many studies of the structure of native lignin and its biosynthesis, in the hope of

identifying a structural weakness that can be exploited or to find plant genes that could be modified to reduce amounts of lignin as a whole or could make the plant produce more uniform versions of the lignin polymer that would be more susceptible to biological or chemical attack. Many of these studies have included radiolabeling experiments and production of transgenic plants, providing improved understanding of the stages of lignin biosynthesis/polymerization, along with new insights into the regulation of these pathways. However, these are not thought to be sufficient to provide a full picture of the processes *in vivo* (Boerjan *et al.*, 2003). The creation of lignin takes place through polymerisation of monolignols. This involves the initial formation of intermediate free radicals created by deprotonation of phenolic hydroxyl groups by peroxidise and lactase enzymes (Whetten and Sederoff, 1995). Two of these radical monomers then couple, forming a dimer then polymerization progresses by the addition of monomers at various positions. The nature of the monomers involved determines which bonds will hold the polymer together. A shematic of the most common bonds and list of these bonds and their prevalence in differnt types of plants are shown in Figure 1-6 and Table 1-1 respectivley. The level of branching caused by these various linkages and the different groups that are involved play a major role in the chemical and biological resistance to degradation of lignin.
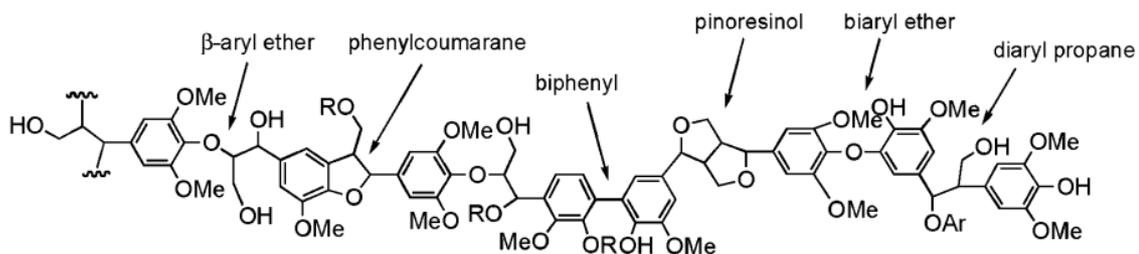
**Figure 1-6 A schematic representation of some of the most common linkages found between lignin monomers**

Source: (Ahmad *et al.*, 2010)

| Linkage type | Softwood (spruce) (%) | Hardwood (birch) (%) |
|---|---|---|
| β-O-4-Aryl ether | 46 | 60 |
| α-O-4-E4Aryl ether | 6-8 | 6-8 |
| 4-O-5-Diaryl ether | 3.5-4 | 6.5 |
| β-5-Phenylcoumaran | 9-12 | 6 |
| 5-5-Biphenyl | 9.5-11 | 4.5 |
| β-1-(1,2-Diarylpropane) | 7 | 7 |
| Pinoresinol | 2 | 3 |
| Others | 13 | 5 |

**Table 1-1    The percentage of possible inter monolignol bonds and their abundance in hardwoods and softwoods**
Source: (Pandey and Kim, 2011)

Despite this extensive research, the native structure of lignin remains unclear. One possible explanation is that its structure may alter during the process of being isolated using current methods (Zakzeski *et al.*, 2010, Pandey and Kim, 2011).

## 1.5  Carbohydrate Active Enymes (CAZymes) and Lignin Degrading Enzymes (Ligninases)

Although carbohydrates have similar chemical compositions, they can form an extremely large number of combinations through the stereochemical variation in hydroxyl positions, the many orders in which monosaccharides can be assembled together and the many non-carbohydrate side groups and substituents that form part of many polysaccharide chains. These carbohydrates are widely distributed in nature and carry out a wide number of functions, with the vast majority (by mass) being used as structural molecules in lignocellulose. The structure of these polysaccharide chains is controlled by the enzymes that build them, the glycosyltransferases, but also those that break them down, which include glycoside hydrolases, polysaccharide lyases and carbohydrate esterases (Lombard *et al.*, 2014). These enzymes collectively are known as the Carbohydrate-Active enZymes (CAZymes). They are split into families relating to structure and activities. To date 133 glycoside hydrolase families, 97 glycosyltransferase families, 23 polysaccharide lyase families and 16 carbohydrate esterase families have been identified (www.cazy.org [accessed 20/01/15]. Many of these groups contain members that have activity against plant cell wall polysaccharides. For example enzymes involved in the complete enzymatic breakdown of cellulose are found in 23 of the glycoside hydrolase groups.

Glycoside hydrolase enzymes are the most numerous and of most use to biotechnology.  They are used in a large number of industries, including the pulp and paper industry, the textile industry, the dairy industry and the brewing industry, to name but a few (Li *et al.*, 2012b). The profitability of the processes in

which they are used has meant that they have been intensively studied and as such there is a very detailed knowledge of their structures, functions and mechanisms.

The modularity of some glycoside hydrolases was first observed after a proteolysis investigation in 1986 (Van Tilbeurgh *et al.*, 1986). These studies led to the discovery of two main functional modules, a catalytic domain responsible for hydrolysis and a carbohydrate binding module (CBM). These CBMs are themselves further categorised into 71 groups (http://www.cazy.org/Carbohydrate-Binding-Modules.html [Accessed 22/01/15]) based on specificity and structure. These CBMs allow recognition and adsorption of the target substrate, causing localisation of the soluble enzyme to the (often) insoluble target, such as to a native cellulose microfibril, thereby facilitating the hydrolytic action of the catalytic module. Structural studies of the CBM show that they are dominated by a "β-Jelly Roll" formation of two β-pleated sheets which contain a hydrophobic planar surface for binding to multiple polysaccharides such as those found in crystalline cellulose, or a cleft region which allows recognition of a single polysaccharide chain (Teeri *et al.*, 2002). The mechanisms by which glycoside hydrolase enzymes catalyse glycosidic bond cleavage are now well understood, with the vast majority falling into two types of mechanism, retaining glycoside hydrolases and inverting glycoside hydrolases. As the name suggests, an inverting mechanism causes inversion of the anomeric configuration of the saccharide being cleaved, which causes the hydroxyl group to assume the opposite orientation (relative to the plane of the ring) from where it began.  While a retaining mechanism results in the cleaved saccharide maintaining its original anomeric configuration, Figure 1-7 demonstrates these processes.

**Retaining glycoside hydrolases:**
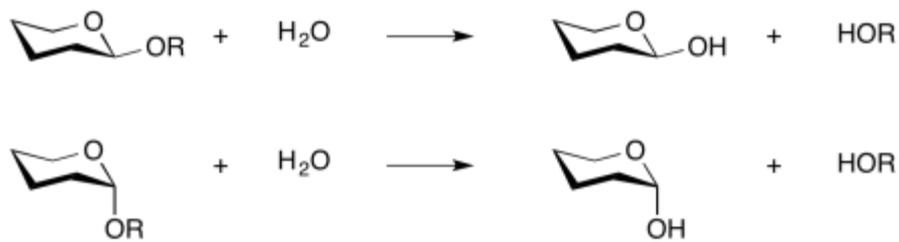


**Inverting glycoside hydrolases:**
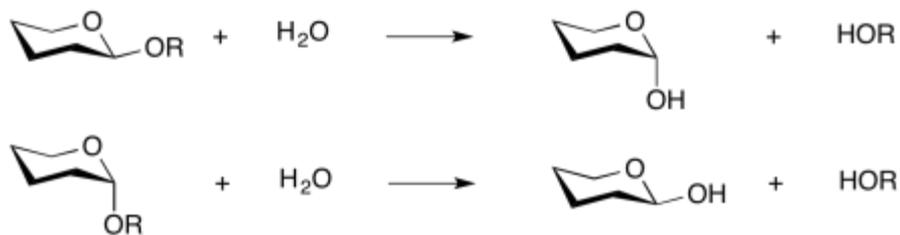


**Figure 1-7 The change in anomeric configuration of a hexose sugar cleaved by both retaining and inverting hydrolysis mechanisms**

**Source**: [http://www.cazypedia.org/index.php/Glycoside_hydrolases accessed 30/01/15]

The mechanism by which inverting glycoside hydrolase activity occurs involves a single step displacement. This is achieved by a nucleophilic substitution of OR by water to form OH, cleaving the saccharide from the chain. This mechanism requires a basic amino acid side chain to donate a pair of electrons to a hydrogen atom and an acidic side chain to act as an electrophile to the oxygen atom in the glycosidic bond (Figure 1-8). The outcome of this mechanism is hydrolysis, with a net inversion of anomeric configuration in the cleaved saccharides.

17

**Inverting mechanism for an α-glycosidase:**
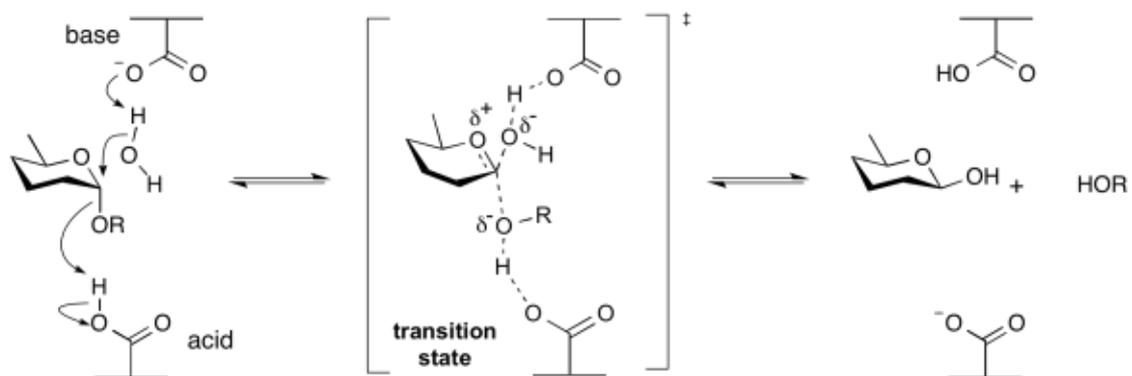


**Inverting mechanism for a β-glycosidase:**
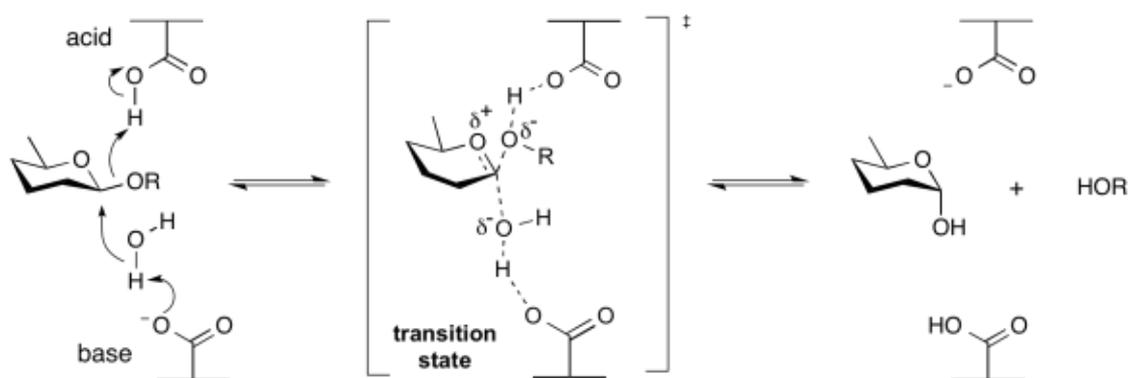


**Figure 1-8 The proposed Inverting mechanism for an α and β glycosidase**

The retaining mechanism of hydrolysis is often called The Classical Koshland mechanism. This mechanism is achieved via two steps; a glycosylation step and a deglycosylation step, through a double displacement mechanism that involves the creation of covalently bonded glycosyl-enzyme intermediates. This mechanism again requires an amino acid acting as an acid and one as a base, often from glutamate and aspartic acid. In the first step, one of the residues acts as a nucleophile attacking the anomeric centre forming the glycosyl intermediate. In the second step, water hydrolyzes the glycosyl intermediate resulting in a cleaved saccharide *(*Figure 1-9*).*

Retaining mechanism for a β-glycosidase:
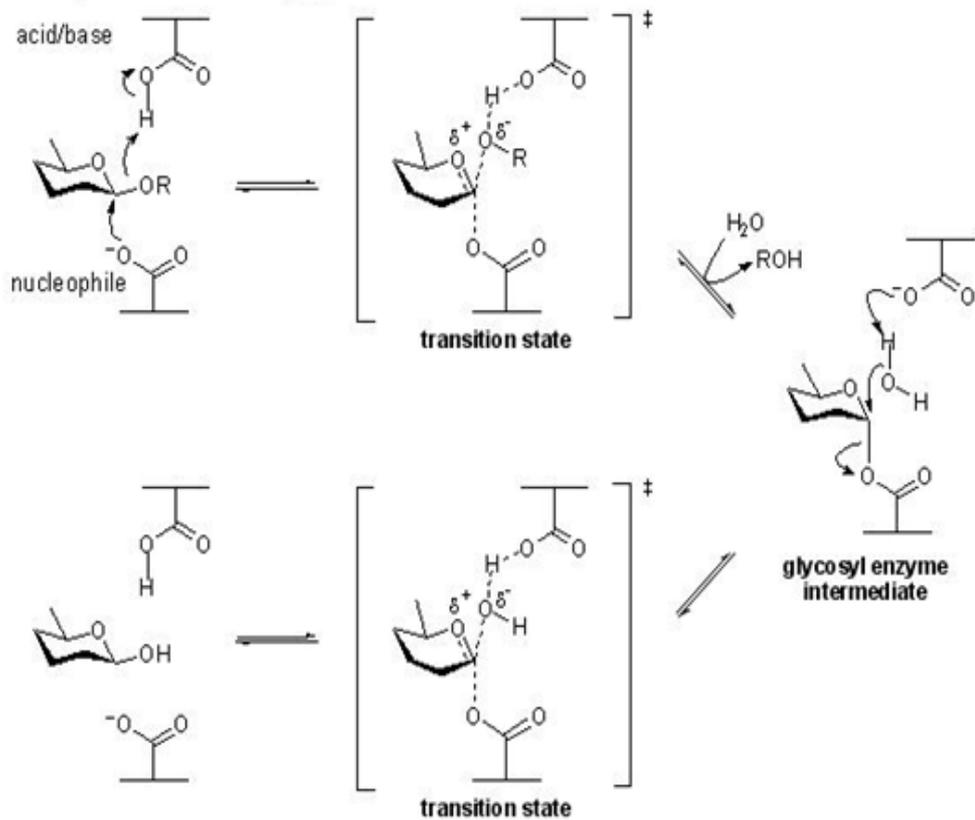
**Figure 1-9 The proposed classical Koshland retaining mechanism for β glycosidase activity**

Source: http://www.cazypedia.org/index.php/File:Retaining_glycosidase_mechanism.png [accessed 22/01/15]

Determination of these mechanisms was assisted by means of visualising x-ray crystal structures that give a snap shot in time of the configuration of proteins, even while they are in complexes with their substrates. Figure 1-10 shows an example of one of these crystal structures showing the 3 dimensional structure of a cellobiohydrolase catalytic domain, in this case from *Trichoderma reesei*.
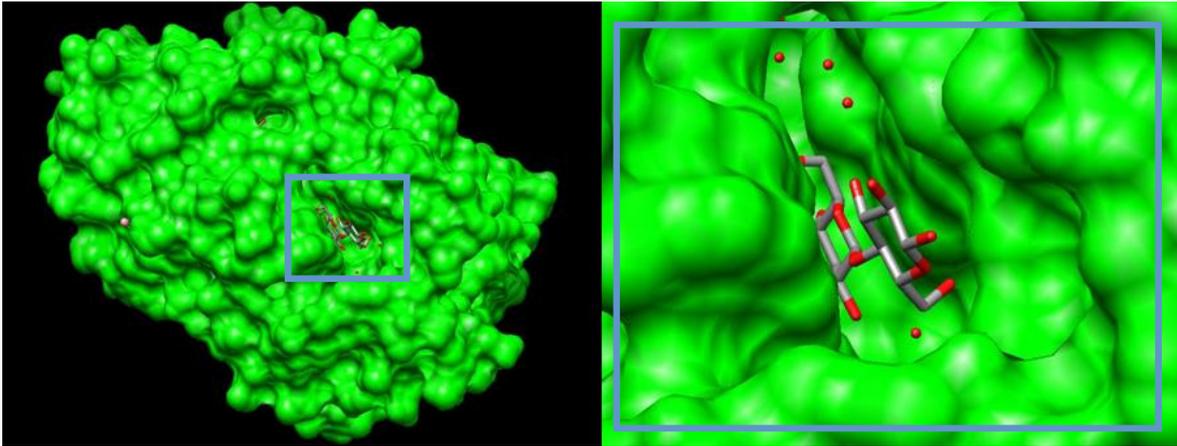
**Figure 1-10 The X-ray crystal structure of a cellobiohydrolase enzyme in complex with a cleaved cellobiose molecule after hydrolysis from a cellulose chain.**
*Source: Image derived from PDB file 7CEL (Divne et al., 1998) manipulated using UCSF Chimera (Pettersen et al., 2004)*

Figure 1-10 shows one of the many x-ray crystal structures now available for CAZyme families, in this case, a cellobiohydrolase from the glycoside hydrolase Family 7 which contains cellobiohydrolses, endoglucanases and chitosanases. The file contains the structure of the enzyme in complex with a short cellulose chain (Divne *et al.*, 1998). The angle at which the image has been taken shows the cellobiose molecule that has been sheared from the reducing end of that chain. The rest of the cellulose chain runs through the centre of the protein. In depth structural analysis of these domains has enabled module designation by analysis of their DNA sequence (Henrissat, 2000).Three dimensional structural analysis using x-ray crystallography has also shown that tertiary structure is conserved within families. Once the tertiary structure of an individual enzyme from any family is derived, that structure can be utilised as a basis for homology modelling and molecular replacement to predict the structure of other members of the same family (Henrissat and Davies, 1997). This is especially useful when the protein of interest is difficult to crystallise or the time and money required to make/process a crystal is not available. The linker regions between the modules

vary in length, from as few as 6 to as great as 69 amino acids in length (Gilkes *et al.*, 1991). Within each group the linker sequences are highly conserved, making them one of the main criteria for group designation. The linker groups are all rich in proline and amino acids with side groups containing hydroxyl groups. Other domains have also been identified but as many as 60 discovered modules have an unknown function and are often termed X modules (Gilkes *et al.*, 1991).

Due to the extensive investigations into lignocellulose pre-treatment that remove much of the native protection of cellulose provided by lignin and hemicellulose, the focus of enzyme research has focused mainly on cellulase enzymes. Currently, it is accepted that the efficient breakdown of native cellulose requires four enzymes that are members of the glycoside hydrolase super family, which work synergistically to hydrolyse glycosidic bonds in cellulose chains producing glucose monosaccharides. Firstly endoglucanase (endocellulase) enzymes (EC 3.2.1.4) are required to decrease the crystallinity of the cellulose microfibrils by causing hydrolysis of glycosidic bonds within the cellulose, which in turn increases the number of free chain ends. These free chain ends can then be targeted by exoglucanase or cellobiohydrolase enzymes (exocellulases) (EC 3.2.1.91) both of which cleave a cellobiose moiety (usually) from the non-reducing end of the sugar chains. The cellobiose disaccharide then goes on to be broken down into two glucose monomers by β-glucosidase enzymes (EC 3.2.1.21)(Talebnia *et al.*, 2010). During this process β-glucosidase plays an especially important role as cellobiose acts as an end product inhibitor to both exo and endoglucanases (Olson *et al.*, 2012).

In essence, this synergistic mechanism facilitates the breakdown of a large crystalline carbohydrate biopolymer into large quantities of simple sugar suitable

for fermentation into bioethanol or other fuels. The steps in cellulose breakdown can be seen in Figure 1-11.
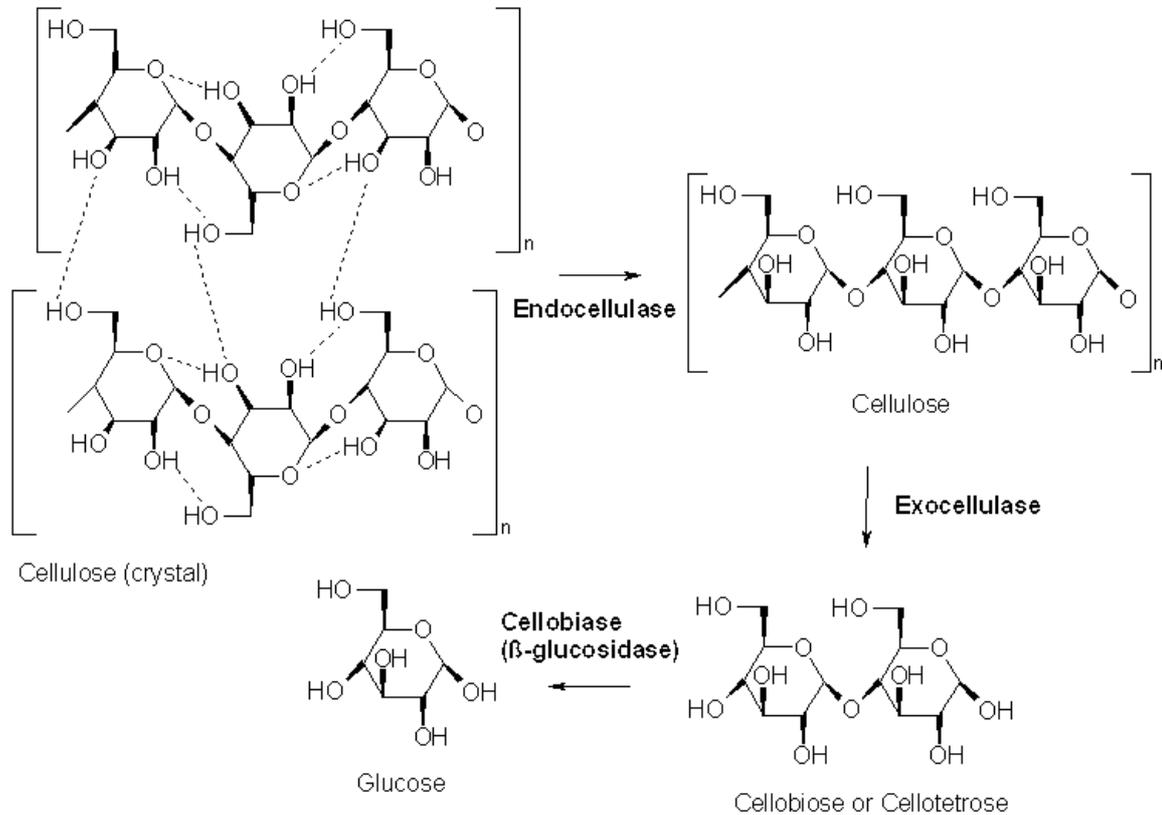


**Figure 1-11 A schematic showing the sequencial breakdown of cellulose to glucose**

For complete breakdown of lignocelluloses, and to gain the highest yields of fermentable sugars possible, the action of hemicellulose enzymes is also required (Menon *et al.*, 2010). Hemicellulase enzymes, unlike cellulases, fall into two super families, the glycoside hydrolase family which break down glycosidic bonds between the monosaccharides and the carbohydrate esterase family which cleaves ester groups found in acetate and ferulic acid linkages, found in locations where the hemicellulose polysaccharide has branched (Shallom and Shoham, 2003). The increased complexity of hemicellulose when compared with cellulose (as previously discussed), means that a wide array of enzymes specific to different sugars and linkage types are required to break bonds between its many different

monomers, including mannoses, xylose, galactose and arabinose monosaccharides in varying orders with various side chains. Figure 1-12 shows the main enzymes required for hemicellulose degradation.
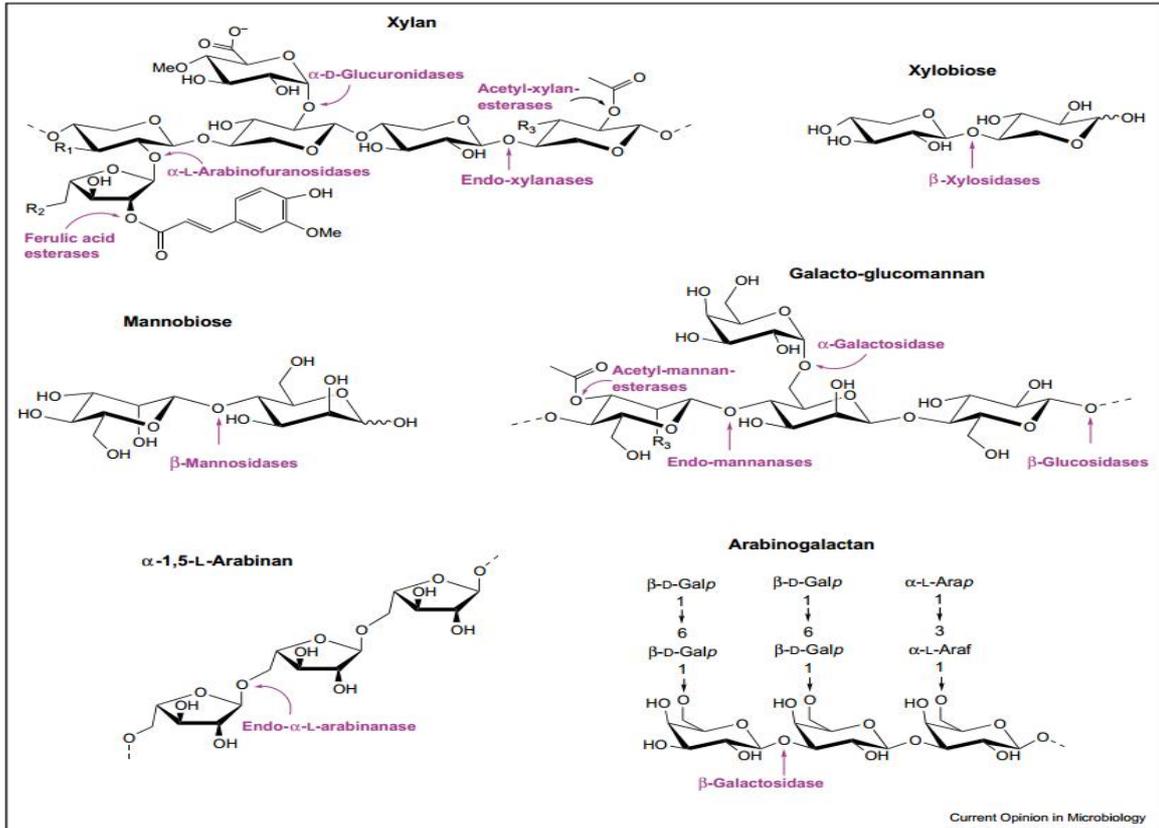


**Figure 1-12 The Enzymes required for complete breakdown of hemicellulose and their respective hydrolysis sites**

**Source**: (Shallom and Shoham, 2003)

Figure 1-12 also shows the numerous side chains found in hemicellulose and demonstrates how they add to the overall structural complexity. With the development of methods that utilize "all in one" reactors (consolidated bioprocessing), the search for effective hemicellulose enzymes has intensified, along with studies to identify microbes that can ferment pentose sugars like xylose in order to utilize these abundant sugars more effctively.

Although many CAZymes are free enzymes that work as independent entities, many anaerobic bacteria also create superstructures called cellulosomes that

contain multiple glycoside hydrolases and carbohydrate binding modules that work sequentially to degrade plant cell wall polysaccharides.

## 1.6  The Cellulosome

Cellulosomes have recently been described as "highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates" (Fontes and Gilbert, 2010). The cellulosome comprises various enzymes of differing specificity that, together, allow the efficient breakdown of cellulose and hemicellulose even in the most complex cell wall systems. The cellulosome system is used by anaerobic bacteria as a nanomachine anchored to the cell surface and held in place by a non-catalytic entity called scaffoldin through highly specific dockerin-cohesin interactions between enzymes and the cell surface (Bayer *et al.*, 1994).

The scaffoldin structure found in the cellulosome was first observed after sequencing of *Clostridium cellulovorans.* This gave valuable information about specificity and lead to a number of other anaerobe scaffoldins being sequenced. This research revealed that the arrangement of these structures was dictated by multi scaffoldin-enzyme gene clusters in the genome of the source microbe (Bayer *et al.*, 2004). This made it possible to ascertain the mechanism by which cellulase and hemicellulase enzymes were ordered in the cellulosome complex. The scaffoldin itself is a relatively large enzyme binding protein that was found to contain multiple copies of cohesion binding regions. These cohesion regions allow specific binding of glycoside hydrolase enzymes and polysaccharide specific modules, such as cellulose binding modules (CBM) (synonymous with cellulose binding domain CBD), all of which themselves contain dockerin binding domains. Scaffoldin proteins and the dockerin cohesin specificity are also important in enabling anchorage, with at least two types of scaffoldin proteins binding to

specific points of the cell surface of the cell holding the cellulosome in the specific orientation required for substrate interaction. Biochemical analysis of the various cohesion-dockerin interactions have shown that the interactions are among the strongest protein-protein interactions observed in nature thus far, causing the cellulosome to have a very stable super macromolecular structure (Mechaly *et al.*, 2001). The conformation of a cellulosome and its components can be seen in Figure 1-13.
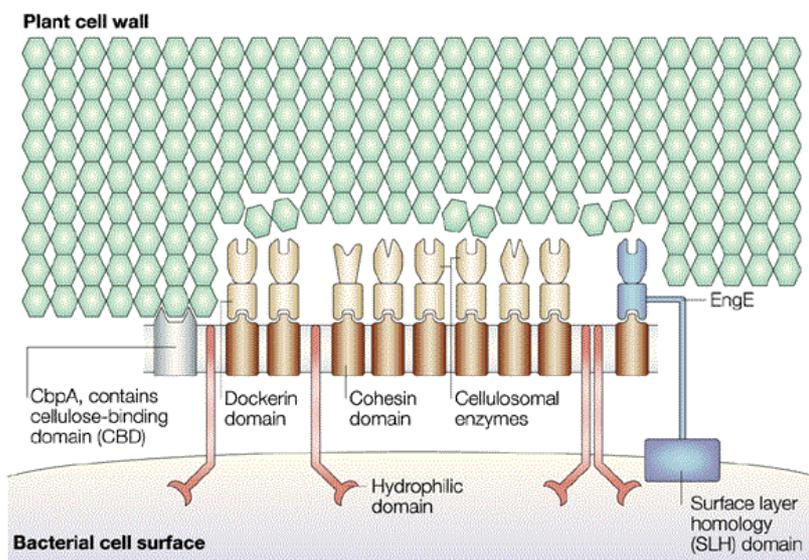
**Figure 1-13 Structure of a typical cellulosome**

Source: (Doi and Kosugi, 2004)

This shows the interactions between the dockerin and cohesion domains of the scaffoldin and the enzymes along with showing the relative location of the CBM and anchorage scaffoldin proteins.

Cellulosomes can integrate various cellulose binding domains bound to the scaffolds to recognise cellulose microfibrils. These interactions have been well studied and are usually of the type 3 CBM. However, in nature, the cellulose microfibrils are often unavailable or out of reach so it is understood that the

cellulosome can undergo super molecular rearrangement, allowing its constituent enzymes to come into contact with their specific substrates. In order to facilitate this, some of the cellulosomal enzymes themselves contain a CBM specific to their substrate (Bayer *et al.*, 2004). After the rearrangement, the cellulosome acts like a small machine that systematically breaks down the surrounding polysaccharide chains. The enzyme subunits of the system are identified by sequencing and subsequent homology searches that match them to the family of carbohydrate hydrolases to which they belong. Many different classes of enzymes have been identified within the cellulosome of different bacterium, including hemicellulases (Murashima *et al.*, 2002), chitinases (Zverlov *et al.*, 2002) and pectate lyases (Tamaru and Doi, 2001). These non-cellulolytic enzymes are even found in cellulosomes of organisms that are known not to utilise hemicellulose or pectin as an energy source, suggesting they are present merely to allow access to the cellulose microfibrils, further facilitating their conversion to simple sugars (Bayer *et al.*, 1994).

The intricate nature of the cellulosome and its ability to digest even the most complex of the protective plant cell wall structures makes them of great interest to the lignocellulosic bioethanol industry.

## 1.7  Lignocellulose For Use In Bioethanol Production

Using conventional biorefining methods, there are three main stages in the production of bioethanol from lignocellulosic feedstocks. The first stage is pre-treatment of the feedstock that increases the availability of plant biomass carbohydrates for hydrolysis into simple sugars. The next stage involves hydrolysing the carbohydrate chains in the pre-treated feedstock material, using either biochemical or physiochemical methods, into simple sugars. The final stage

utilizes specially selected and often modified yeasts and bacteria that ferment the resulting simple sugars to produce ethanol.

## 1.7.1 Pre-treatment

Pre-treatment of the feedstock material is carried out in order to maximise the efficiency of utilization of sugars during the hydrolysis and fermentation stages. This is achieved by removing lignin and often hemicellulose, reducing the level of crystallinity in cellulose fibrils, and increasing the porosity of the material. To be effective, pre-treatments must; increase the formation of simple sugars in subsequent enzymatic treatments, avoiding loss of carbohydrate chains through degradation and minimise the production of by-products that will inhibit enzymatic hydrolysis, while remaining cost effective (Sun and Cheng, 2002). The complexity and intricacy of the cell wall matrix, as discussed previously, results in a low porosity and makes lignocelluloses very resistant to hydrolysis by either physico-chemical and enzymatic methods. The importance of carrying out pre-treatments is apparent from studies that have shown how the theoretical yield of ethanol may be reduced by 80% if pre-treatment of the biomass is not carried out (Zheng *et al.*, 2009). This means that, although these treatments are often time consuming and somewhat expensive, they are very likely to be worthwhile. However, given the many different feedstocks available and the individual requirements of each of them, there is no "perfect" pre-treatment method (Kurian *et al.*, 2013) leading to many different methods being developed. These include physical, physico-chemical, chemical based methods, each of which is discussed below.

### 1.7.1.1 Physical Pre-treatments

Physical pre-treatment methods are sometimes used in the first instance to simply reduce the size of the starting material. Mechanical comminution combines

chipping, milling and grinding of feedstocks to reduce particle size. Chipping will reduce particles to 10-30 mm while milling may reduce it to as low as 0.2mm (Sánchez and Cardona, 2008), considerably increasing surface area and reducing cellulose crystallinity. Reducing particle size has enabled substantial reductions in subsequent hydrolysis time, of between 23 and 59%, depending on feedstock (Hartmann *et al.*, 2000). This increases efficiency with respect to time and yield but the large amount of energy required to power large milling machinery can outweigh the advantages, making the process economically unviable (Ghosh and Ghose, 2003) especially for hardwoods that can require 130 kWh/ton of wood to produce a particle size of 1.6 mm whereas, for the same results, corn stover (the leaves and stalks of maize) only requires 14 kWh/ton (Cadoche and Lopez, 1989). An alternative to grinding and milling is pyrolysis, where high temperatures are used to increase cellulose breakdown and chemical degradation. At temperatures greater than 150°C, hemicellulose begins to solubilise and break down, an exothermic process that causes partial hemicellulose hydrolysis and the formation of acidic products that are thought to act as a catalyst for further break down of the cell wall matrix (Zhu *et al.*, 2004, Liu and Wyman, 2003). The temperatures required to initiate this process depend on the level of branching and components of the hemicellulose chains, which, as previously mentioned, are highly variable. At temperatures near 160°C the lignin component of the plant cell also begins to solubilise and degrade, which may seem to be an advantage but, in fact, creates phenolic breakdown products that are highly toxic to many bacteria, fungi and archaea, causing downstream problems when fermentation is being undertaken (Gossett *et al.*, 1982). Furthermore, if not removed rapidly, these phenolic products can react with hemicellulose hydrolysates to form various compounds

that then precipitate. This is especially problematic during high temperature pre-treatments. Laser *et al.* (2002) reported that when solid precipitates such as furfurals and compounds like vanillin or vanillin alcohol exceed 3% during pre-treatment at 220°C subsequent ethanol production was almost completely inhibited. A common name for lignin breakdown products is lignin carbohydrate complexes or LCCs, where the phenolic components have formed complexes with hemicellulose oligomers (Gupta and Lee, 2010).

### *1.7.1.2 Physico-chemical Pre-treatments*

One of the first physico-chemical pre-treatments developed involved the use of steam. High pressure steam is utilised to create temperatures of up to 270°C for an appropriate time (determined by water content of the biomass source) (Hendriks and Zeeman, 2009). A modern extension to this method is called steam explosion (SE) treatment. This method uses chipped feedstocks that are treated with high-pressure steam. The pressure is then lowered very rapidly, which causes water within the biomass to expand rapidly or 'explode' which is thought to increase further the accessibility of cellulose to enzymes (Ruiz *et al.*, 2008). This method also allows some carbohydrate hydrolysis to occur when acetic acid is liberated from hemicellulose and water and its weak acidic nature at high temperatures facilitates the degradation of both hemicellulose and cellulose. This method is carried out at between 190-270 °C for periods of between 1 and 10 minutes (Kurian *et al.*, 2013). Although steam explosion is much more energy efficient than physical methods, it results in the loss of much of the hemicellulose portion of the biomass, incomplete separation of lignin, and production of many compounds thought to inhibit microbes used at the later hydrolysis and fermentation stages (Sun and Cheng, 2002) . Ammonia fibre explosion (AFEX)

works on the same principle as steam explosion. Feedstocks are subjected to ammonia at high temperatures and pressures, followed by sudden pressure reduction leading to similar rapid decompression that causes separation of plant cell wall components and their hydrolysis. This has advantages over steam explosion because it is only necessary to reach temperatures of 90°C (Menon and Rao, 2012).

### 1.7.1.3  Chemical Pre-treatment

Chemical pre-treatment methods are the most successful and, thus, the most common pre-treatment methods used in lignocellulose processing to date, again with the purpose of solubilising lignin along with a portion of the hemicellulose biomass. Weak or strong acids can be used in the process but both methods use low temperatures. This saves energy but biomass slurries must then be neutralised for downstream processing (Sun and Cheng, 2002). In the acidic conditions created in this process hemicellulose begins to hydrolyse into short oligomers and monomers, mainly due to the instability of xylan monomers at acidic pHs (Saha *et al.*, 2005). Dilute acid treatment also helps to solubilize lignin, which in turn improves yields during downstream enzymatic hydrolysis of cellulose (Menon and Rao, 2012).

Alkaline conditions can also be employed in lignocelluloses degradation. Weak alkaline processing uses sodium hydroxide (NaOH) or ammonia ($NH_3$). The means of optimizing alkaline pre-treatment varies according to the feedstock being used. For example, ammonia is more effective when using feed stocks with low lignin concentrations whereas NaOH is more suitable for materials with more lignin, such as hard woods (Gupta and Lee, 2010). The aim of weak alkaline degradation is the selective removal of lignin, while preserving the hemicellulose

and cellulose, from which many the monosaccharide components can be converted into fuel. The reactions involved in lignin alkaline pre-treatment are mainly saponification, where ester bonds within the lignin superstructure break in the presence of bases such as NaOH (Hendriks and Zeeman, 2009). The use of ammonia is preferable to NaOH as it selectively reacts with lignin over hemicellulose, leaving hemicellulose intact for more efficient extraction while still increasing accessibility of cellulose. Another consideration is base concentration. At very high pH values some degradation selectivity is lost and a reaction called "endwise peeling" occurs. Endwise peeling describes the step-by-step degradation of cellulose and hemicellulose where a monosaccharide unit is removed in a rearrangement and elimination reaction mechanism; the resulting product is described as a "meta-saccharinic acid". This reaction continues until the reducing end of the cellulose chain is no longer present (Yoneda *et al.*, 2008) and is, for example, replaced by a carboxylic acid group.

Each of these methods has their advantages and disadvantages for certain feedstocks and situations. With the vast majority of these methods one of the most important considerations is the compatibility with enzymatic hydrolysis and fermentation steps, which are required after all pre-treatment steps to complete the breakdown of the plant biomass to yield the highest possible proportion of simple sugars. A summary of the individual and combinatorial effects of each pre-treatment can be seen in Table 1-2, summarised by (Hendriks and Zeeman, 2009). One limitation each of the methods described share is high costs.

Effects of the different pretreatments on the physical/chemical composition or structure of lignocellulose

| | Increase accessible surface area | Decrystallization cellulose | Solubilization hemicellulose | Solubilization lignin | Formation furfural/ HMF | Alteration lignin structure |
|---|---|---|---|---|---|---|
| Mechanical | + | + | | | | |
| ST/SE | + | | + | − | + | + |
| LHW (batch) | + | ND | + | − | − | − |
| LHW (flow through) | + | ND | + | +/− | − | − |
| Acid | + | | + | − | + | + |
| Alkaline | + | | − | +/− | − | + |
| Oxidative | + | ND | | +/− | − | + |
| Thermal + acid | + | ND | + | +/− | + | + |
| Thermal + alkaline (lime) | + | ND | − | +/− | − | + |
| Thermal + oxidative | + | ND | − | +/− | − | + |
| Thermal + alkaline + oxidative | + | ND | − | +/− | - | + |
| Ammonia (AFEX) | + | + | − | + | − | + |
| CO$_2$ explosion | + | | + | | | |

+ = major effect.
− = minor effect.
ND = not determined.

**Table 1-2 Summary table of lignocellulose pre-treatment approaches**

**Source**: (Hendriks and Zeeman, 2009)

The requirement of these expensive pre-treatments could be reduced significantly if highly active, stable ligninase enzymes were discovered that could facilitate the delignification of feedstocks. These enzymes could potentially be incorporated into the enzyme cocktails used during hydrolysis stages, or incorporated into the genomes of fermentative microbes alongside CAZymes to remove the need of pre-treatment for both traditional bioethanol production methods and in consolidated bioprocessing methods.

### 1.7.2 Hydrolysis

Pre-treatments create slurries that comprise plant biomass that has been made highly accessible to hydrolysis into monosaccharides that can be fermented in later stages. At present there are two main methods used for hydrolysis, one using acids and the other using microbial enzyme cocktails, with enzymes known to produce much higher overall yields of ethanol (Galbe and Zacchi, 2002). During enzymatic hydrolysis, slurries are treated with enzyme cocktails containing a multitude of carbohydrate active enzymes. The content of these cocktails can be tailored to the feedstock and pre-treatment method used. For example, if

hemicellulose and cellulose are not separated by pre-treatment, cocktails will contain both hemicellulose degrading and cellulase enzymes amongst others. The specific enzymes required for total breakdown of lignocellulose carbohydrates and the actions associated with them were discussed in detail in section 1.5. Briefly, cellulose degrading enzymes will include endocellulases targeting long chain cellulose, exocellulases and cellobiohydrolases, which release cellobiose disaccharides from chain ends which are then broken down into two glucose molecules by β-glucosidase enzymes. The heterogeneous nature of hemicellulose means many more enzyme types are required for its complete degradation. These include enzymes that target glycosidic bonds between xylose, mannose, arabinose and galactose as well as carbohydrate esterase enzymes that target the many sidechains found in hemicellulose. After incubation with these enzyme cocktails, the slurry becomes rich in simple sugars. These include 6 membered ring sugars such as glucose, galactose and mannose and with 5 membered ring sugars such as xylose. Although enzymatic hydrolysis provides much better yields than other methods, the process is still time consuming, which increases overall costs by decreasing productivity. The cost of these enzyme cocktails is also quite considerable, with estimates of between US$0.30/Gallon to US$0.40/Gallon (Klein-Marcuschamer *et al.*, 2012) and up to greater than US$1/Gal (Petiot, 2008). This makes enzyme production one of the most expensive components of the whole process. It is widely thought that significant saving could be made if better lignocellulose degrading organisms/enzymes were developed or discovered (Lynd *et al.*, 2008).

### 1.7.3 Fermentation

The simple sugars produced by hydrolysis of cellulose and hemicellulose are then used as a substrate to create ethanol by means of microbial fermentation. Fermentation takes place in a bioreactor containing modified yeast strains such as *Saccharomyces cerevisiae* (Romaní *et al.*, 2015) or bacteria such as *Clostridium* species (Jin *et al.*, 2014) and *Zymonas mobilis* (Claassen *et al.*, 1999). These can produce ethanol very efficiently from hexose sugars with yields of up to 97% of theoretical maximum being achieved (Talebnia *et al.*, 2010). Yeast fermentation will produce 2 ethanol molecules from a single glucose molecule, with 2 $CO_2$ molecules made as a by-product (Equation 1).

**Equation 1-1 Chemical equation for the fermentation of a single glucose monosaccharide to 2 ethanol molecules by yeast**

$C_6H_{12}O_6 \rightarrow 2\ C_2H_5OH + 2\ CO_2$

Glucose is subjected to glycolysis to produce 2 pyruvate molecules; these are then decarboxylated by pyruvate decarboxylase to acetaldehyde, producing carbon dioxide. NADH then reduces acetaldehyde to ethanol. It is important to note, however, that many of the simple sugars will be pentose sugars like xylose, the most abundant component of hemicellulose. However few native microorganisms are able to ferment both hexose and pentose sugars efficiently without producing undesirable by products (Stambuk *et al.*, 2008). For industrial ethanol production an ideal microorganism should be able to ferment a broad range of substrates, have a high ethanol production yield and rate, have high tolerance to inhibitors produced during pre-treatment stages, and be tolerant to high ethanol concentrations while exerting a high level of cellulolytic/hemicellulolytic activity (Hahn-Hagerdal *et al.*, 2007).

### 1.7.4 Separate Hydrolysis and Fermentation (SHF) and Simultaneous Saccharification and Fermentation (SSF)

Enzymatic hydrolysis and microbial fermentation steps can either be carried out independently, using separate hydrolysis and fermentation (SHF), or at the same time using simultaneous saccharification and fermentation (SSF) (Peng and Chen, 2011). SHF involves, as preliminary step, an incubation period after the addition of enzyme cocktails to the biomass slurry. During this process polysaccharides are degraded into simple sugars. These are then transported to a biofermenter to be used as substrate for microbial fermentation to ethanol by yeast or bacterial strains.

Initially, separate hydrolysis and fermentation was employed because this allowed for each stage to be fully optimized for conditions such as temperature and pH, as the requirements can vary greatly for each process (Erdei *et al.*, 2012). Another advantage of SHF is that the yeast produced during the separate fermentation step can be purified and reused whereas, once SSF has ended, the yeast is difficult to separate from lignin solids and is lost (Olofsson *et al.*, 2008).

SSF technology has been in development since the late 1970s (Blotkamp *et al.*, 1978) and has become popular because of its much greater efficiency and ethanol production yields than SHF and also for its overall cost effectiveness (Tomas-Pejo *et al.*, 2008, Wingren *et al.*, 2003). This was primarily because SSF overcame the end product inhibition of β-glucosidase enzymes by glucose, which greatly limited hydrolysis in SHF production (Ballesteros *et al.*, 2004). During SSF, as glucose is liberated from the slurry biomass it is fermented into ethanol, preventing end product build up and the resulting end product inhibition.  This allows the process

to achieve much higher ethanol yields, even reaching theoretical maximum yields for some feedstocks.

The main issue for both of SHF and SSF processes is the presence of inhibitory compounds in the biomass slurries produced. The formation and concentration of these compounds is controlled by a multitude of factors including; feedstock type, amount of solids in the reactor, the pre-treatment conditions and durations. This makes it very difficult to standardise the reaction conditions for downstream hydrolysis and fermentation processes. These inhibitory compounds are commonly categorised into three groups: phenolic compounds such as vanillin, syringaldehyde and conferyl aldehyde; weak acids such as acetic acid, formic acid and levulinic acid; and furaldehydes including furfural and 5-hyroxymethyl-2-furaldehyde (HMF). All of these have a negative impact on the hydrolysis and fermentation processes (Parawira and Tekere, 2011). Furaldehydes are especially detrimental to SSF processing as they inhibit protein and RNA synthesis in fermenting microbes, thereby reducing enzymatic activity (Liu *et al.*, 2004, Modig *et al.*, 2002). The presence of these by-products is often unavoidable so the enzymes used in hydrolysis and the microbes used during fermentation should, ideally, be highly resistant to their effects as well as showing heat stability and high levels of activity. These requirements have triggered many studies to discover new highly active and stable lignocellulolytic enzymes from nature (Pope *et al.*, 2012, Tokuda and Watanabe, 2007, Watanabe *et al.*, 1998, Cardoso *et al.*, 2012a, Peng and Chen, 2011, Xia *et al.*, 2013) along with research on metabolic engineering of microbes designed to improve the overall yields of ethanol (Ingram *et al.*, 1998, Edwards *et al.*, 2011, Zhang *et al.*, 1995, Wang *et al.*, 2013) .

## 1.8 Consolidated Bioprocessing and Metabolic Engineering

SHF and SSF both have advantages and disadvantages but what they both have in common is the added costs of multiple steps, including off-site production of cellulase cocktails that are manufactured in expensive, time consuming processes in separate reactors. As previously discussed the cost of offsite enzyme production can exceed US$1/Gallon with lower estimates of around US$0.50/Gallon that in many cases is comparable or even exceeds the cost of the feedstock. Olson *et al.* (2012) comments that in almost no commodity process are the costs of catalysts used comparable to the costs of the raw feedstock materials in the process. This has been a long standing barrier to the industrialisation of bioethanol and considerable research has been undertaken to circumvent the costs of multiple stage processes.

Consolidated bioprocessing (CBP) is the conversion of lignocellulose into the desired products, such as ethanol, in a single step without addition of externally produced enzymes. This involves combining the four biological events required to produce ethanol from pre-treated biomass; production of required CAZymes (cellulases and hemicellulases), hydrolysis of plant cell wall polysaccharides into simple sugars, fermentation of hexose (glucose, galactose and mannose) and pentose (xylose and arabinose) sugars in one bioreactor (van Zyl *et al.*, 2007). This requires microbes that can carry out all of these processes quickly and efficiently while also being resistant to the build-up of inhibitory compounds and toxic products, such as ethanol itself. To date no natural microbe has been discovered that can fulfil all of these desired properties, although many bacteria and fungi are known to be able to perform some of the necessary processes (van Zyl *et al.*, 2007). In order to harness all of the properties that are seen in several

different microbes in a single organism capable of being used in industry, genetic engineering will be required (Olson *et al.*, 2012).

There are two main genetic engineering strategies that can be used to produce CBP microbes. The native strategy involves metabolic engineering steps carried out on naturally occurring CAZyme-producing microbes in order to improve their production-related properties such as yields. The recombinant strategy involves engineering non-cellulolytic microbes that can produce very good product yields naturally, to express the required repertoire of cellulase and hemicellulose enzymes that enable efficient hydrolysis of plant cell wall components (Lynd *et al.*, 2005). The recombinant strategy is the focus of most genetic engineering research, mainly due to the relative ease of inserting individual CAZymes compared to metabolic pathway engineering for fermentation. A summary of these processes can be seen in Figure 1-14.
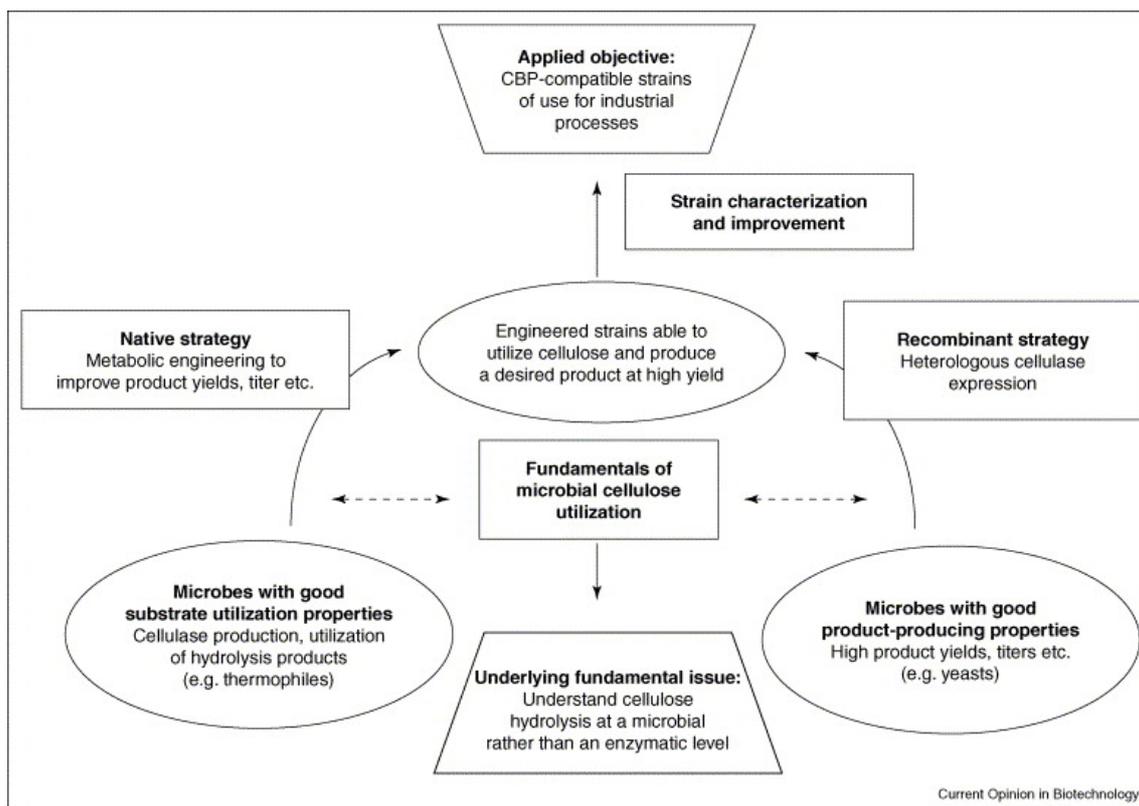
**Figure 1-14 The genetic engineering strategies used in CBP microbe development**

**Source**: (Lynd *et al.*, 2005)

In recent years there have been great advances in the metabolic engineering of microbes with the goal of increasing ethanol production. These include improvements in ethanol production in yeast species such as *Fusarium oxysporum* (Xiros and Christakopoulos, 2009) and *Trichoderma reesei* (Xu *et al.*, 2009) and also in bacterial systems using Clostridium species (Tolonen *et al.*, 2009) and *Thermanaerobacterium* species (Shaw *et al.*, 2011). There have also been significant improvements to many native ethanol producing organisms using recombinant strategies. These include introduction of new CAZymes, increases in levels of overall CAZyme expression, and improvements in specific activity levels. These studies have focussed on engineering anaerobic microbes including *S. cerevisiae*, *E. coli* and *Klebsiella oxytoca* (Lynd *et al.*, 2005) in order to avoid the additional costs of slurry aeration and loss of biomass through oxidative

metabolism. Studies involved in the addition of CAZymes to the enzymatic systems of *S. cerevisiae* have been the most successful, with several studies succeeding in inserting and expressing the genes of enzymes from 30 CAZy families (Olson *et al.*, 2012). In addition, studies inserting CAZyme genes have also been successful in many bacterial species. Multiple xylan degrading *E. coli* species have also been created, with one strain breaking down and fermenting 63% of xylan in a birchwood feedstock without addition of external enzymes (Shin *et al.*, 2010). The highly ethanologenic species *Klebsiella oxytoca* has also been engineered to express cellulase enzymes from the non enthanologenic bacteria *Erwinia chrysanthemi* (Zhou *et al.*, 2001) and it has subsequently been proved to hydrolyse cellulose and ferment the resulting sugars without supplementary CAZymes (Zhou and Ingram, 2001). Although consolidated bioprocessing is thought to offer a solution to many of the issues faced by SSF and SHF, it still not a mature technology and requires more research to discover adequate CAZymes and to develop microbes that can fulfil the many requirements of industrial processing of plant biomass.

## 1.9 Lignocellulose Feeding Organisms: Exploiting Successful Natural Mechanisms

Despite the recalcitrance of lignocellulose, many animals rely on plant biomass as their sole source of food. The inherent ability of many animals to utilize plant cell wall materials in their diet has stimulated a very large number of studies into the mechanisms by which herbivores manage to breakdown lignocellulose and their ecological origins, a method known as "bioprospecting" (Mateo *et al.*, 2001). Researchers have sought to study the breakdown of cellulose in the gut of lignocellulose feeding organisms in order to identify the enzymes responsible for

efficient breakdown into simple sugars, with the hope of harnessing biochemical systems that have developed over millions of years to improve lignocellulose breakdown on an industrial scale. Many studies have looked at the cellulolytic capability of arthropods, since they are widely considered to be the most successful group of organisms in the world and many rely on lignocellulose as a main part of their diet. Early strategies involved research into plant pest species, including termites (Watanabe *et al.*, 1998) and cockroaches (Scrivener and Slaytor, 1994). Following the success of these initial studies, many more ensued and CAZymes were found to be present in the gut of 20 insect families (Sun and Scharf, 2010). The majority of these studies showed that breakdown was carried out predominantly by symbiotic microorganisms but some endogenous cellulase enzymes have been identified in the orders *Orthoptera* (Kim *et al.*, 2008) , *Coleoptera (Wei et al., 2006) an*d *Isoptera* (Zhang *et al.*, 2009). These studies only found endoglucanase enzymes, with no evidence that any insects produced cellobiose hydrolase enzymes. This led researchers to believe that even species from these orders must depend upon symbiotic degradation of lignocellulose by microbial enzymes (Willis *et al.*, 2010b).

Probably the best studied wood feeding arthropods are the termites, with the first successful investigation of their lignocellulose breakdown by Cleveland in 1924 and the first sequenced endogenous CAZyme gene being sequenced in 1998 by Watanabe *et al.* (1998). Termites were chosen as a target for enzyme prospecting due to their diet, which consists solely of lignocellulose and due to their overall importance to the global turnover of carbon in many environments (Warnecke *et al.*, 2007). It was found that many higher termites can themselves secrete CAZymes as could many microorganisms found in the termites gut (Tokuda and

41

Watanabe, 2007, Warnecke *et al.*, 2007). However, it has been found that the lower termites rely completely on gut microbes for the production of CAZymes (Ohkuma, 2008).

Another organism of interest in the past 2 years has been the gribble *(Limnoria quadripunctata) of the order Isopoda.* This is a wood boring organism, surviving solely on a lignoceullosic diet. These crustaceans were previously considered a virulent pest by sailors due to the great amounts of damage they can cause to wooden ships. As with many other pests, it was the gribble's reputations that lead to studies of the mechanism it used to degrade lignocellulose. Interestingly, these studies showed that breakdown was not by microbial gut flora, instead finding that the gut content of the *limnoriids* was bereft of microorganisms. *This led* to the conclusion that degradation must be due to the production of CAZymes by the crustaceans themselves (King *et al.*, 2010)*.* This hypothesis was tested and proved after analysis of the transcriptome of the hepatopancreas, an organ involved in enzyme production and nutrient uptake in crustaceans. The transcriptome that was produced showed that as many as 27% of the genes being expressed were members of the glycoside hydrolase super family (King *et al.*, 2010). The most interesting discovery in the gribble was the identification of many enzymes of the GH7 sub family, which are usually cellobiohydrolases that had not been reported previously in animal genomes but only in fungi (Teeri *et al.*, 1998) and other microorganisms (Todaka *et al.*, 2007). A second interesting discovery was also made byKing *et al.* (2010), with the identification of high percentages (17%) of hemocyanin transcripts. These enzymes are now being implicated in the breakdown and modification of lignin following discovery of their phenol oxidase

activity on aromatic groups in lignin, observed in the *Amphipoda, Gammarus pulex* (Zimmer and Bartholmé, 2003).

Further studies have been carried out that aim to modify newly discovered CAZymes to enhance specific activity and temperature stability to make them compatible with industrial processes. Random mutagenesis of non-conserved residues can be applied to native protein sequences in the hope of improving their properties. One such study by Ni *et al.* (2007) successfully modified a native termite enzyme and increased its thermal stability by 10°C resulting in 90% and 54% retention of activity at 50°C after 30 and 150 minutes respectively. The specificity and Vmax of the protein was also improved subsequently using family gene shuffling, combining parts of gene sequences of enzymes from the same family. This has been observed to increase activity by as much as 13 times that of the parental enzymes (Ni *et al.*, 2007). These studies suggested that the gut of herbivores can be a treasure trove of novel CAZymes, harbouring enzymes of both animal and microbial origin. It is therefore clear that further study of these environments could allow us to uncover novel CAZymes and novel lignocellulose degradation mechanisms that have allowed herbivores to thrive worldwide. If these hidden systems could be characterized they could be exploited to improve conversion of lignocellulose into biofuels.

## 1.10 Microbial Lignocellulose Degradation

The small number of animal cellulases discovered has led to considerable interest in the potential for microbial enzymes (cellulases, hemicellulases and lignases) to bring about the biological breakdown of lignocellulose. Of particular interest is the scope for degradation by the symbiont microbiota in wood/plant feeding invertebrates. Mutualisms between microbes and invertebrates have been widely

studied and are found in almost every case, they facilitate exploitation of many different food sources by hosts, including plant cell walls which are difficult and sometimes impossible for most animals to digest (Watanabe and Tokuda, 2010). However the enzymatic contributions of microbes to the herbivory of some insect orders is still unclear. Some herbivorous insects possess genes encoding plant cell wall degrading enzymes including a termite which produces its own cellulase (Watanabe *et al.*, 1998), but the overall structural complexity of the plant cell wall superstructure requires a multitude of enzyme classes which gut microbes contribute to. It is therefore thought that the interactions of host and microbe has had a direct impact on the evolutionary transitions in diet in many herbivorous eukaryotes, including some invertebrates (Hansen and Moran, 2014). Recent studies into the gut microbiome of termites using metagenomics identified a large number of plant cell wall degrading CAZymes (Warnecke *et al.*, 2007). Metagenomics has also allowed the identification of a large number of microbial CAZymes in mammalian herbivore guts, including reindeer (Pope *et al.*, 2012) and cattle (Hess *et al.*, 2011), with microbial CAZyme identifications numbering in the thousands. Enzymatic activity has been studied extensively in the digestive fluid of various insects including members of the orders *Isoptera (Konig et al., 2013), Coleoptera* (Dojnov *et al.*, 2013) and *Othoptera* (Shi *et al.*, 2011)*,* all of which have a high lignocellulose diet. However, this focus on arthropods and mammals has been at the expense of other groups such as gastropods. Specifically, there has not yet been a definitive characterisation of the origin of cellulolytic activity in the gut of the common garden slug, *Arion ater,* a significant pest species throughout Europe.

## 1.11 The Black Slug: *Arion ater*

Slugs are commonly encountered organisms that are members of the order *Pulmonata*, and found in many terrestrial and aquatic ecosystems worldwide. The common black slug, *Arion ater*, is particularly prevalent in Western Europe and North America. Figure 1-15 shows the brown variant of *A. ater*.



**Figure 1-15 The common black slug, Arion ater (Brown variant common to North West England)**

**Source**: http://www.english-country-garden.com/ [accessed 20/01/15]

These slugs live nocturnal lives, emerging to feed on a variety of foodstuffs including vegetation (both live and decaying), carrion and fungi. Terrestrial slugs require moist environments for locomotion, preferring to shelter in hedgerows and leaf litter during daylight hours, and traveling relatively long distances during the night in search of food. They use a tongue-like appendage containing over 27,000 barb-like teeth - the *radula* - to shred their food into uniformly sized pieces, increasing the surface area for enzymatic degradation in the gut. These slugs feed actively down to temperatures approaching 0°C, and adults and eggs have been observed to survive freezing at -3°C for 3 days or more (Slotsbo *et al.*, 2011). It is therefore believed that slugs survive seasonal weather either by preservation of

buried eggs or through migration to areas unaffected by frosts, such as deep in compost heaps and underground in leaf litter. Slugs are also known to be resistant to high concentrations of toxic metals, and are often used in studies of the environmental effects of pollution (Ireland, 1979, Seric Jelaska *et al.*, 2014). The ability to consume a wide range of food types, and their physiological robustness to environmental challenges are thought to be amongst the many reasons slugs are such a successful group of organisms, despite the best efforts of humans to eradicate them from agricultural and suburban land.

The *Arion ater* species is commonly named "the common black slug" but, although a high proportion of its members are black, numerous colour and pattern variants exist (Kennedy, 1959). Taxonomic classification of slugs is difficult, and there are suggestions that species such as *A. ater* and *A. rufus* could be classified as either a single species or be rearranged as subspecies (Cain and Williamson, 1958, Evans, 1986). However, a recent study has observed notable genetic differences between many species and estimates the number of distinct species across Britain and Ireland to be 44, an increase from previous estimates of 36 species in that region (Rowson *et al.*, 2014). This study also estimates that there may be as many as 11 haplotypes of *A. ater*, and provides evidence of interspecific hybridization within the large Arion species (Family *Arionidae*). The lack of a coherent species classification for even these common species is indicative of how much remains to be understood about some mollusc families.

It is now well established that the gut microbiome plays a pivotal role in digestion in many invertebrates and vertebrates such as termites (Brune, 2014), cockroaches(Bertino-Grimaldi *et al.*, 2013), cattle (Hess *et al.*, 2011) and humans (Qin *et al.*, 2010). The gut microbiomes of members of the gastropod class are still

largely unstudied, despite their ability to digest a wide range of materials efficiently. One recent study has demonstrated the ecological richness of the gut microbiome of the gastropod *Achatina fulica* (giant snail), highlighting its metabolic capabilities, with greater than 2,700 plant cell wall degrading CAZymes being observed (Cardoso *et al.*, 2012a). In a previous study we demonstrated that the gut microbial consortium of *A. ater* is directly involved in breakdown of the lignocellulose portion of its diet (Joynson *et al.*, 2014), while showing that this activity is stable at a broad range of temperatures and pH levels, which leads us to believe that the gut environment of *A. ater* could harbour microbial consortia of considerable ecological and economic importance, with specific respect in relation to the degradation of lignocellulose.

In this thesis we present a comprehensive examination of the lignocellulolytic capability of the *A. ater* gut environment, with a focus on the microbial population, using extensive biochemical and molecular methods along with state of the art sequencing techniques and bioinformatics analyses. This research has allowed us to characterize the composition of gut microbial consortium in *A. ater*, and their metabolic capability. There are three reasons why this research is of particular importance. First, the gut of *A. ater* is an understudied environment that we have shown to be a rich reservoir of previously unobserved enzymes from groups, such as glycoside hydrolases, that are of great biotechnological importance to industries such as the bioethanol industry. Second, it may offer insight into the survivability and feeding ability of slug species. This is now especially important following the European Union ban on traditional molluscicide pellets, in force from September 2014 (Commission Implementing Regulation 187/2014), which was introduced because of the rapid build-up of traditional molluscicide metabolites in

water sources (Kay and Grayson, 2013). Finally, the microbiological profile of the slug gut may also provide a target for future bacterial crop pathogen diagnostics, tracking, and control measures in agriculture, as slugs have recently been proposed as vectors for the transmission of bacterial pathogens (Gismervik *et al.*, 2014).

# Chapter 2. Characterization of Cellulolytic Activity In The Gut of The Terrestrial Land slug *Arion ater* Using Gel Zymography

## 2.1 Abstract

Many studies to identify novel, highly cellulolytic environments have focused on analysis of soils and the herbivore gut, with the vast majority of gut based studies focusing on insect groups or ruminants. Initial success with studies of termites and cattle has meant that there have been few studies of other plant eating groups of organisms such as gastropods. Consequently, even the most common of the gastropods, such as *Arion ater*, has received little attention. In light of this we have further characterized the cellulolytic activity in different areas of the gut of the terrestrial slug *Arion ater* in order to assess its viability as a target environment for future CAZyme bioprospecting studies. To do this, slugs were dissected and initial identifications confirmed through observation of internal anatomy. Whole gut tracts were excised, from which crude gut proteins were isolated and tested for the presence of β-1,4-endoglucanase and β-glucosidase enzymes using in gel carboxymethyl cellulose (CMC) zymography and esculin hydrate-ferric ammonium citrate activity gel assays. Zymograms and activity gel studies revealed multiple endoglucanase and β-glucosidase enzymes present throughout the length of the gut. This indicated that the *A. ater* gut environment harbours enzymes that have the ability to degrade long chain cellulose and also to breakdown the resulting oligosaccharides into simple sugars. These findings show that enzymes of interest to the biofuel industry are present, but further study into their origin is required in order to tailor future in depth bioprospecting methods.

## 2.2 Introduction

There is considerable interest in the identification of novel CAZymes that exist in nature. In order to identify targets for in depth study using methods such as next generation sequencing and proteomics the gut environment can be screened using biochemical methods to ensure the activities sought after are present to a degree that call for further study (Oppert *et al.*, 2010). Many studies have characterized the gut fluid cellulolytic activity of insects, including termites (Tokuda and Watanabe, 2007), grasshoppers (Willis *et al.*, 2010a) cockroaches (Gijzen *et al.*, 1994) along with locusts and many beetle species (Cazemier *et al.*, 1997) in order to assess their ability to degrade plant biomass. Yet while many insects have now been studied, the numbers of gastropod species that have been studied in detail are much fewer, mainly due to the initial successes seen with insect species. A study by (James *et al.*, 1997) observed the cellulolytic properties of the gut fluids of *A. ater* and the banana slug, *Ariolimax columbianus,* of American origin, noting high levels of cellulolytic activity for both. A study was conducted to characterize the cellulolytic activity of *A. ater* of English origin (Joynson *et al.*, 2014), that successfully identified high levels of cellulolytic activity along with characterizing the pH and temperature stability profile for the crude extracts using crude protein spectrophotometry based biochemical methods. These studies confirm that there is cellulolytic activity within the *A. ater* gut. However further characterization is necessary, firstly identify some of the enzyme classes present and also to identify the origin of the activity. To gain a better understanding of the enzymes present in the gut we carried out acrylamide gel based studies to specifically confirm the presence of both β-1,4-endoglucanase enzymes that breakdown long chain cellulose and β-glucosidase enzymes that breakdown

cellobiose into glucose. This study will confirm whether the *A. ater* gut environment harbours the enzymes required for cellulose degradation to simple sugars, which will indicate its potential as a target for bioprospecting for enzymes of use in the biofuel industry.

## 2.3 Materials and Methods

### 2.3.1 Slug Collection and Dissection

Slugs were collected from a suburban area in North Cheshire (53.391463 N, 2.211214 W) 2 hours after last light when they emerged from hedgerows to feed. Individuals were allowed to feed on celery/lettuce cores for 12 hours. Slugs were then placed into sterile petri dishes before being cooled at 4 °C for 30 minutes prior to dissection. This reduces motility and reduces spontaneous production of thick mucus during dissection, which can hinder the separation of the internal components. Slugs were then cut with a scalpel across the "foot" from mouth to tail. Internal organs were then allowed to rest on the petri dish floor while the outer tissue was cut away. Whole gut tracts were then removed from mouth to anus, avoiding rupture that would result in loss and contamination of gut juices. Mucus that might interfere with the assays was removed by blotting. Total guts were further separated into 'crop', which denotes the region from the mouth up to and including the stomach, and the 'gut', which corresponds to the digestive tract after the stomach up to the rectum (Figure 2-2). After division into the crop and gut, each section was transferred to separate petri dishes to avoid mixing of gut juices from each section. Gut and crop samples were cut into small pieces using a scalpel in a petri dish and then homogenised with a sterile glass rod in 1.5 mL tubes containing 200 µL of 0.2 M sodium acetate buffer (pH 5.2) followed by

51

vigorous vortexing. Any liquid from the gut pieces that was left in the petri dishes was collected using a pipette and placed into the 1.5mL tube. To clear tissue debris and food matter, samples were centrifuged at 13.3 Krpm for 5 minutes. Supernatants were then aspirated, pooled (subsequently referred to as crop/gut 'crude protein samples') and stored at -80 °C if not used immediately. Protein content of each of the crude samples was estimated using a standard Bradford assay (Bradford, 1976) using bovine serum albumin (BSA) (Sigma-Aldrich, UK) to construct the standard curve.

## 2.3.2  Identification of Endoglucanases Using CMC SDS PAGE Zymography

To detect proteins in the gut fluids that exhibit β-1,4-Endoglucanase activity,  SDS PAGE zymography was carried out following the procedure of Schwarz *et al.* (1987) and Willis *et al.* (2010a) using carboxymethyl cellulose (CMC) as a substrate. Crude gut proteins were separated using a 12% acrylamide SDS PAGE gel containing 0.2% CMC. Before polymerisation of the resolving gel was induced, solutions were heated to 30 °C and CMC was added slowly to the resolving gel mixture in order to dissolve the substrate and prevent aggregation. After the CMC had dissolved polymerisation was induced by adding APS and TEMED. Components of these gels and volumes are shown in Table 2-1.

| Component | Volume/weight |
|---|---|
| **Deionised water** | 4.9 mL |
| **30% Acrylamide solution** (Sigma-Aldrich, UK) | 6.0 mL |
| **1.5 M Tris (pH 8.8)** (Fisher Scientific, UK) | 3.8 mL |
| **10% SDS** (Fisher Scientific, UK) | 0.15 mL |
| **Carboxymethyl cellulose (CMC)** (Sigma-Aldrich, UK) | 0.03 g |
| **10% Ammonium persulphate** (Fisher Scientific, UK) | 0.15 mL |
| **TEMED** (Sigma-Aldrich, UK) | 0.006 mL |

**Table 2-1 The components and volumes required to create the resolving gels for 2 12% acrylamide, 0.2% CMC SDS PAGE mini gels**

Gels were overlaid with deionised water and allowed to polymerize for 2 hours.

The stacking gel solution was then created using the components in Table 2-2.

Gels were used the same day to prevent CMC degradation.

| Component | Volume |
|---|---|
| **Deionised water** | 4.1 mL |
| **30% Acrylamide solution** (Sigma-Aldrich, UK) | 1.0 mL |
| **1.0 M Tris (pH 6.8)** (Fisher Scientific, UK) | 0.75 mL |
| **10% SDS** (Fisher Scientific, UK) | 0.06 mL |
| **10% Ammonium persulphate** (Fisher Scientific, UK) | 0.06 mL |
| **TEMED** (Sigma-Aldrich, UK) | 0.006 mL |

**Table 2-2 The components and volumes required to create a 6% stacking gel for 2 SDS PAGE mini gels**

The inner and outer chambers of the SDS PAGE tank were filled with 1x SDS PAGE running buffer (25 mM Tris, 192 mM Glycine, 0.1% SDS, pH 8.3) a Crop and gut crude protein samples were thawed on ice . A modified 4x Laemmli loading buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 0.01% bromophenol blue (Laemmli, 1970)) was added to 50 µg of each crude protein sample. As a positive control, 2 µg of *Aspergillus niger* cellulase (Sigma-Aldrich, UK) was ran alongside the crude gut proteins. All samples were heated to 80 °C for 10 minutes followed by pulse centrifugation to partially denature the proteins and thus increase separation and limit substrate digestion during electrophoresis. The crude protein sample-loading buffer mixes and positive control were then loaded on duplicate gels along with 15 µL of SeeBlue® Plus2 Pre-Stained Standard (Invitrogen, UK). The gel tank was placed into an ice bath and gels were then ran at a constant 100 V for 4 hours 30 minutes at 4°C to reduce cellulose breakdown by enzymes as they migrate through the gel.

For size estimation and total protein visualisation, the duplicate gel was stained in Coomassie Brilliant Blue stain (0.1% Coomassie Brilliant Blue (Bio-Rad, UK), 50% Methanol [v/v], 10% acetic acid [v/v], 40% deionised water [v/v]) for 1 hour. The gel was then destained using a 40% methanol [v/v] 10% acetic acid [v/v] 50% water [v/v] destaining solution until the gel background was clear, refreshing the destain every 30 minutes. Prior to incubation and staining of the zymogram gel, the distances travelled by the pre-stained standard bands were measured as this became difficult to visualise after treatment. The zymogram gel was washed in a 5% [v/v] Tritron X-100 (Sigma-Adrich, UK) 95% water [v/v] solution for 30 minutes (repeated 5 times) to remove SDS. The gel was then rinsed with deionised water, placed in sodium phosphate buffer (50 mM, pH 6.5) and incubated for 2 hours at 4

°C to exchange the buffer system and to allow renaturation of the proteins in the gel. The phosphate buffer was then refreshed and the gel was then incubated at 37 °C overnight. Following incubation, the gel was stained with 0.1% (w/v) Congo Red solution in water for 1 hour. The gel was rinsed with deionised water then destained with a 1 M sodium chloride solution for 3 hours, refreshing the destain solution every hour. To enhance visualisation of clear zones, acetic acid was added drop wise to the 1 M sodium chloride solution containing the gel, until the pH change causes the Congo Red stain to change from a red colour to deep purple.

### 2.3.3 Identification of β-glucosidase Enzymes Using Esculin Hydrate – Ferric Ammonium Citrate Native PAGE Activity Gel Assays

In order to detect β-glucosidase enzymes in the crude gut and crop protein samples, native PAGE activity gel assays were carried out as described by Kwon *et al.* (1994) using 8% and 12% acrylamide native Tris-glycine gels. In this assay, proteins were separated on a native PAGE gel and activity visualised by incubating the gel in an solution containing a celliobiose mimic, esculin hydrate, and ferric ammonium citrate.

The components and volumes required to create the 8% and 12% native PAGE gels are shown in Table 2-3.

| Component | 8% Native resolving gel | 12% Native resolving gel |
|---|---|---|
| Deionised water | 7.0 mL | 5.0 mL |
| 30% Acrylamide solution (Fisher Scientific, UK) | 4.0 mL | 6 mL |
| 1.5 M Tris (pH 7.8) (Fisher Scientific, UK) | 3.8 mL | 3.8 mL |
| 10% Ammonium persulphate (Fisher Scientific, UK) | 0.15 mL | 0.15 mL |
| TEMED (Sigma-Aldrich, UK) | 0.009 mL | 0.006 mL |

**Table 2-3 The components and volumes required to create the resolving gels for 2 8%/12% native PAGE mini gels**

After pouring the resolving gels were over laid with deionised water and left to polymerize for 2 hours. The stacking gel solution was then created as seen in Table 2-4.

| Component | 6% Native stacking gel |
|---|---|
| Deionised water | 4.1 mL |
| 30% Acrylamide solution (Fisher Scientific, UK) | 1.0 mL |
| 1.0 M Tris (pH 6.8) (Fisher Scientific, UK) | 0.75 mL |
| 10% Ammonium persulphate (Fisher Scientific, UK) | 0.06 mL |
| TEMED (Sigma-Aldrich, UK) | 0.006 mL |

**Table 2-4 The components and volumes required to create a 6% stacking gel solution for 2 native PAGE mini gels**

The inner chamber of the SDS PAGE tank (cathode) was then filled with the "cathode running buffer" (53 mM Tris (Fisher Scientific, UK), 68 mM glycine (Fisher Scientific, UK), pH 8.9) and the outer chamber filled with "anode running buffer" (0.1 M Tris (Fisher Scientific, UK), pH 7.8).

Crop and gut crude protein samples were thawed on ice. A native 4x Laemmli loading buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 0.01% bromophenol blue (Laemmli, 1970)) was added to between 10 µg and 50 µg of each crude protein sample. Samples were then loaded on duplicate gels. Gels were ran at 100 V for 4 hours at room temperature. One of the gels was then soaked in a 0.2 M sodium acetate (Sigma-Aldrich, UK) buffer in water (pH 5.5) for 10 minutes to exchange the buffer system and then placed in a fresh 0.2 M sodium acetate buffer (pH 5.5) containing 0.1% (w/v) esculin hydrate (Sigma-Aldrich, UK) and 0.03% (w/v) ferric ammonium citrate (Sigma-Aldrich, UK) in water and incubated for 3 hours- overnight at 37 °C to allow in gel hydrolytic activity. Where β-glucosidase enzymes are present esculin is cleaved producing esculitin which then reacts with ferric iron present in the ferric ammonium citrate, resulting in the formation of a black precipitate. To stop the reaction, the gel was placed into a 10% glucose solution, preventing further colour development through end product enzyme inhibition of β-glucosidase enzymes. For total protein visualisation, the duplicate gel was stained in Coomassie Brilliant Blue stain (0.1% Coomassie Brilliant Blue (Bio-Rad, UK), 50% Methanol [v/v], 10% acetic acid [v/v] (Sigma-Aldrich, UK), 40% deionised water [v/v]) for 1 hour. The gel was then destained using a 40% methanol [v/v] 10% acetic acid [v/v] 50% water [v/v] destaining solution until the gel background was clear, refreshing the destain every 30 minutes.

## 2.4  Results

### 2.4.1  *A. ater* Dissection and Whole Gut Extraction

*A. ater* whole gut tracts were removed by dissection. Through dissection we also confirmed initial taxonomic identification from external feature of the slug by observation of the morphology of the internal organs, including the organisation of the digestive system and the genitals (Figure 2-1). The entire gut was carefully unwound from the digestive cecum and surrounding organs then lay out in order to identify "crop" and "gut" regions clearly (Figure 2-2). The crop region was initially obscured by the digestive cecum and was thicker than the gut region. In many cases the stomach, at the interface between the crop and the gut was inflated with liquid produced by the digestive gland. The gut wall appeared transparent, with individual shredded food particles visible in the crop section.
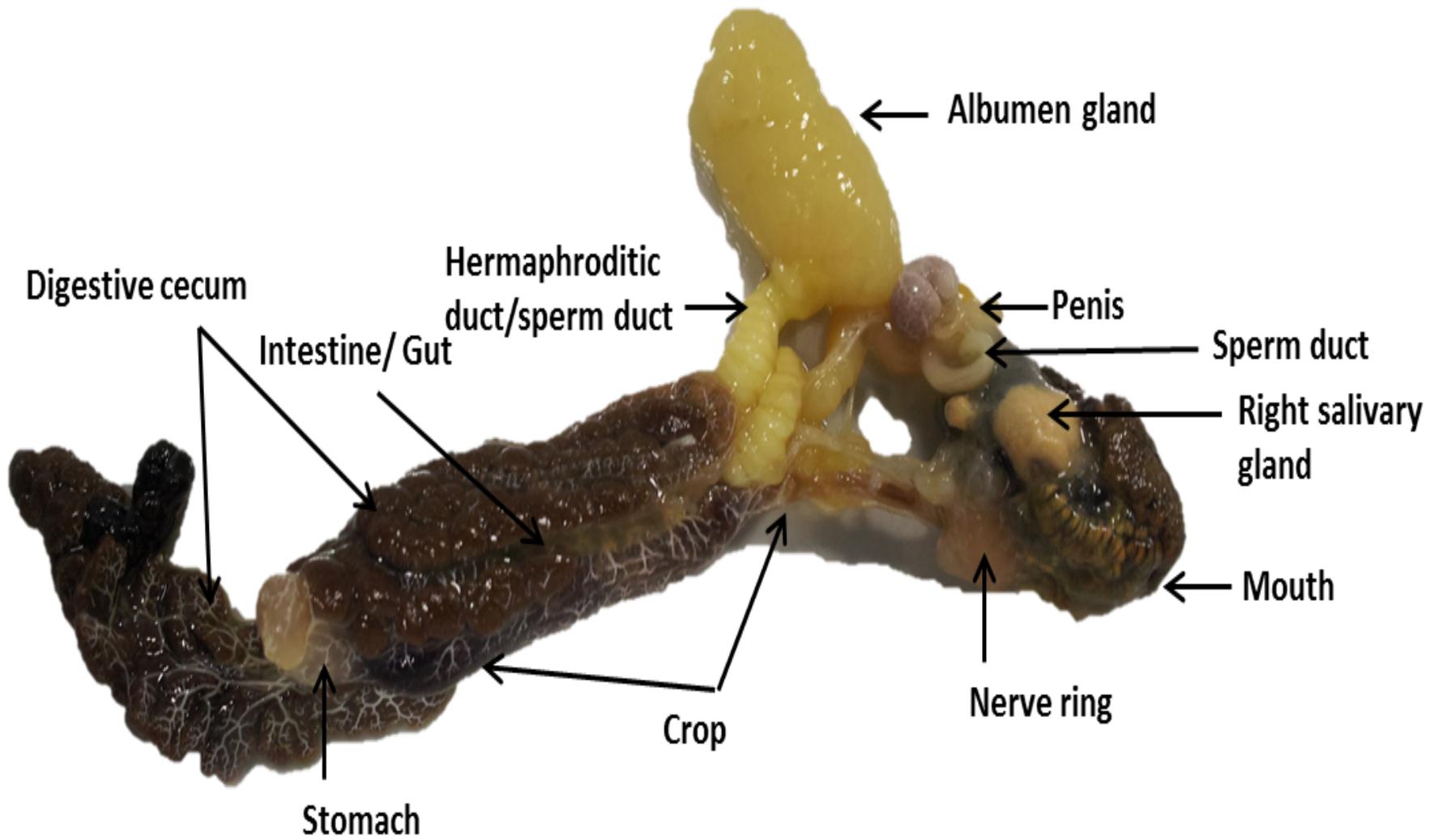
**Figure 2-1 The internal anatomy of the black slug *Arion ater***
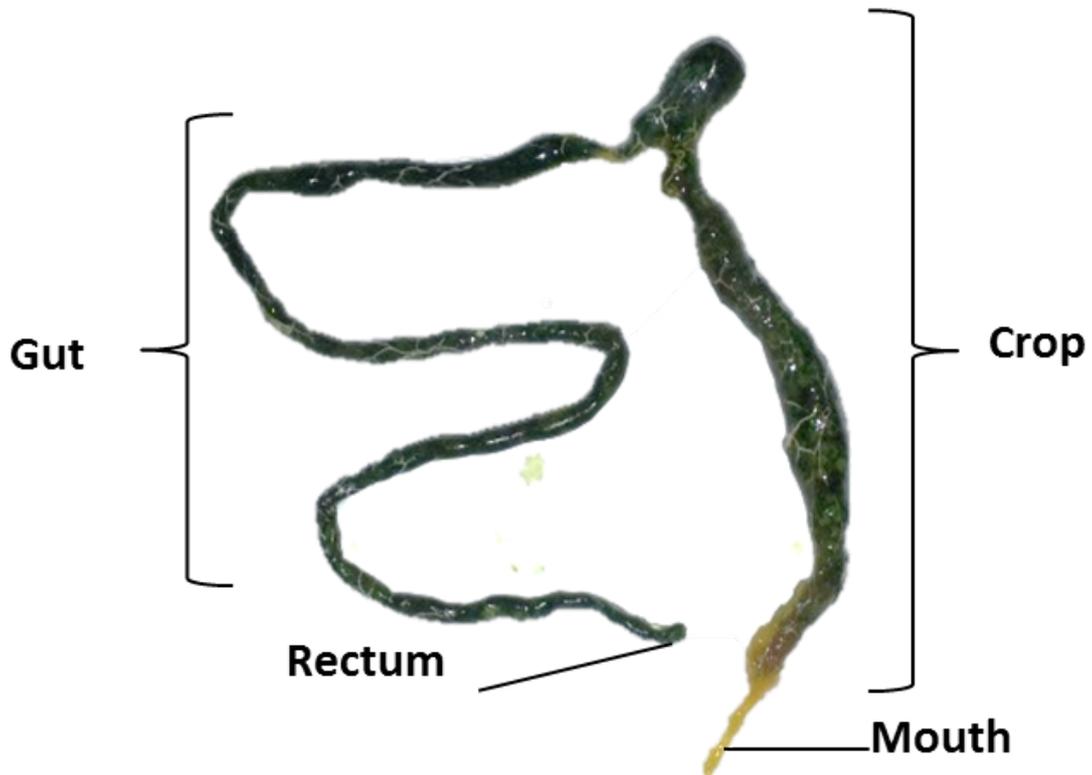
**Figure 2-2 The dissected whole gut tract of the black slug, *A. ater***
*Figure oriented by labels for mouth and rectum and showing the designated areas included in the "crop" and "gut" samples*

Figure 2-2 shows the entire gut tract of *A. ater*. The "crop" samples were designated as all sections of the gut between the nerve ring (Figure 2-1) up to and including the enlarged stomach region, and "gut" samples included the intestine from the stomach intestine interface to the rectum/anus.

### 2.4.2  Identification of Endoglucanases Using CMC SDS PAGE Zymography

To gain an insight into the cellulolytic activity previously detected in the *A. ater* gut juices, crude protein extracts from the crop and gut portions of the digestive tract were separated using 12% SDS PAGE gels containing 0.2% CMC followed by renaturing of the proteins and staining to detect protein bands that exhibit β-1,4-endoglucanase activity.
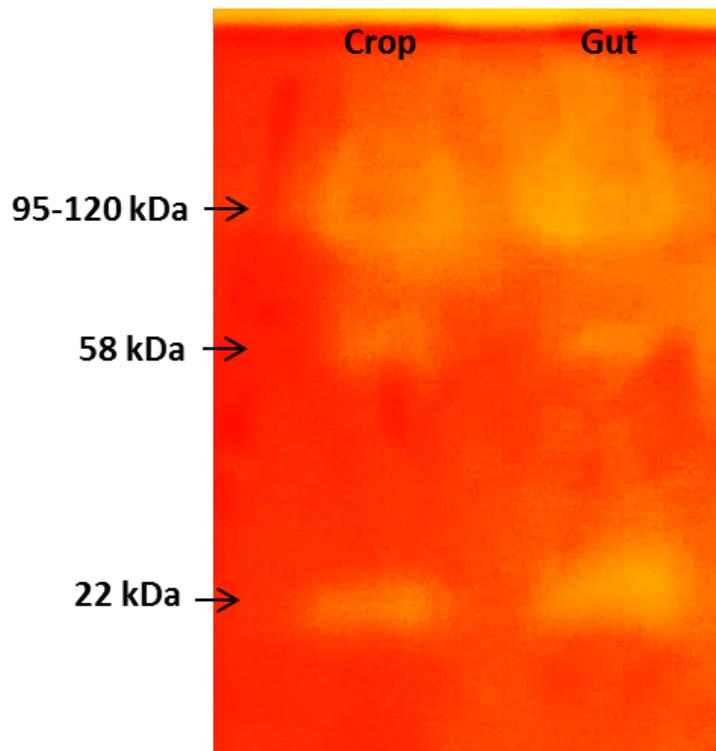
**Figure 2-3 Identification of cellulolytic proteins using CMC SDS PAGE zymography**
*A 12% SDS PAGE zymograom showing identification of ß1-4-endoglucanase activity in 50µg of Crop and Gut crude protein extracts after electrophoretic separation at room temperature.*

Figure 2-3 shows a zymogram stained with Congo Red followed by destaining with NaCl, clear zones (yellow areas) in the gel image indicate areas in which CMC has been degraded, allowing the Congo Red stain to be washed away. There are three main areas of activity in both the crop and gut samples at 95-120 kDa, 58 kDa and 22 kDa. The gel shows smearing in the high molecular weight region in both lanes, suggesting CMC degradation during migration. To limit this and to improve band resolution, the procedure was repeated with the temperature control measures described in section 2.3.2 (Figure 2-4).
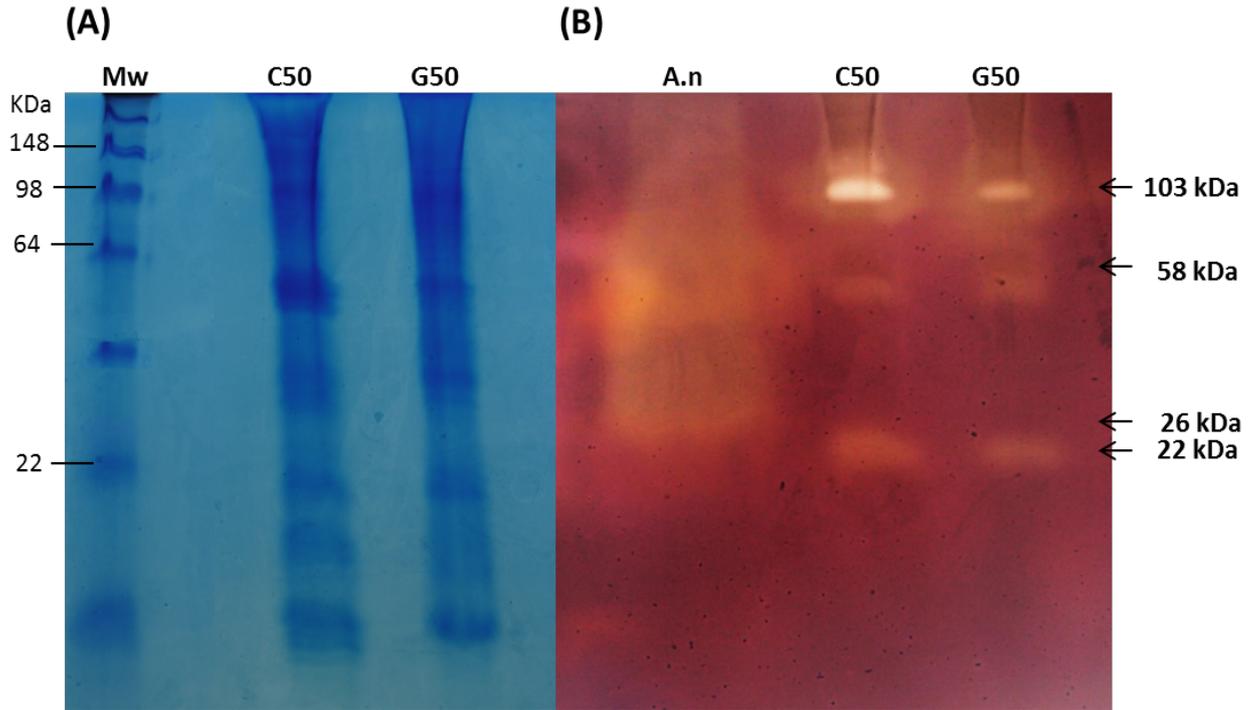
**Figure 2-4 Identification of cellulolytic proteins using temperature controlled CMC SDS PAGE zymography**
*(A) A 12% SDS PAGE coomassie stained gel showing separation of 50 µg crop and gut crude protein extracts and (B) a 12% SDS PAGE CMC zymorgam that has been subjected to endoglucanse activity staining showing separation of 50 µg crop and gut crude protein extracts and a positive control A. niger cellulose shown in gel B. Both gels were ran at 4 °C*

Figure 2-4 shows duplicate sample runs of 50 µg of both crop and gut crude protein samples. Gel (A) shows the total proteins separated after Coomassie Brilliant Blue staining. Gel (B) shows the zymogram after staining/destaining. Visualization of bands on the zymogram was enhanced by acidifying the gel with acetic acid in a NaCl solution post destaining. Lanes C50 and G50 on both gels show separation of crop and gut samples respectively. The same three zones of activity are visible with greater resolution than in Figure 2-3, with activity observed at ~103 kDa, ~58 kDa and ~22 kDA. Lane A.n shows the migration of 2 µg of commercial *Apergillus niger* cellulase (A 26 kDa protein) smearing can be seen in this lane indicating CMC degradation during migration.

## 2.4.3 Identification of β-glucosidase Enzymes Using Esculin Hydrate – Ferric ammonium Citrate Native PAGE Activity Gel Assays

Post electrophoretic activity staining was conducted to identify the presence of β-glucosidase enzymes that breakdown the cellulose degradation product cellobiose into glucose. Figure 2-5 shows the results of an initial test where we identified the appropriate protein amount to run for adequate activity levels for visualisation.
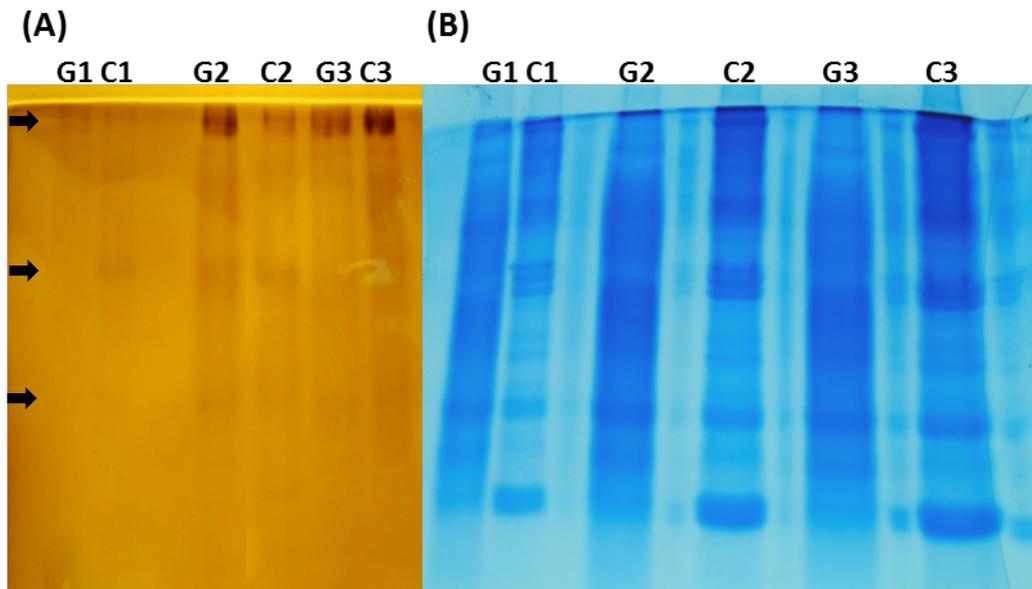


**Figure 2-5 Identification of the presence of β-glucosidase activity in crude gut protein samples using 8% native PAGE esculin hydrate activity gels**
*(A) An 8% native PAGE gel showing separation of 10, 30 and 50μg of crude protein samples after activity staining using an esculin hydrate, ferric ammonium citrate activity buffer. (B) A duplicate gel showing total protein separation of 10, 30 and 50μg of crude protein samples stained with Coomassie Brilliant Blue.*

Figure 2-5 (A) and (B) show separation of 10 μg (G1 and C1), 30 μg (G2 and C2) and 50 μg (G3 and C3) of gut and crop crude protein samples. Gel (A) shows the activity stained gel, where multiple areas of activity were visible (annotated with black arrows). The strongest bands of activity are located in the high molecular weight region for each sample. Gel (A) shows that 10 μg of crude samples loaded appears to be at the lower limit of detection, with 30 μg an 50 μg both showing

activity that may be visualised clearly. Gel (B) shows the total protein ran on the gel, significant streaking can be seen in both samples, with the gut sample most affected. Thus only fresh (unfrozen) crude protein extracts were used for further tests.
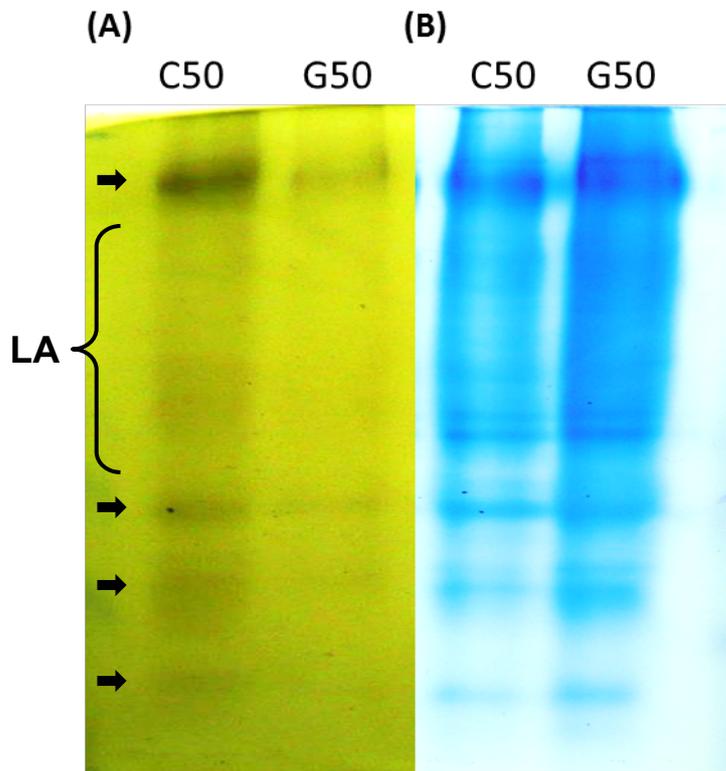


**Figure 2-6 Identification of the presence of β-glucosidase activity in unfrozen crude gut protein samples using 8% native PAGE esculin hydrate activity gels**
*(A) An 8% native PAGE gel showing separation of 50μg of freshly (unfrozen) crude protein samples after activity staining using an esculin hydrate, ferric ammonium citrate activity buffer. (B) A duplicate gel showing total protein separation of 50μg crude protein samples stained with Coomassie Brilliant Blue*

Figure 2-6 (A) shows at least 4 zones of activity for both crop (C50) and gut (G50) samples (indicated by arrows). Low levels of activity are also seen in a large proportion of the crop sample lane (labelled LA). The use of fresh gut protein samples resulted in a greater number of zones of activity. Most of the zones of activity are seen in the lower molecular size range, thus a further test was carried

out to increase separation of these bands using a 12% acrylamide gel (Figure 2-7).
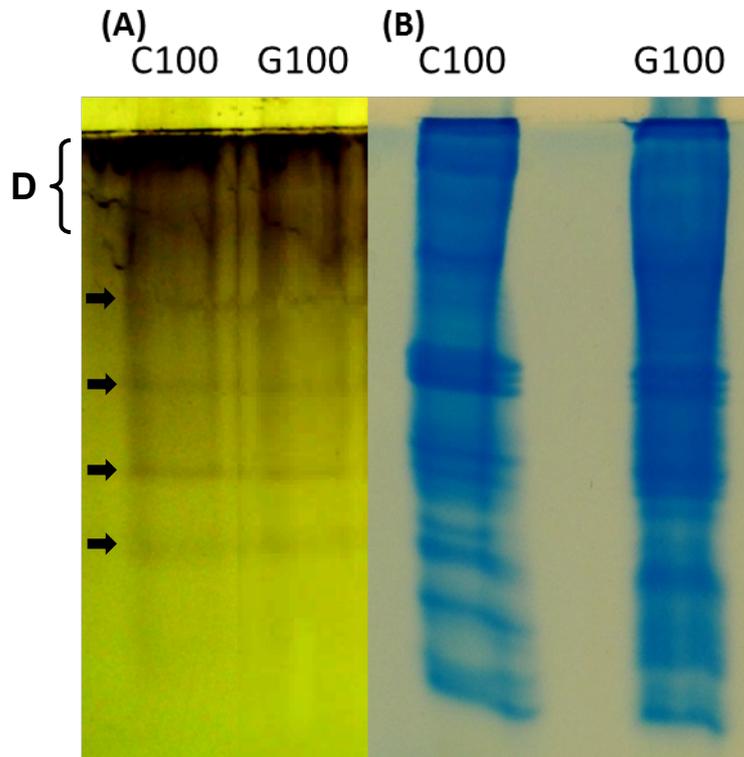


**Figure 2-7 Identification of the presence of β-glucosidase activity in unfrozen crude gut protein samples using 12% native PAGE esculin hydrate activity gels**
*(A) A 12% native PAGE gel showing separation of 100µg of freshly prepared (unfrozen) crude protein samples after activity staining using an esculin hydrate, ferric ammonium citrate activity buffer. (B) A duplicate gel showing total protein separation of 100µg crude protein samples stained with Coomassie Brilliant Blue*

Figure 2-7 (A) shows activity staining which reveals 4 distinct bands in gel (indicated by black arrows). The dark zone labelled D indicates high levels of activity in the high molecular weight region, that could be caused by large proteins, aggreagates or cell surface bound constructs. Utilization of a 12% gel increased separation of the lower bands and increasing the loading amount to 100 µg increased activity in these regions, leading to better visualization.

## 2.5 Discussion

In this study we observed multiple β-1,4-endoglucanase and β-glucosidase enzymes in the crude gut protein extracts from both the gut and crop portions of the *A. ater* digestive tract. This indicates that the gut environment harbours enzymes with the ability to break down cellulose into its simple sugar components (glucose) completely using enzymes of either microbial or animal origin. Each enzyme was observed using polyacrylamide gel based biochemical assays to further characterize the cellulolytic activity that was detected and quantified by Joynson *et al.* (2014).

The study into endocellulase activity using modified SDS PAGE allowed successful identification of multiple β-1,4-endoglucanase enzymes. Using carboxymethyl cellulose we observed three endoglucanase enzymes in the crude gut protein extractions (Figure 2-3, Figure 2-4), indicated by three resolute zones of clearance seen at corresponding migratory distances in both the crop and gut samples. This could indicate a lack of compartmentalisation of the gut tract, meaning that the same cellulolytic enzymes are active throughout the entire digestive tract. Similar results were observed in a study by Willis *et al.* (2010a) which used CMC zymography to observe the endoglucanase activity of crude gut fluids of the grasshopper *Dissosteira carolina,* after extraction of the gut and separation of the fore, mid and hindgut regions. They identified activity at four regions in the gel that were identical in the fore and mid gut sections but with almost complete loss of activity in the hind gut section. With the exception of the difference in the hind gut region (where enzymes appear to denature), as in our study, there were no detectable changes in observable endoglucanase enzymes

between gut regions. This suggests that it is not uncommon for a similar repertoire of digestive proteins to be present throughout the entire digestive tract.

During this study multiple β-glucosidase enzymes were also observed. To visualise these enzymes, post electrophoresis esculin hydrate –ferric ammonium citrate in gel activity assays were carried out. Initial tests using frozen protein samples were relatively unsuccessful, showing low levels of activity largely in the high molecular weight region (Figure 2-5). In subsequent studies only fresh protein extracts were used, which resulted in observation of a higher number of areas of activity with greater levels of activity seen at each location in general (Figure 2-6Figure 2-7). Figure 2-6 shows a distinct clear zone of activity close to the top of the gel, indicating that the protein is large, cell surface bound or have formed aggregates. Along with this large protein, three other resolved areas of activity can be seen further down the gel, suggesting that they are much smaller, although size determination using native PAGE is problematic due to factors such as protein folding and conformation. To increase the resolution of these smaller proteins, crude gut samples were then separated on 12% acrylamide gels (Figure 2-7). This alteration allowed identification of four areas of activity that were highly resolved in the lower molecular weight region. This did however result in a large area of activity at the very top of the gel at the stacking gel- resolving gel interface where the larger enzymes, observed clearly in Figure 2-6, where unable to migrate into the 12% gel due to the lower pore size.

The minimum detectable amount of active enzyme in the esculin hydrate activity gel assay was relatively high at >10 ng (Kwon *et al.*, 1994), which means that this type of study is limited to detection of proteins expressed at high levels by the slug itself or by dominant species of gut bacteria.

Although the enzymes detected could be of microbial origin, the endocellulases observed in the study by Willis *et al.* (2010a) were subject to N-terminal protein sequencing and identified as of animal origin. However a study by (Tokuda and Watanabe (2007)), of the cellulolytic activity of termite species *Nasutitermes takasagoensis* and *Nasutitermes walkeri*, showed that the cellulolytic proteins (observed using zymograms) were of microbial origin. This was achieved by showing a significant reduction in cellulolytic activity after antibiotic feeding.

In this opening study we have identified the presence of enzymes that are capable of total degradation of cellulose in the gut of *A. ater*. This provides good evidence that the environment may be of interest as a source of novel CAZymes. However, in order to ascertain the methods that should be implemented in the next steps, identification of the origin of the observed cellulolytic activity is required. Although some herbivorous insects possess genes encoding some plant cell wall degrading enzymes, such as higher termites (Watanabe *et al.*, 1998), the structural complexity of the plant cell wall superstructure requires a large number of enzyme classes for complete degradation, some of which have only been observed of microbial origin, suggesting a general reliance of herbivores on symbiont microbes for complete degradation of lignocellulose. It is also becoming clear that the interactions of host and microbe has had a direct impact on the evolutionary transitions in diet in many herbivorous eukaryotes, including insects (Hansen and Moran, 2014). The microbial contribution to the breakdown of lignocellulose in the gut has been observed in insects (Mikaelyan *et al.*, 2014, Cruden and Markovetz, 1979, Shi *et al.*, 2013), ruminants (Flint *et al.*, 2008), fish (Li *et al.*, 2014) and also in gastropods (Cardoso *et al.*, 2012a). Thus further studies were conducted

focusing on detection and characterization of the microbial contribution to lignocellulose degradation in the gut of *A. ater*.

# Chapter 3. Isolation and Identification of *A. ater* Gut Microorganisms: Identification of The Source of Gut Cellulolytic Activity

## 3.1 Abstract

Since the development of modern high throughput sequencing methods it is becoming ever clearer that microbial gut symbionts have had a great impact on the evolution of plant utilization by eukaryotes. In recent years genomic studies have elucidated the roles of microbes in the breakdown of each component of the plant cell wall within the gut environments of various herbivores. We seek to ascertain whether the microbial consortium in the gut of *A. ater* has the potential to contribute to the breakdown of lignocellulose and to discover if the cellulolytic activity previously observed in the gut of *A. ater* is of microbial origin. To do this, microbes were isolated and cultured from the gut. Each isolate was tested for endoglucanase and β-glucosidase activity using growth plate activity assays. 12 cellulolytic microbes were isolated and identified by amplification and sequencing of their 16s rRNA gene. These included members of the genera *Buttiauxella, Enterobacter, Citrobacter, Serratia* and *Klebsiella*. To further characterize the gut microbial population, metagenomic DNA was extracted and subjected to 16s rRNA gene amplification targeting a 400 bp region. The amplified genes from the multiple microbial genomes present were subsequently separated by sequence variation using differential gradient gel electrophoresis (DGGE). These separated bands were then isolated and sequenced. This identified members of the genera *Citrobacter, Serratia, Pectobacterium, Acinetobacter, Mycoplasma, Pantoea* and *Erwinia*. Identification of multiple cellulolytic bacteria, some of which could not be

accurately identified, indicates that the gut microbiome of *A. ater* contributes to the breakdown of plant biomass in the gut and also suggests that the microbiome contains undescribed cellulolytic systems that could be exploited to improve commercial degradation of lignocellulose.

## 3.2 Introduction

Bacteria are the most numerous and one of the oldest groups of organisms on the planet. Over billions of years the catabolic diversity of these microorganisms has become extremely broad, with species being able utilize a multitude of different organic compounds as primary sources of carbon for energy production. These include highly specialized microbes that are able to utilize only a few substrates and others that use over 100 organic compounds as sources of carbon and energy (Lynd *et al.*, 2002). In order to use multiple energy sources, microbes have developed a large metabolic repertoire that facilitates the breakdown of a wide range of organic compounds. Lignocellulose, being the most abundant source of continually regenerating organic carbon on the planet, is a common energy source for microbial life and its degradation by microbes is a key step in the carbon cycle (Zeikus, 1981, Štursová *et al.*, 2012). This degradation occurs both in the soil and in the guts of higher organisms. One of the best studied herbivore gut environments is that of the cow rumen, which contains a complex mixture of anaerobic bacteria, of which 10% are thought to be cellulolytic (Russell *et al.*, 2009b). Some early research studies isolated microbes from the cow rumen and tested them for cellulolytic activity using growth plate assays (Teather and Wood, 1982). Identification of the presence of cellulolytic microbes led to the use of modern high throughput sequencing to assess the cow rumen microbiome in greater detail. This allowed identified of 27,755 putative carbohydrate active genes along with a large amount of data on other metabolic processes (Hess *et al.*, 2011). Similar research was conducted on the termite gut, where cellulolytic microbes were initially cultured using traditional microbiology methods (Wenzel *et*

*al.*, 2002) and then using modern sequencing methods(Warnecke *et al.*, 2007), again  identifying a great number of CAZyme genes.

Based on the success of these projects, we conducted a study to isolate and culture cellulolytic microbes from the *A. ater* gut in order to ascertain whether cellulolytic symbionts were present in the gut environment. We also carried out a brief study to identify members of the gut microbiome using culture independent methods to gain an insight into microbial ecology of the *A. ater* gut. This enabled us to deduce if the *A. ater* gut would be a viable target environment for metagenomic study using high throughput methods to search for novel carbohydrate active enzymes.

## 3.3 Materials and Methods

### 3.3.1 Identification of Culturable Celluloytic Bacteria

#### 3.3.1.1 Dissection and Gut Microbe Isolation

Adult slugs were collected from the wild in a pesticide free area. Live samples were kept at room temperature and provided with lettuce cores for food overnight. Slugs were dissected as previously described in chapter 2. Briefly, slugs were cooled to 4 °C prior to dissection, and then individuals were placed in clean petri dishes and cut on the foot from mouth to tail using a fresh scalpel each for each sample. The entire gut was extracted and cut into pieces. All gut pieces from a single individual were quickly transferred to sterile 2 mL eppendorf tubes. A ¼ strength Ringer's solution was created by dissolving one ringers solution tablet (Sigma-Aldrich, UK) in 500 mL of distilled water. This isotonic solution was used to prevent rupture of bacterial cells due to osmotic pressue. The solution was autoclaved and left to cool. 500 µL of the ¼ strength solution was added to the Eppendorf tubes contain the gut pieces, the gut was then homogenized with a sterile glass rod to encourage release of microbes from the gut material. A 5-fold serial dilution was then created with 4 more tubes containing 400µl of Ringer's solution (100µl transferred for each dilution step). A single sterile loop (~2 µL) of each Ringer's solution gut dilution was streaked on Luria-Bertani (LB) (Sigma-Aldrich, UK) agar (Sigma-Aldrich, UK) plates containing 0.5% (w/v) carboxymethyl cellulose (CMC) (Sigma-Aldrich, UK) in order to ensure that on at least one plate bacterial concentrations were adequate for individual colony isolation. After 18 hours of incubation at 25°C (approximate temperature of slug gut), 2 replica plates were created using microfiber cloth squares to transfer plated microbes. This was

done to avoid false identification of cellulolytic microbes due to the activity of cellulase enzymes present in the crude gut mixture that was plated. Replica plates were incubated for a further 18 hours at 25°C. One plate was then subjected to Congo Red activity staining (described in 3.3.1.2). Individual colonies were then taken from the non-stained replica plate that corresponded to the zones of clearance seen in the activity stained replica. Isolates were then streaked onto fresh CMC agar plates and incubated at 25°C for 16 hours. A replica of this streak plate was then created and incubated at 25°C for a further 16 hours and the original subjected to Congo Red staining to ensure correct selection of a cellulolytic microbe. A single colony was taken from the replica and again re streaked. This was repeated once more to ensure isolation of a single microbial species from each plate. Frozen stocks were created for each isolated bacterial strain by dispersing a single colony into 5ml of LB broth (Sigma-Aldrich, UK) and incubating overnight at 25°C in a shaking incubator at 250rpm followed by addition of glycerol (15% v/v), aliquots were then immediately frozen in liquid nitrogen and stored at -80°C. In order to ensure isolates had survived freezing and to reconfirm that isolates frozen did indeed have cellulolytic activity, each frozen isolate was plated on another CMC LB agar plate and incubated at 25°C for 24 hours, followed by staining and destaining as described in section 3.3.1.2. For these confirmatory assays, isolates were plated alongside un-transformed Top10 competent *E. coli* (Life Technologies, UK) as a negative control and 2 µL of a 1 mg/mL *Aspergillus niger* cellulase (Sigma-Aldrich, UK) solution as a positive control.

### 3.3.1.2 Carboxymethyl Cellulose (CMC)- Congo Red Growth Plate Activity Assays

In order to determine cellulolytic activity of the microbial isolates, CMC LB agar plates on which isolates were cultured were stained with a 0.1% (w/v) Congo Red solution (Sigma-Aldrich, UK) for one hour. Plates were then destained with a 1 M sodium chloride solution for 3 hours, refreshed at 1 hour intervals. Congo Red binds strongly to intact cellulose chains but can be washed away from areas of the plate in which the cellulose chains have been degraded. Clear zones around areas of microbial growth therefore indicate cellulolytic activity. To enhance visualisation of clear zones, acetic acid was added drop wise to the NaCl destaining solution, which turned the Congo Red stain from red to a deep purple.

### 3.3.1.3 Esculin Hydrate – Ferric Ammonium Citrate Growth Plate Activity Assays

Microbes that exhibited cellulolytic activity using CMC growth assay plates were also tested for β-glucosidase activity using an esculin hydrate – ferric ammonium citrate growth plate assays. Isolates were plated on LB agar plates containing 0.1% (w/v) of the cellobiose mimic esculin hydrate (Sigma Aldrich, UK) and 0.03% (w/v) ferric ammonium citrate (Sigma Aldrich, UK) and grown for between 3 and 16 hours (checked at regular intervals). Appearance of a dark black precipitate around areas of microbial growth indicate cleavage of esculin into esculitin, which then reacts with iron in ferric ammonium citrate forming a black precipitate. Un-transformed top10 competent *E. coli* (Life Technologies, UK) were grown alongside isolates as a negative control.

### 3.3.1.4 Microbial Genomic DNA Extraction

Firstly, isolates were removed from the -80 °C freezer and placed into dry ice to prevent thawing. A sterile loop was then used to scrape a small amount of frozen culture that was used to inoculate 10 mL of LB broth (Sigma-Aldrich, UK) in a

sterile 30 mL universal tube. Isolates, along with a control tube containing only LB broth, were incubated at 25°C overnight for 16 hours. If the broth in the control tube remained free of growth, bacteria from 5 mL of each overnight culture was pelleted by centrifugation at 11,000 x g for 1 minute. Genomic DNA was then extracted from the pellets using the spin column based method of the DNAeasy Blood and tissue kit (Qiagen, UK) following the standard protocol for bacterial DNA extraction. DNA sample concentration and purity was then assessed using a nanodrop spectrophotometer (Thermo Fisher, UK). DNA extracts were subsequently stored at -20°C for future use.

### 3.3.1.5  PCR Amplification of The 16s rRNA Gene

Extracted genomic DNA from each sample was then used as template for PCR targeting the 16S ribosomal RNA (rRNA) gene using a bacterial 16s universal primer set (Table 3-1).

| Name | Forward/Reverse | Sequence |
|---|---|---|
| Weisburg bacterial universal primers (Weisburg *et al.*, 1991) | Forward | 5' AGA GTT TGA TCC TGG CTC- 3' |
| | Reverse | 5' ACG GCT ACC TTG TTA CGA-3' |

**Table 3-1 Primers used to amplify the full 16s rRNA gene for culturable bacteria identification**

Table 3-1 shows the primer sets used to amplify the 16s rRNA gene from each of the isolated microbes. To amplify the 16s rRNA genes a standard Taq based PCR procedure was used. PCR reactions were set up using the reagents and volumes set out in Table 3-2 and the PCR conditions and cycles seen in Table 3-2 The PCR reaction mixture for 16s rRNA gene amplification of cultured isolates.

77

| Component | Amount |
|---|---|
| **DreamTaq mix** (Life Technologies, UK) | 25 µL |
| **Weisberg Forward Primer** | 1 µL(0.2 µM) |
| **Weisberg Reverse Primer** | 1 µL (0.2 µM) |
| **Template DNA** | 10ng |
| **Nuclease-free water** (Fisher Scientific, UK) | Up to 50 µL |
| **Total volume** | 50 µL |

**Table 3-2 The PCR reaction mixture for 16s rRNA gene amplification of cultured isolates**

| Stage | Temperature | Time (MyTaq®) | Cycles |
|---|---|---|---|
| **Initial denaturation** | 95°C | 1 minute | 1 |
| **Denaturation** | 95°C | 30 seconds | |
| **Annealing** | 53°C | 30 seconds | 25 |
| **Elongation** | 72°C | 1 minute | |
| **Final** | 72°C | 1 minute | 1 |

**Table 3-3 16s rRNA PCR cycling stages, temperatures and times**

After thermocycling, PCR products were immediately stored on ice. A 1% agarose gel was created using 0.5X TBE (40 mM Tris-HCl (Fisher Scientific, UK), pH 8.3 45 mM boric acid (Fisher Scientific, UK), 1mM EDTA (Fisher Scientific, UK)) and 1 g of agarose (Bioline, UK). When agarose gel mixture was at approximately 50°C, 4 µL of Gel Red fluorescent DNA stain (10,000x concentrate) (Biotum, USA) was added and the liquid swirled gently. The gel was placed in the gel tank and 10 µL of each PCR product was then loaded alongside a size marker mixture containing 2 µL of HyperLadder I  (Bioline, UK), 2 µL loading buffer (Bioline, UK) and 2 µL of

nuclease free water (total volume 6 µL). Samples were ran at 100v until the coloured markers (Red for the MyTAq samples and blue for the HyperLadder) were 3 cm from the end of the gel (~75 minutes). The gel was then viewed in a G:Box Chemi transilluminator (Syngene, UK) to check for successful amplification and correct sizes of amplified gene fragments by comparison to the sized DNA ladder, images were taken as appropriate.

### 3.3.1.6 Molecular Cloning and Transformation of Amplified 16s rRNA Gene Fragments

In order to isolate the amplified 16s rRNA genes for cloning bands from the agarose gel that corresponded to the 1500bp predicted amplicon size were excised, using a fresh scalpel for each band, in a dark room using a UV light source for visualisation. The gel was kept on the UV light source for the minimal time to limit DNA damage and cleavage of the 3' adenine from the Taq amplified products. DNA was then extracted from the agarose gel using the Wizard® SV Gel and PCR Clean-Up System (Promega, UK) following the "Gel slice" procedure that utilizes a column based method through which DNA is initially bound and then washed and eluted into water. A modification was made to the protocol, eluting the DNA from the column with only 15 µL of water in order to increase the DNA concentration of the sample for use in a cloning reaction. The purity and concentration of the extracted DNA was then determined using a Nanodrop spectrophotometer.

Purified PCR products were then cloned into the pCR®2.1 cloning vector using TA cloning methods set out in the TA cloning® Kit (Life Technologies, U.K.) This method utilises TA cloning as a means of inserting Taq amplified PCR product into the vector and uses antibiotic selection for detection of microbes containing the

79

vector and blue white screening to detect if an insert is present. The amount of PCR product required for ligation to 50ng of pCR®2.1 vector was calculated using Equation 3-1.

**Equation 3-1 Calculation of the number of ng of PCR product required to create a 1:1 (vector: insert) molar ratio**

$$X \text{ ng PCR product} = \frac{(y \text{ bp PCR product})(50 \text{ ng pCR®2.1 vector})}{(\text{size in bp of pCR®2.1 vector} \sim 3900)}$$

In all cases described, the amplified products were ~1500bp, which requires 19.2ng of PCR product in the cloning reaction mixture to create the optimum vector-insert ratio of 1:1. The cloning reaction mixture was created and incubated at 22°C (room temperature) for 60 minutes. Incubations were carried out in a thermocycler for consistency. The components and volumes for this TA cloning reaction can be found in Table 3-4.

| Component | Volume |
|---|---|
| Extracted PCR product (~19.2 ng) | *x* µL |
| 5X T4 DNA Ligase Reaction Buffer | 2 µL |
| pCR®2.1 vector (25 ng/µL) | 2 µL |
| Water  (make up to 9 µL) | *x* µL |
| ExpressLink™ T4 DNA Ligase (5 units) | 1 µL |
| Total volume | 10 µL |

**Table 3-4 The components and volumes required for creation of the TA cloning reaction mixture**

After the incubation period, cloning reactions were kept on ice until use.

These closed vectors were then transformed into TOP10 competent *E. coli* (Life Technologies, UK) as follows. Firstly, two LB agar plates containing 100 µg/mL

ampicillin (Sigma-Adrich, UK) were made for each cloning reaction/transformation. Plates were stored at 4°C until use. Plates were equilibrated for 30 minutes at 37°C and 40 µL of 40 mg/mL X-gal (Bioline, UK), dissolved in Dimethylformamide (DMF) (Sigma-Aldrich, UK), was spread onto each plate and allowed to soak in. One 50 µL vial of One Shot® TOP10 competent cells was thawed on ice for each transformation. 2 µL of each cloning reaction mixture was then pipetted into a vial of cells and mixed gently, without pipetting, and incubated on ice for 30 minutes. The remaining cloning reaction was stored at -20°C. Cells were then subjected to heat shock at 42°C by submerging the vials in a water bath for 30 seconds, taking care not to agitate the vials at this point. The vials were then placed back on ice and 250 µL of super optimal broth with catabolite repression (S.O.C) medium (Life Technologies, UK) was added to each vial. Vials were then shaken horizontally in a shaking incubator at 225 rpm, 37°C for 1 hour. After incubation the content of each vial was plated onto two preheated agar plates, 50 µL on one plate and 200 µL on a second plate to ensure at least one plate with colonies with adequate spacing to allow them to be selected. Agar plates were incubated overnight at 37°C followed by incubation at 4°C for 2 hours to encourage maximal colour development of negative clones. A single white colony was selected from each set of transformants and grown in 5 mL of LB broth containing 100 µg/mL ampicillin overnight at 37°C, alongside a control containing uninoculated LB broth 100 µg/mL ampicillin.

Plasmids were extracted from these overnight cultures as follows. Initially, 2 mL of each overnight culture was pelleted at 13,000 xg for 1 minute. LB broth media was then removed by pipetting, ensuring as much was aspirated as possible. Plasmids were extracted using the PureYield™ plasmid miniprep system (Promega, UK)

following the "DNA purification by centrifugation" protocol, with nuclease free water being used for elution. Extracted plasmids were stored on ice for immediate use or at -20°C for long term storage.

As a diagnostic step, plasmids were digested with EcoRI (New England Biolabs, UK) for 1 hour at 37°C to ensure an insert of the correct size was present. Restriction digest reactions were created as described in Table 3-5.

| Component | Positive sample | Plasmid DNA control |
|---|---|---|
| Plasmid DNA (~500ng) | x µL | x µL |
| 10X EcoRI Reaction Buffer | 2 µL | 2 µL |
| Nuclease free water | x µL (up to 19) µL) | x µL (up to 20 µL) |
| EcoRI (20 U) | 1 µL | 0 µL |
| Total volume | 20 µL | 20 µL |

**Table 3-5 The components and volumes required for creation of the pCR2.1 EcoRI restriction digest reaction mixture**

After incubation, agarose gel electrophoresis was carried out to separate DNA fragments. A 1% agarose gel was created as described in section 3.3.1.5, 3 µL of loading buffer was added to tubes containing 10 µL of each restriction digest and control sample each sample was then mixed by pipetting. All 13 µL was then loaded onto the gel and ran alongside HyperLadder I at 100V for 75 minutes.

Plasmids seen to contain an insert of the correct size (~1500 bp) were then subject to sequencing reactions targeting the insert sequence. Inserts were sequenced using the BigDye v3.1 system (Applied biosystems, USA). Components for the BigDye reaction mixture can be seen in Table 3-6 and the thermocycling programme in Table 3-7.

| Component | Volume |
|---|---|
| Weisburg primer (10 pM/μL) | 1 μL |
| 5X BigDye 3.1 buffer | 1 μL |
| Plasmid DNA  (~400 ng ) | x μL |
| Nuclease free water | x μL (up to 19) μL |
| BigDye v3.1 | 1 μL |
| Total volume | 20 μL |

**Table 3-6 The components and volumes required for creation of the BigDye v3.1 sequencing reaction mixture**

| Stage | Temperature | Time (BigDye v3.1) | Cycles |
|---|---|---|---|
| Initial denaturation | 96°C | 5 minutes | 1 |
| Denaturation | 96°C | 10 seconds | |
| Annealing | 53°C | 5 seconds | 35 |
| Elongation | 60°C | 4 minute | |

**Table 3-7 The thermocycling stages, temperatures and times required for the BigDye v3.1 sequencing reaction**

After thermocycling, DNA was precipitated from the reaction mixture using ethanol precipitation. The reaction mixtures were transferred to 1.5 mL Eppendorf tubes and the following reagents added: 2 μL sodium acetate (3 M, pH 4.5), 50 μL of 96% ethanol and 1 μL of Glycoblue (Ambion, USA). Solutions were vortexed briefly and incubated at room temperature for 10 minutes, followed by centrifugation at 16,000 x g for 30 minutes to pellet the precipitated DNA. The supernatant was then aspirated, leaving behind a small blue pellet. To wash the pellet, 100 μL of 70% ethanol was added to each tube, mixed by inversion and incubated for 2 minutes at room temperature. Solutions were then centrifuged at 16,000 xg for 10 minutes and the supernatant aspirated. Tubes were then left open for approximately 20 minutes to allow the DNA pellet to dry, leaving a small,

dry, blue pellet at the bottom of each tube. Tubes were then stored at room temperature and sent to the University of Manchester Sequencing facility (Manchester, UK) where the sequence reaction pellets were rehydrated, ran and read using an ABI 3730 Genetic analyser (Applied Biosystems Inc, USA). DNA sequences were extracted from the output chromatograms and used as a query in an alignment search using Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) available at http://blast.ncbi.nlm.nih.gov/Blast.cgi [accessed 02/02/15]. Sequences were queried against the "16s ribosomal RNA sequence" database using the megaBLAST algorithm in order to predict the taxonomy of the bacterial gut isolates.

### 3.3.2  Identification of Gut Microbes Using Culture Independent Methods

To gain an insight into other members of the gut microbiome that were not cultured, a study using 16s rRNA amplification from a gut metagenomic DNA sample was undertaken. This involved isolation of metagenomic DNA, amplification of a short region of the 16s rRNA gene and the optimization of denaturing gradient gel electrophoresis, which separates DNA fragments of similar size and similar sequence by small regions of sequence variation that cause amplicons to retard in the gel at different points, therefore allowing separation of 16s rRNA amplicons by sequence variation and therefore, by species of origin.

### 3.3.2.1  *Microbial Metagenomic DNA Extraction:*

Gut metagenomic DNA was extracted from the *A. ater* gut using the Meta-G-Nome DNA isolation kit (Epicentre Biotechnology, USA) using the extraction from soil protocol. The whole gut tracts of 3 slugs were extracted as described previously. All three guts were cut into pieces in a single petri dish and added to a 50 mL falcon tube containing 10 mL of extraction buffer (0.1% Tween 20 (Life

Technologies, UK)) and vortexed at maximum speed for 3 minutes to disperse gut material. The gut suspension was then centrifuged at 1,600 x g for 4 minutes to pellet the larger gut pieces, leaving microbial matter in suspension. The supernatant was then collected in a sterile 50 mL falcon tube. A further 20 mL of extraction buffer was then added to the gut component pellet and the pellet resuspended by vortexing for 1 minute. The suspended mixture was then centrifuged at 900 x g for 3 minutes and the supernatant combined with the supernatant collected previously. The gut component pellet was then re-extracted by adding another 20 mL of extraction buffer and repeating centrifugation. The resulting pooled 50 mL supernatant solution was then centrifuged for 2 minutes at 900 x g and transferred to another sterile falcon tube, being careful not to disturb any pellet that formed. The supernatant was then filtered through a 1.2 µm filter membrane in order to remove any gut debris including host and plant tissue. Filtered supernatant was then passed through a 0.45 µm filter to trap microbial mass on the filter. The microbial mass was then washed off the filter membrane by placing it into a sterile 50 mL falcon tube along with 1.5 mL of filter wash buffer (0.1% tween 20), the solution was then vortexed at increasing speeds for 2 minutes to wash the trapped microbes into the filter wash buffer. The wash buffer was then pipetted into a 1.5 mL Eppendorf tube and centrifuged at 16,000 x g for 2 minutes to pellet suspended microbes. The wash buffer was aspirated and DNA extracted from the microbial mass using lysozyme/proteinase k lysis, protein precipitation and ispropanol/ethanol DNA precipitation solutions provided in the kit following the manufacturers instructions. The concentration and quality of the extracted metagenomic DNA was assessed using a Nanodrop spectrophotometer. All DNA samples were stored at -20°C until use.

### 3.3.2.2 PCR Amplification of 16s rRNA Gene Fragments From Gut Metagenomic DNA

Microbial metagenomic DNA from slug gut content was isolated as detailed above and PCR conducted with two sets of primers, designed specifically for DGGE studies. Primer set 1 was used for the initial optimization stages of the DGGE study. The poor separation of PCR products produced using primer set 1 lead to the use of primer set 2 which produced a longer amplicon which was more easily separable using DGGE. Table 3-8 contains primer sets used to amplify partial 16s rRNA sequences for separation using DGGE. Each forward primer contains a GC clamp at the 5' end that helps to promote specific binding. Primer set FP338c1 and RP534 were selected due to their use in previous study to target cellulose degrading bacteria specifically and Primer set 2 was chosen because of their larger amplicon size (~400bp) and broad bacterial specificity.

| Primer set | Primer | Sequence | Target/source |
|---|---|---|---|
| Primer set 1 | FP338c1 | 5'- CGC CCG CCG CGC CCC GCG CCC GGC CCG CCG CCG CCG CCG CAT CCT ACG GGA GGC AGC AG -3' | Cellulose degrading bacteria Specific (Vlková et al., 2005, Kopečný et al., 2004) |
| | RP534 | 5'- ATT ACC GCG GCT GCT GG -3' | |
| Primer set 2 | F984GC | 5'- CGC CCG CCG CGC GCG GCG GGC GGG GCG GGG GCA CGG GGG GAA CGC GAA GAA CCT TAC -3' | General bacterial (Nübel et al., 1996, Zhang and Jackson, 2008) |
| | R1378 | 5'- CGG TGT GTA CAA GAC CC -3' | |

**Table 3-8 Primer sets used to amply partial 16s rRNA sequences for separation using DGGE**

The reaction mixture for the PCR reactions carried out using both primer sets 1 and 2 is shown in Table 3-9 and PCR thermocycling times for each primer set shown in Table 3-10 and Table 3-11.

| Component | Volume |
|---|---|
| MyTaq red 2x mix (Bioline, UK) | 25 µL |
| Forward Primer | 1 µL(0.2 µM) |
| Reverse Primer | 1 µL (0.2 µM) |
| Template DNA | 30ng |
| Nuclease-free water | Up to 50 µL |
| Total volume | 50 µL |

**Table 3-9 The PCR reaction mixture for 16s rRNA gene fragment amplification for DGGE analysis**

Cycle 1 – Touch down

| Stage | Temperature | Time (My Taq®) | |
|---|---|---|---|
| **Initial denaturation** | 95°C | 3 Minutes | |
| **Denaturation** | 95°C | 10 Seconds | X20 (annealing decreasing by 0.5°C every cycle until 55°C) |
| **Annealing** | 65°C | 10 Seconds | |
| **Elongation** | 72°C | 10 Seconds | |

Cycle 2 – Constant annealing temperature

| Stage | Temperature | Time (My Taq®) | |
|---|---|---|---|
| **Denaturation** | 95°C | 10 Seconds | X30 (Constant 55°C annealing) |
| **Annealing** | 55°C | 10 Seconds | |
| **Elongation** | 72°C | 10 Seconds | |
| **Final** | 72°C | 2 Minutes | |

**Table 3-10 Thermocyling temperatures and times for amplification of 16s rRNA gene fragments using DGGE primer set 1**

| Stage | Temperature | Time (Dream Taq®) | Time (My Taq®) | |
|---|---|---|---|---|
| Initial | 94°C | 5 Minutes | 3 Minute | |
| Denaturation | 94°C | 30 Seconds | 10 Seconds | ⎫ |
| Annealing | 56°C | 30 Seconds | 10 Seconds | ⎬ X35 |
| Elongation | 72°C | 30 Seconds | 10 Seconds | ⎭ |
| Final | 72°C | 10 Minutes | 2 Minutes | |

**Table 3-11 Thermocyling temperatures and times for amplification of 16s rRNA gene fragments using DGGE primer set 2**

### 3.3.2.3 Creation of DGGE Optimization Sample '16s Gene Mix'

Genomic DNA extracts from 10 bacteria isolated during the culturable microbe study were subjected to PCR with primer set 1. The resultant amplification products from each of the 10 bacterial DNA samples were then pooled to create a sample that could be used for DGGE optimization. This sample would imitate a mixture of 16s rRNA gene fragments that would be produced using a metagenomic DNA sample as template, preventing wastage of 'precious' metagenomic DNA extracts.

### 3.3.2.4 Denaturing Gradient Gel Electrophoresis

The following procedure for DGGE is the final result of multiple optimization stages. These stages and alterations made are outlined in the results section, showing clear stepwise modifications that lead to optimal separation of amplicon 16s rRNA gene fragments. For clarity, the following procedure is the final

procedure that yielded successful separation of DNA bands that were excised and sequenced successfully.

DGGE was carried out using an SDS PAGE protean I (Biorad) gel tank apparatus. Firstly, two sets of glass plates were extensively cleaned, with soap and a soft sponge, followed by rinsing with MQ water. Next the plates were cleaned with 70% ethanol, allowed to air dry. Plates were then finally cleaned using a KOH/methanol solution (5 g/100 mL). Vacuum grease was applied sparingly to the inter plate spaces which were then placed at the outer edges of the plates to make the gel sandwich. The casting apparatus clamps were then screwed onto the plate sandwich and vacuum grease applied to the lower gasket to prevent leakage of acrylamide mixtures. The gel sandwich was then locked in to the casting apparatus using the cams.

Gradient gels were formed using a gradient mixer (Hoefer, USA) and a peristaltic pump. Two denaturing 6% acrylamide solutions containing urea and formamide (UF) were used to create a linear gradient, one containing a high concentration (60% w/v) and the other a low concentration (30% w/v) of urea and formamide, the components for these solutions can be seen in Table 3-12.

| | 30% UF solution (Low) | 60% UF Solution (High) | Stacking gel solution |
|---|---|---|---|
| Urea (Fisher, Scientific, UK) | 31.5 g | 63 g | n/a |
| Formamide (Fisher, Scientific, UK) | 30 mL | 60 mL | n/a |
| Acylamide (40%) (Sigma-Aldrich, UK) | 37.5 mL | 37.5 mL | 40 mL |
| 50X TAE buffer (Life Technologies, UK) | 2.5 mL | 2.5 mL | 2.0 mL |
| Deionised water | to 250 mL | to 250 mL | to 200 mL |

**Table 3-12 The components to create 30% and 60% UF DGGE resolving gel solutions and stacking gel solution**

The volumes of both the high and low acrylamide solutions for the resolving gel and the volumes required for the stacking gel are outlined in Table 3-13

| UF Solution | % UF | Volume UF solution (ml) | Volume APS (10%) (µl) | Volume TEMED (µl) |
|---|---|---|---|---|
| High | 60% | 13 | 78 | 6 |
| Low | 30% | 13 | 78 | 6 |
| Stacking | 0% | 8 | 80 | 11 |

**Table 3-13 Solutions and volumes required for creation of a DGGE gel (60% - 30% gradient)**

To prepare the gel, one 800µl aliquot of 10% APS (Fisher Scientific, UK) was thawed. The gradient mixer (containing a magnetic flea in the chamber closest to the outlet tap) was placed above a magnetic stirrer with a tube leading from the outlet tap to a peristaltic pump. On the end of the tube leading from the pump a needle was fixed and placed between the glass plates.

The required volumes of 30%, 60% and 0% acrylamide solutions were placed into separate 50ml falcon tubes. Ammonium persuphate (APS) and TEMED (Sigma-Aldrich,UK) were added to the 30% and 60% tubes and the tubes inverted twice, the 'High' solution was poured into the gradient mixer chamber closest to the outlet tap and the 'low' solution into the other with the central chamber connector closed. The peristaltic pump and magnetic stirrer were then switched on and both taps on the gradient mixer opened simultaneously. Before the resolving gel was set, APS and TEMED was added to the stacking gel mixture which was mixed by inversion and pipetted slowly into the gel sandwich to avoid mixing and a 16 well comb inserted. Gels were left for 1 hour to polymerise. The gradient mixer and pump were flushed out with deionised water immediately after use to prevent blockages.

During polymerisation the gel tank was filled with 0.5X TAE (40 mM Tris, 20 mM acetic acid and 1 mM EDTA) running buffer and placed into a large water bath at 60°C to create the temperature required to aid DNA denaturation. Gels were then fixed into the gel tank and the inner chamber filled with previously warmed 0.5X TAE running buffer. Wells were flushed using a LUERlock syringe and needle and 20µl of PCR product amplified from metagenomic DNA template using DGGE primer set 2 was loaded into two wells in each gel. The gels were then run at 100v 60°C for 20 hours. After electrophoresis each gel was carefully removed from the plates and immersed in a 3x gel red staining solution containing 0.1 M NaCl for 30 minutes, followed by visualisation in a transilluminator.

### 3.3.2.5 DGGE Band Processing, Sequencing and Microbe Identification

Upon successful separation of metagenomic template derived 16s rRNA products, gels were visualised using a UV box and working quickly to minimize UV exposure, individual bands were cut out of the gel using a fresh scalpel for each band. Gel slices were then stored in Eppendorf tubes on ice. To facilitate gel extraction, bands were then placed into wells of a 1% agarose gel, sealed in with liquid agarose, and electrophoresed into the agarose gel at 100v for 75 minutes alongside HyperLadder I. Agarose gel slices were then taken and DNA extracted using the *Wizard® SV* Gel and PCR Clean-Up System (Promega, UK). Gel extraction eluents were then used as template for another round of PCR using the same primers as with which the sample of origin was made using the same PCR conditions (DGGE primer set 2, components: Table 3-9, conditions:Table 3-11). PCR products were ran on a 1% agarose gel, excised and DNA extracted using *Wizard® SV* Gel and PCR Clean-Up System (method previously described) and the eluted DNA used as template for big dye V3.1 sequencing. The reaction

components and volumes for sequencing using PCR products can be found in Table 3-14. The thermocylcing programme used for sequencing can be found in Table 3-7, but in this case the primer specific annealing temperature used was 56°C. DNA was then precipitated using the ethanol precipitation procedure explained in section 3.3.1.6 and sent to the University of Manchester for sequencing. Sequences received were then queried against the 16s rRNA database using the megaBLAST algorithm (NCBI).

| Component | Volume |
|---|---|
| **DGGE Primer 2 (10 pM/µL)** | 1 µL |
| **5X BigDye 3.1 buffer** | 1 µL |
| **Re-amplified, gel extracted DNA  (~50 ng )** | $x$ µL |
| **Nuclease free water** | $x$ µL (up to 19) µL) |
| **BigDye v3.1** | 1 µL |
| **Total volume** | 20 µL |

**Table 3-14 The components and volumes required for creation of the BigDye v3.1 sequencing reaction mixture using PCR product as template**

## 3.4 Results

### 3.4.1 Culturable Cellulolytic Gut Microbe Identification:

In this section shows the results of the study into isolation and identification of cellulolytic bacteria. This includes slug dissection and microbial isolation and culturing, PCR amplification of the microbial 16s rRNA genes and subsequent cloning, transformation, DNA sequencing and finally, identification through alignment of the gene sequences to database entries. An image depicting the gut tract of *A. ater* can be seen in chapter 2 (Figure 2-2). Figure 3-1 shows the initial culturing of gut microbes from the homogenised gut samples.
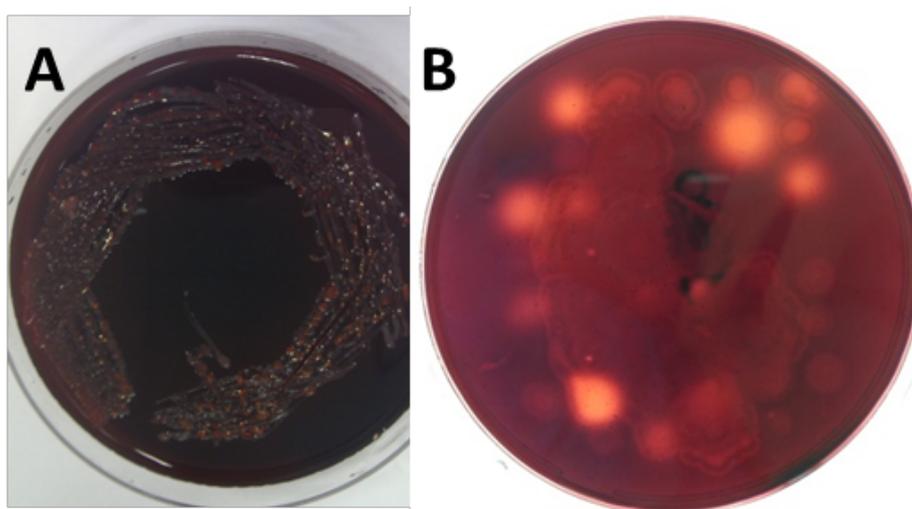


**Figure 3-1 Initial gut microorganism culturing on CMC (0.5%) LB agar plates and identification of cellulolytic colonies**
*A) A CMC LB agar plate showing microbial mixtures plated after dilution in ringers solution B) A replica CMC LB agar plate after treatment with Congo Red staining and NaCl destaining*

The plate in Figure 3-1 (A) shows growth of microbes after plating the gut homogenate, diluted in Ringer's solution. Multiple colonies can be seen with many individual colonies indicating adequate dilution. Plate (B) shows a replica plate after staining with Congo Red and destaining with NaCl. The clear zones here indicate areas on which cellulolytic microorganisms were growing, where the

cellulose in the media has been broken down leading to that area of the plate not retaining the Congo Red stain. Colonies from those locations were then streak plated and tested for cellulolytic activity, this process is depicted in Figure 3-2.
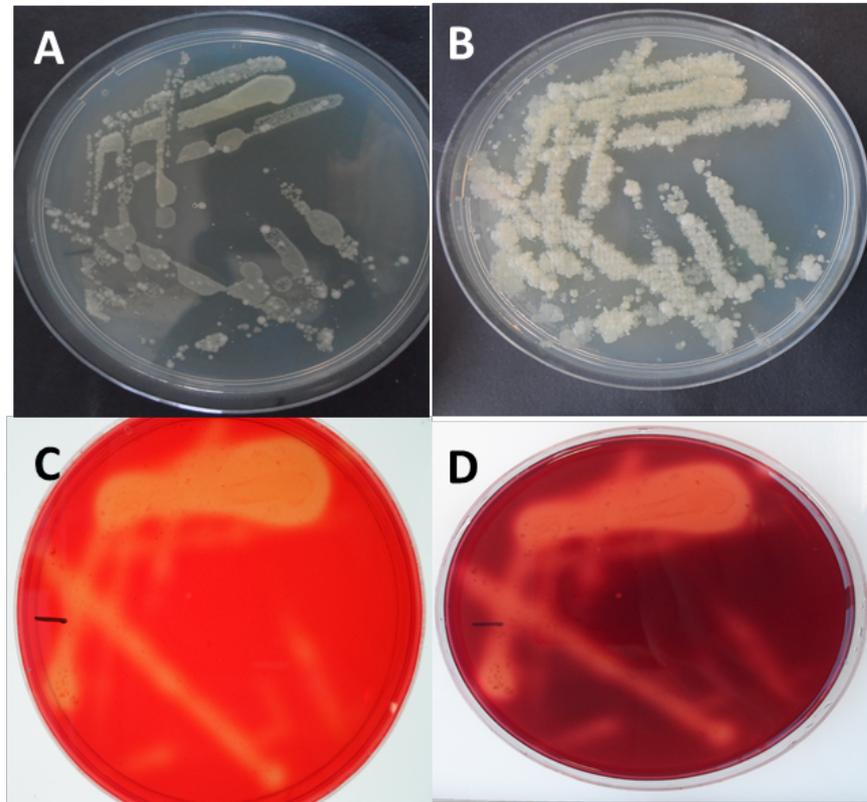


**Figure 3-2  Streak plate isolation of cellulolytic gut bacteria using CMC LB agar plates and congo red staining**
*A) Streak plate isolation of suspected cellulolytic microbes B) A replica plate created using sterile microfiber cloth C) The agar plate seen in A after treatment with Congo Red and NaCl destaining D) Shows the plate in C after treatment with acidic NaCl*

The agar plates in Figure 3-2 depict the stages of streak isolation of celluloytic microbes (A) shows a streaked isolated bacteria growing on a CMC LB agar plate. Plate (B) shows successful transfer of these microbes onto a replica plate. This was done in order to have a copy of the original plate; the replica plate was used to select bacteria for further study after staining of the original plate revealed areas in which cellulolytic microbes were growing. (C) shows the original plate after staining and destaining, revealing multiple clear zones indicating areas of

cellulolytic activity. (D) shows a plate that was treated with an acidic NaCl solution that enhanced visualisation of clear zones by turning the Congo Red stain still bound to the plate a deep purple, revealing more areas of cellulolytic activity. Single colonies from replica plates were then grown in liquid media and stored at -80°C as previously described. These frozen microbes were then re-tested for cellulolytic activity using CMC growth plate assays and for β-glucosidase activity using esculin hydrate-ferric ammonium citrate growth plate assays to confirm isolation of the correct, cellulolytic microbe.
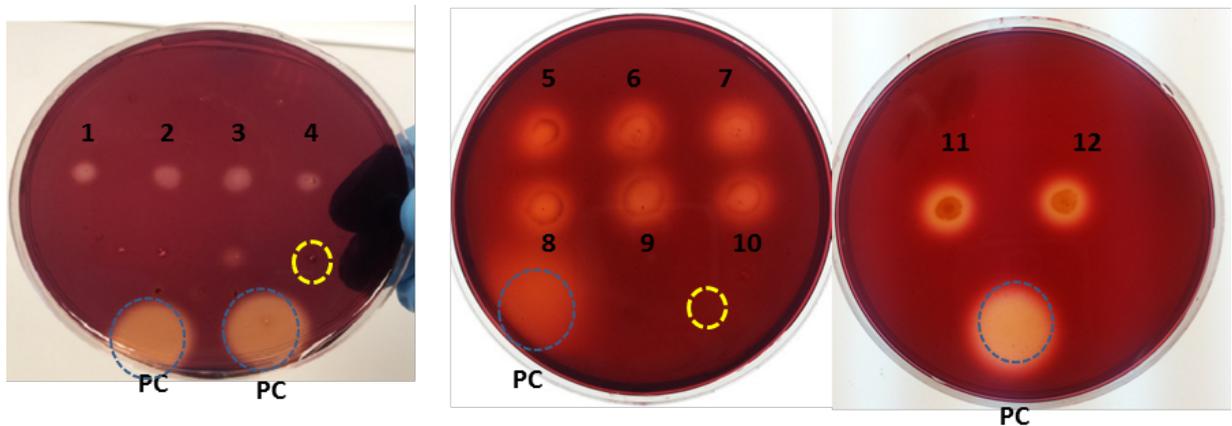


**Figure 3-3 CMC growth assay plates showing β-1,4-endoglucanse activity for 12 isolated bacteria**
*Three CMC (0.5%) LB agar growth plates after frozen isolates were regrown for 16 hours and subsequently stained using Congo Red. Blue circles depict apergillus niger positive control (4 mg). Yellow circles depict area that untransformed TOP10 E.coli were grown as a negative control.*

Figure 3-3 shows that the frozen microbes could be regrown from frozen stocks and still exhibit cellulolytic activity, while also reconfirming successful isolation of celluloytic bacteria.
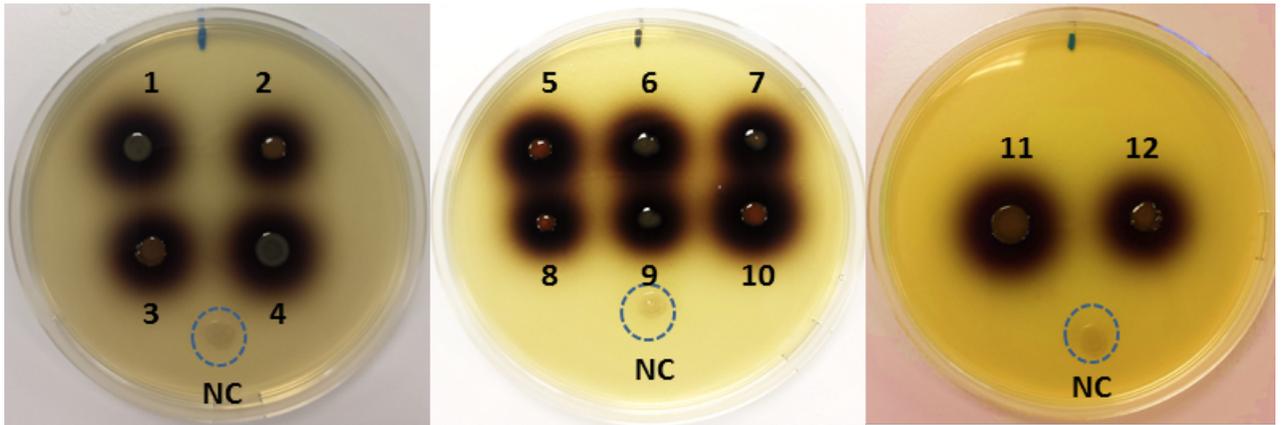
**Figure 3-4 Esculin hydrate-ferric ammonium citrate growth assay plates showing β-glucosidase activity for 12 isolated bacteria**
*Three growth plates showing regrowth of frozen celluloytic bacterial gut isolated. Untransformed TOP10 E.coli were used as a negative control depicted by blue circles*

Figure 3-4 shows the 12 cellulolytic isolates grown on LB agar containing esculin hydrate and ferric ammonium citrate. The black precipitate around the colonies indicates β-glucosidase activity.
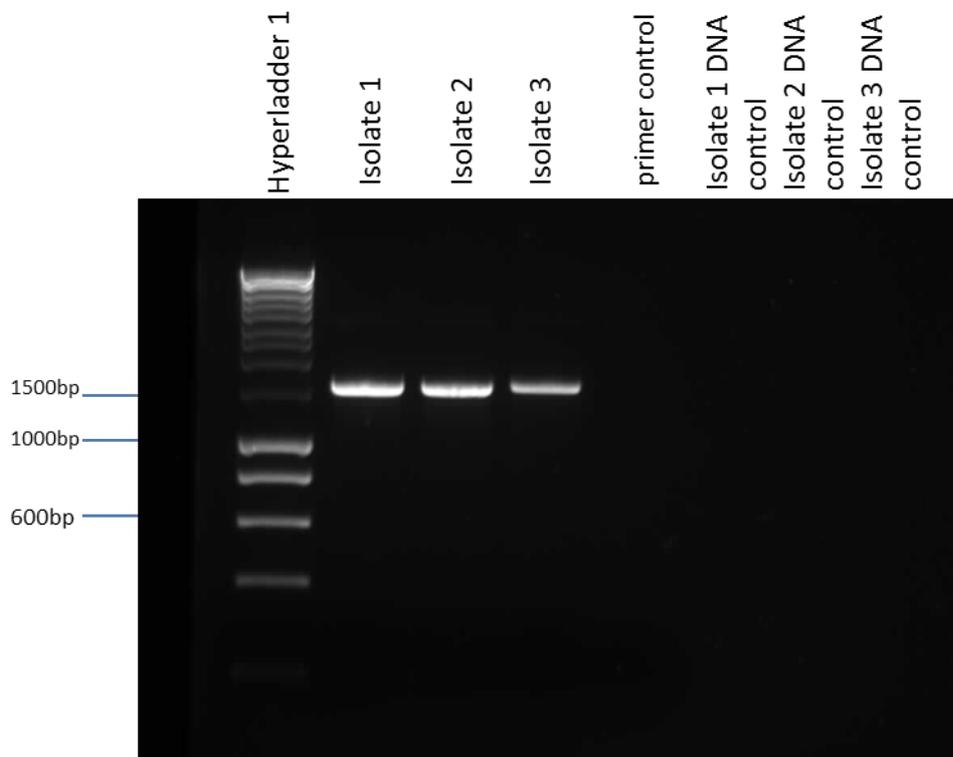
**Figure 3-5  A 1% agarose gel showing the PCR amplification of the 16s rRNA gene of 3 cellulolytic bacteria isolated from the slug gut**

*Lane 1: HyperLadder I, lanes 2-4: The amplification products of isolates 1-3, lanes 5: A primer control sample, Lanes 6-8 DNA control samples for isolates 1-3*

Figure 3-5 shows that the PCR reactions were successful, indicated by amplification of a single product of expected length (~1500bp). The gel shows the amplification products of PCR targeting the 16s rRNA genes for 3 cellulolytic isolates. Lane 5 appears devoid of amplification products showing that none the components were contaminated with bacterial DNA and that the primer set is not forming any dimers. Lanes 6-8 are also clear indicating that the bands seen in 1-3 are indeed the result of targeted PCR amplification that are not present in the genomic DNA sample.

These ~1500bp products were excised and extracted from the gel slices as described previously. In this case half of the extracted products were run once again on a separate 1% agarose gel to confirm extraction success (Figure 3-6).
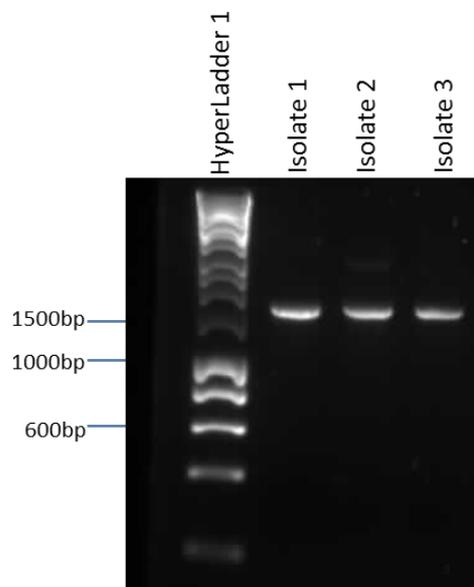


**Figure 3-6 A 1% agarose gel showing the successful extraction of amplification products of isolates 1-3 from 1% agarose gel slices**

97

Figure 3-6 confirms successful extraction of 1500bp PCR products from the gel in Figure 3-5. These products were subsequently cloned using the original TA cloning kit. The diagnostic restriction digest of plasmids extracted from successful transformants with EcoRI is shown in Figure 3-7.
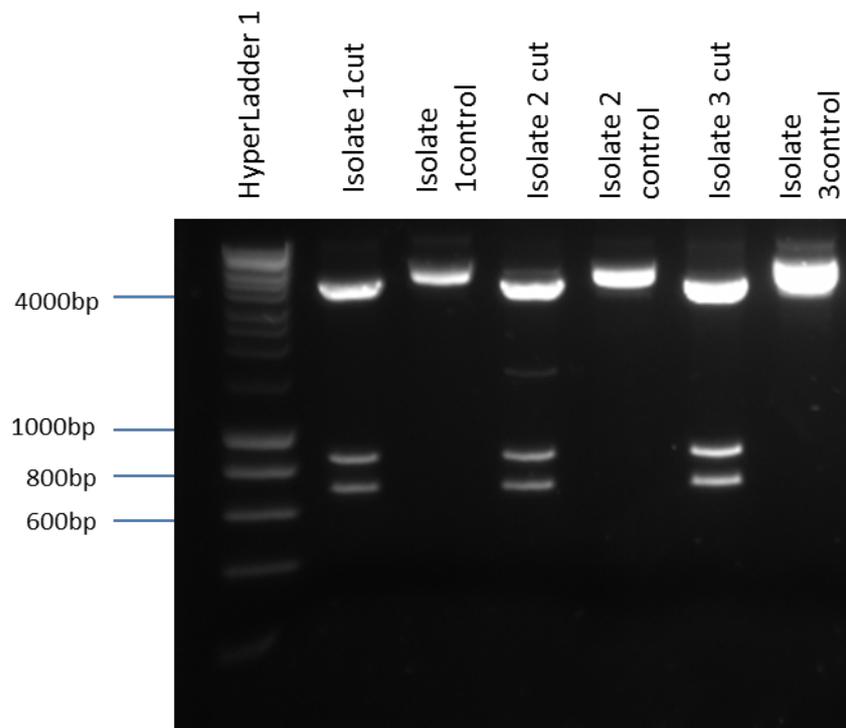


**Figure 3-7 Restriction digest of pCR2.1 plasmids with EcoRI to confirm the presence of 16s rRNA PCR product inserts**
*Lane 1: HyperLadder I, Lanes 2, 4 and 6: Isolates.1-3 plasmids cut using EcoRI, Lanes 3, 6 and 7: Negative control samples containing uncut Isolate 1-3 plasmids.*

Figure 3-7 shows bands at >4000bp in all control lanes, which correspond to closed supercoiled plasmids that cannot be sized correctly against a linear molecular size marker. Lanes 2, 4 and 6 show 3 predominant bands, the largest corresponding to the linearized vector (~3900bp). In each case two bands are seen at ~700 bp and at ~850 bp. Another band is also visible at ~1500bp in lane 4. EcoRI only cuts in two places on the plasmid, so presence of three bands suggests an internal EcoRI recognition site within the 16s rRNA inserts as the two bands at approximately 700 bp and 850 bp added up to the ~1500 bp expected

product. This was confirmed following sequencing of the plasmid inserts. After sequencing, a plasmid map of pCR®2.1 was created containing a complete 16s rRNA sequence using SnapGene (USA) software (Figure 3-8).
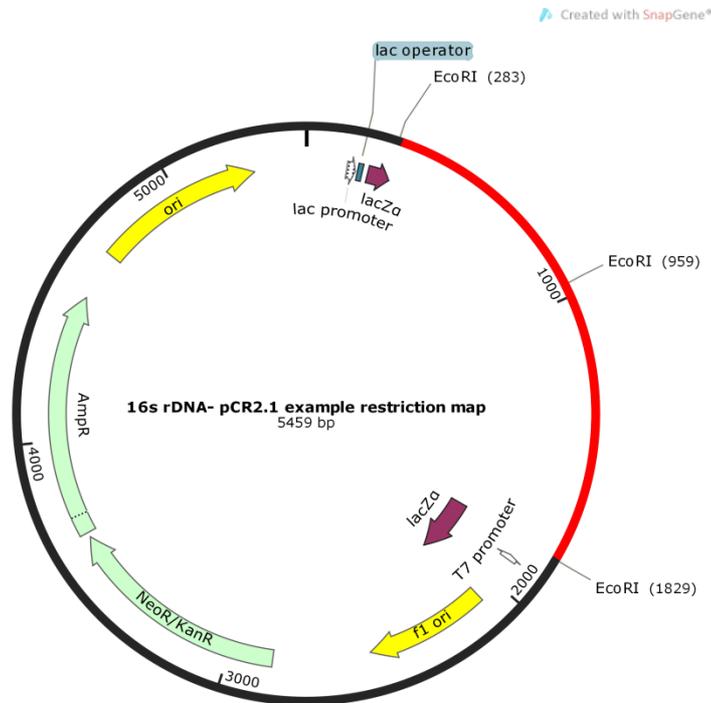


**Figure 3-8 A plasmid map showing a 16s rRNA gene in a pCR2.1 vector produced using SnapGene**

Figure 3-8 shows the plasmid map for pCR®2.1 with a full length 16s rRNA gene inserted. The map shows three cut sites for EcoRI, 2 which are within the vector that would allow diagnostic release of the insert and 1 within the 16s rRNA gene.

The inserts from successful clones were then sequenced using Bigdye v3.1 (Life Technologies, UK) and the sequences queried against the NCBI 16s rRNA database for identification using a BLAST search. The outputs of these searches for the 12 cellulolytic isolates can be seen in Table 3-15. All 16s rRNA sequences can be found in Appendix 1.

Each of the 9 remaining isolates were sequenced using the direct sequencing method using PCR product as template for sequencing as described in Table 3-14. The 16s rRNA gene PCR amplifications for each of the 9 other isolates can be seen in Figure 3-9 and Figure 3-10.
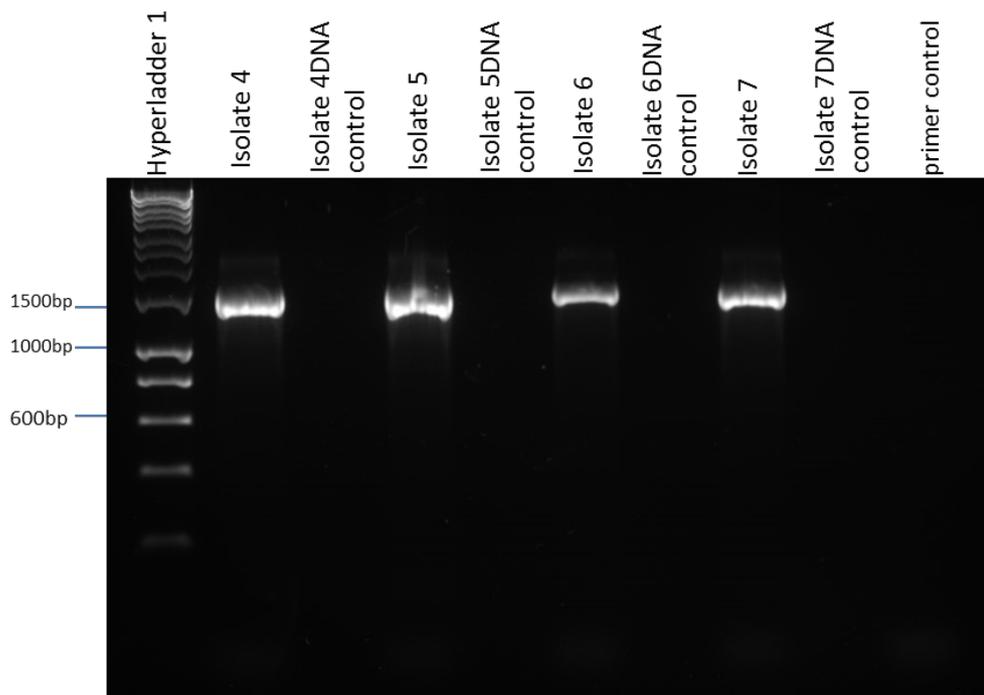


**Figure 3-9 A 1% agarose gel showing the PCR amplification of the 16s rRNA gene of isolates 4-7**

*Lane 1: HyperLadder I, Lanes 2, 4, 6 and 8 : The amplification products of isolates 4-7, lane 10: A primer control sample, Lanes 3, 5, 7 and 9: DNA control PCR samples*
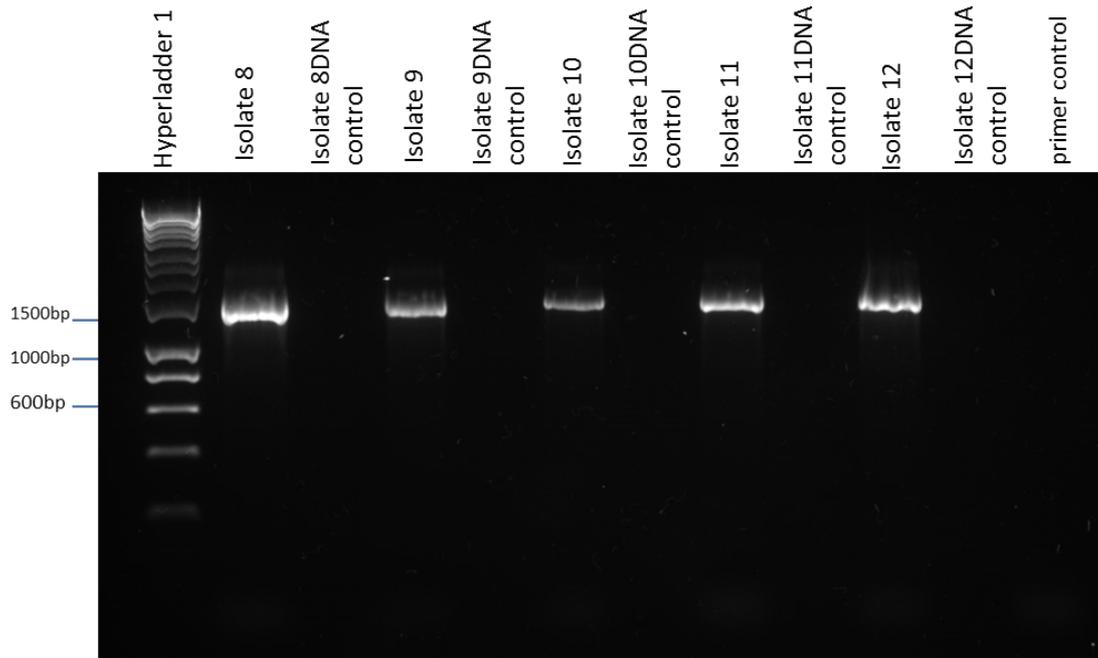
**Figure 3-10 A 1% agarose gel showing the PCR amplification of the 16s rRNA gene of isolates 8-12**
*Lane 1: HyperLadder I, Lanes 2, 4, 6, 8 and 10: The amplification products of isolates 8-12, lane 12: A primer control sample, Lanes 3, 5, 7, 9 and11: DNA control PCR samples*

Figure 3-9 and Figure 3-10 show successful amplification of the 16s rRNA gene for each of the 9 remaining isolates, with clear bands visible at the expected 1500 bp region. No bands can be seen in the DNA control or primer control showing that amplification was not due to contamination of materials.

| Name | Closest database match | E-value | Identity | Accession |
|---|---|---|---|---|
| **Isolate 1** | *Citrobacter braakii* | 0 | 85% | NR_028687.1 |
| **Isolate 2** | *Salmonella enterica* | 0 | 91% | NR_044371.1 |
| **Isolate 3** | *Klebsiella pneumonia* | 0 | 96% | NR_037084.1 |
| **Isolate 4** | *Serratia marcescens* | 0 | 91% | NR_036886.1 |
| **Isolate 5** | *Buttiauxella agrestis* | 0 | 79% | DQ440549.1 |
| **Isolate 6** | *Serratia liquefaciens* | 0 | 86% | GU586145.1 |
| **Isolate 7** | *Aeromonas hydrophila* | 0 | 99% | NR_104824.1 |
| **Isolate 8** | *Acinetobacter calcoaceticus* | 0 | 92% | NR_042387.1 |
| **Isolate 9** | *Kluyvera intermedia* | 0 | 99% | KF724024.1 |
| **Isolate 10** | *Buttiauxella agrestis* | 0 | 99% | NR_041968.1 |
| **Isolate 11** | *Citrobacter freundii* | 0 | 99% | NR_028894.1 |
| **Isolate 12** | *Enterobacter sp. E6-PCAi* | 0 | 94% | JN853247.1 |

**Table 3-15 NCBI BLASTn search results for each amplified 16s rRNA gene from cultured cellulolytic microbes**

*Sequences obtained for each isolate were queried against the NCBI 16s rRNA database or the nr database if no match was found*

### 3.4.2 Identification of gut microbes using culture independent methods

In order to identify uncultured microbes, a DGGE study was carried out using gut microbial metagenomic DNA as template for 16s rRNA amplification.
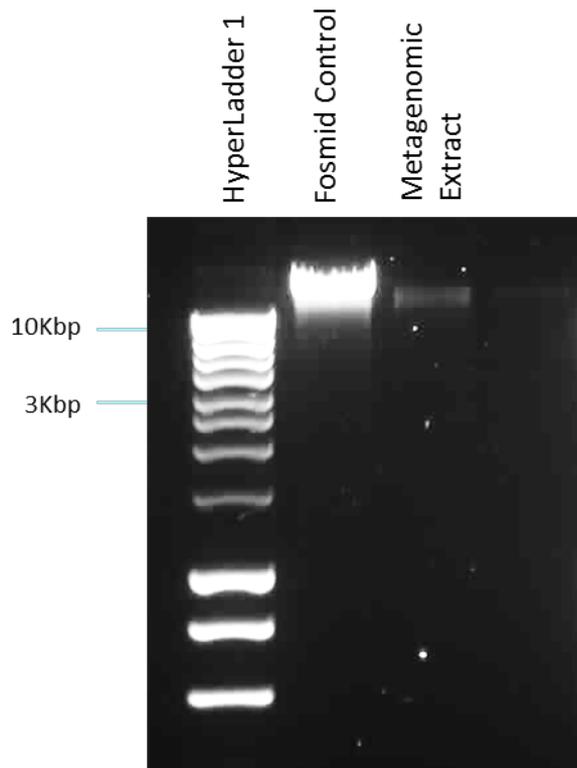
**Figure 3-11 A 1% agarose gel showing successful extraction of metagenomic DNA from the slug gut**

*Showing intact genomic DNA from microbial mass extracted from the slug gut environment using filtration methods. Lane 1 shows HyperLadder I with the largest band being 10Kbp. Lane 2 shows 100 ng of a 40Kbp fosmid control provided as part of the Meta-G-Nome extraction kit. Lane 3 shows 2 μL of gut metagenomic DNA extract*

Figure 3-11 shows that the size of extracted DNA is comparable to the 40 Kbp fosmid control and therefore that intact genomic DNA was successfully extracted.

### 3.4.3 Creation of DGGE Optimization Sample 'Cultured16s Gene Mix'

In order to optimize the conditions of DGGE separation using the apparatus available, a mixture of 16s rRNA genes was created by carrying out PCR reactions using DGGE primer set 1 using DNA from individual cultured microbes as template. These individual amplicons were then pooled to create a mock up sample that could be representative of a PCR product mixture using precious

103

metagenomic DNA as template. Figure 3-12 shows the successful amplification of partial 16s rRNA sequences from 10 microbes using primer set 1.
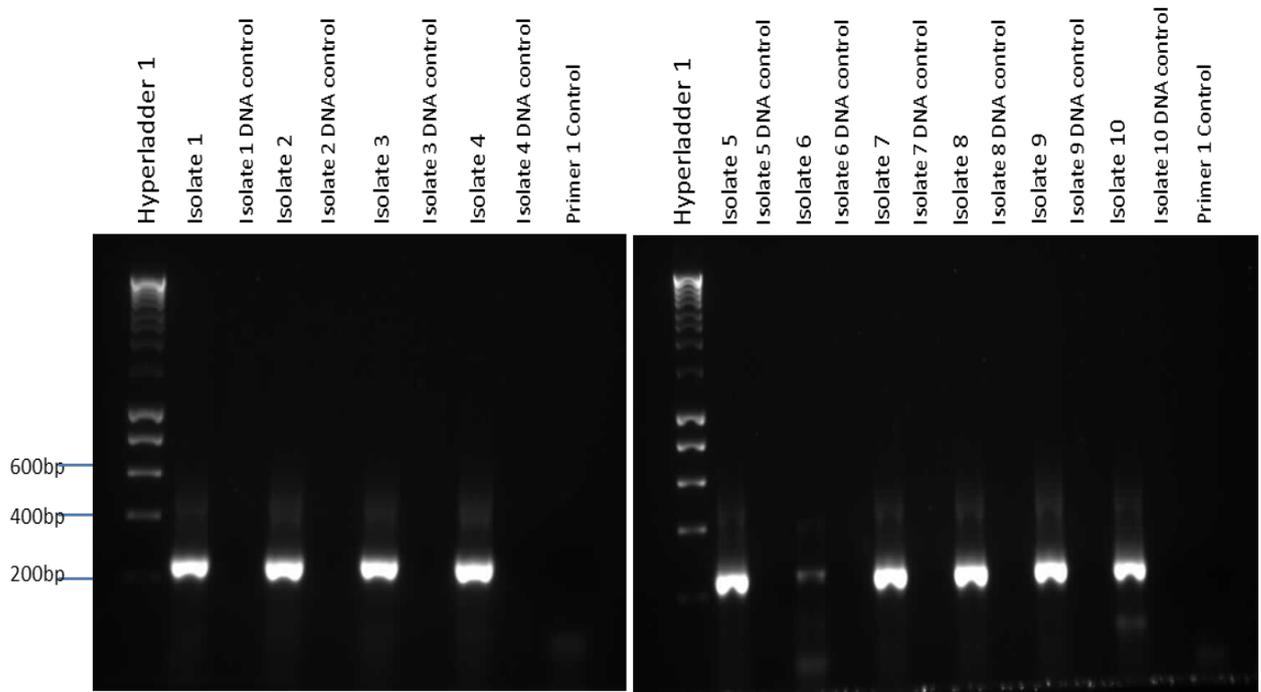


**Figure 3-12 Two 1% agarose gels showing amplification of a ~200bp fragment of the 16s rRNA gene of 10 cultured bacteria using primer 1.**

The amplification products in Figure 3-12 were pooled to create the cultured16s rRNA gene mix that was used in DGGE optimization steps.

### 3.4.4 DGGE Gel Optimisation Steps

Separation of the 16s rRNA gene mix with DGGE was then attempted using the various conditions outlined below. Initially gels with a denaturant concentration range of 45-60% were used and subsequently, using a gradient of 30-60%.

Figure 3-13 to Figure 3-17 show the results and the stages of optimization.
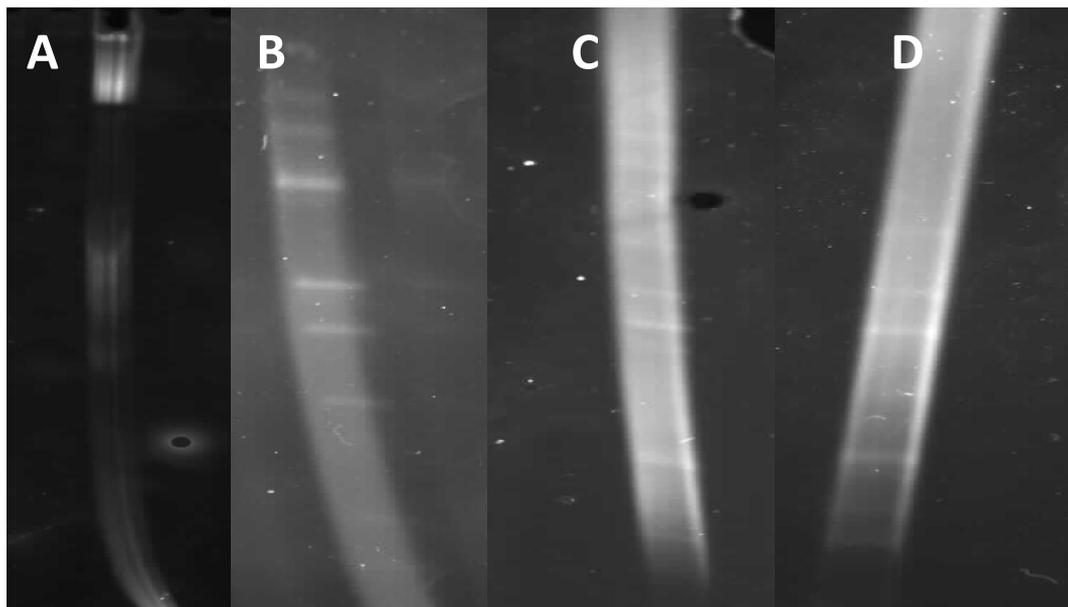
**Figure 3-13 Gel images of initial pilot DGGE gels and optimization steps. Showing separation of the cultured 16s rRNA gene mix**

Figure 3-13 A shows separation of 10ul of primer 1 16s gene mix after 14 hours of electrophoresis at 60°C at 150 volts in a 45-60% UF 6% acrylamide gel. B shows 10μl of 16s gene mix separated with the same conditions as A but at only 100v, some separation could be seen so 100v was adopted from here onwards. C and D show separation with the same conditions as in B but with gels created using different pumping apparatus, feint bands can be seen in both but there is no notable difference in resolution. These pump tests proved inconclusive so other conditions and components were altered, such as the gradient of denaturant. The results of these steps can be seen in Figure 3-14.
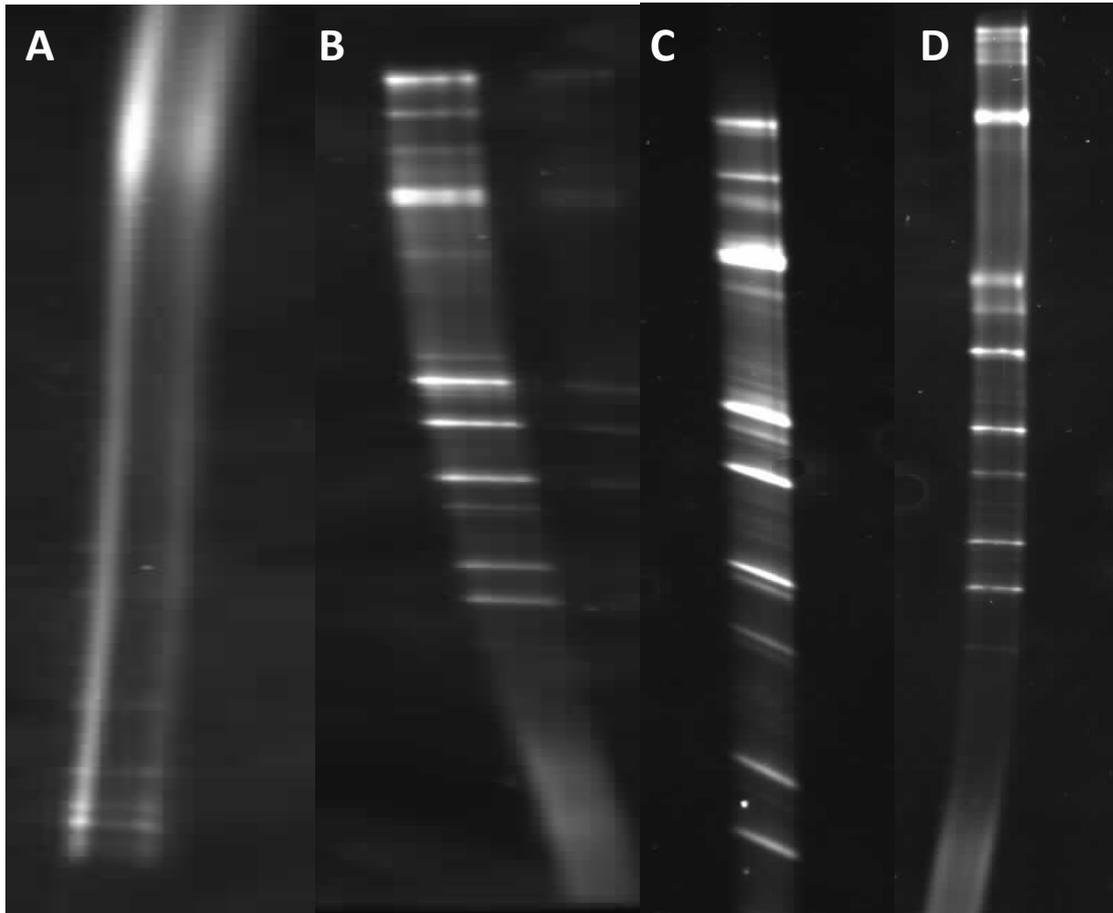
**Figure 3-14 DGGE Acrylamide gels showing separation of 16s Gene mix and results of optimization steps**

Figure 3-14 (A) shows separation of 16s gene mix using a 45-60% UF gradient 6% acrylamide gel ran at 100v for 14 hours at 60°C. The image shows little separation with few distinct bands. (B) shows a gel ran in tandem with (A) with the same conditions but using a 30-60% UF gradient resolving gel; multiple individual bands can be seen. Gels (C) and (D) were ran with the same conditions as in (B) with a 30-60% gel however gel (C) was created the night before use and (D) was created the day of use, (D) shows much higher levels of band resolution and resulted in bands which could be deemed resolved enough for gel extraction and identification. A test with conditions used for gel (D) were repeated and showed the same level of band resolution, which confirms optimization steps carried out produced a robust method for separating 200bp fragments of DNA that was also reproducible. At this point the conditions used for Figure 3-14 (D) (30-60% UF

106

gradient 6% acrylamide gel ran at 100v for 14 hours at 60°C) were therefore taken forward for tests using 16s gene fragment mixtures created using slug gut metagenomic DNA as template for PCR. Figure 3-15 shows an attempt to separate a 16s gene mix derived from PCR (as conducted for cultured 16s gene mix) using metagenomic DNA as template.
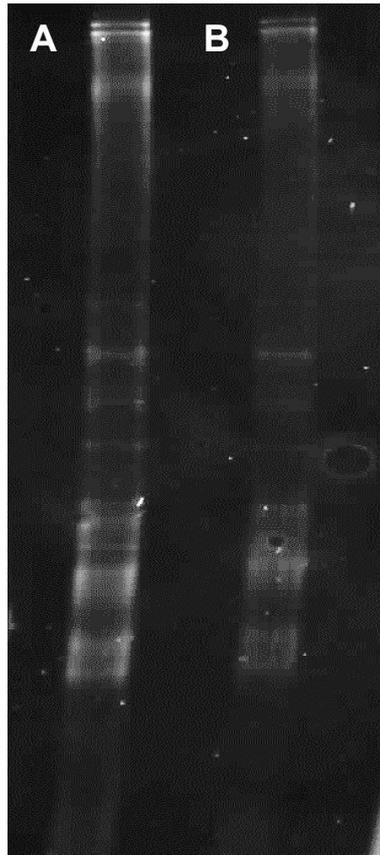


**Figure 3-15 A 30-60% UF gradient DGGE gel showing separation of 20 μl (A) and 10 μl (B) of metagenomic 16s gene mix PCR product**

Figure 3-15 shows the attempted separation with a 30-60% UF gradient 6% acrylamide gel ran at 100v for 14 hours at 60°C with A, 10μl of sample and B, 20 μl of sample. Both show very faint bands which were not visible when attempts were made to extract the bands on a UV box. Separation of PCR products was also not thought to be adequate. The experiment was repeated multiple times to no avail and the quality of PCR was deemed to be the cause. This was deduced

because the only change to the experiment seen in Figure 3-14 D was the 16s

DNA amplicon sample being separated. To that end metagenomic DNA was used

as template for PCR using primer set 2 as described in section 3.3.4 This gene

mix was then separated using the same optimized conditions as in Figure 3-14D.

The results can be seen in Figure 3-16.
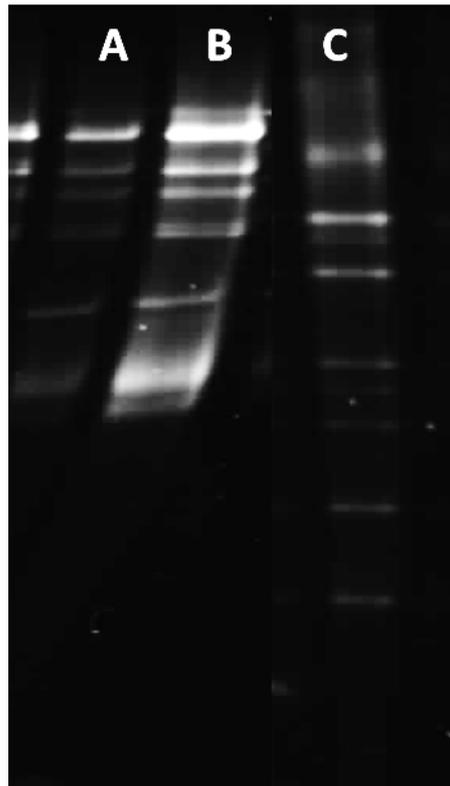


**Figure 3-16 A DGGE gel showing separation of partial 16s gene PCR products amplified using primer set 2 and slug gut metagenomic template DNA**

Figure 3-16  A and B shows separation of 10 and 20µl of metagenomic 16s rRNA

gene mix created using primer set 2. Separation was limited due to slower

migration of the larger 400bp products produced by primer set 2. Figure 3-16 C

shows successful separation of cultured 16s gene mix produced using primers set

1. Due to the inadequate separation of the larger products, the electrophoresis

time was increased to 16 hours. The results can be seen in Figure 3-17.
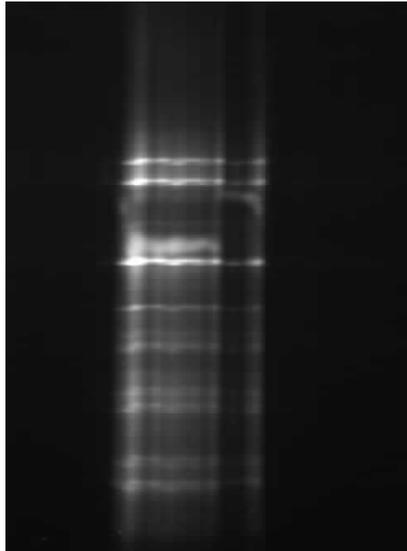
**Figure 3-17 A DGGE gel showing separation of partial 16s gene PCR products amplified using primer set 2 and slug gut metagenomic template DNA**
*Figure shows separation of 20μl of metagenomic 16s gene mix produced using primer 2 in a 30-60% UF gradient 6% acrylamide gel, after electrophoresis for 16 hours at 100v and 60°C*

Figure 3-17 shows successful separation of the 16s rDNA amplicons with multiple

individual bands visible. Each band was cut out of the acrylamide gel, placed in a

1% agarose gel well, and electrophoresed into the agarose to facilitate DNA

extraction. The results of this experiment are shown in

Figure 3-18.
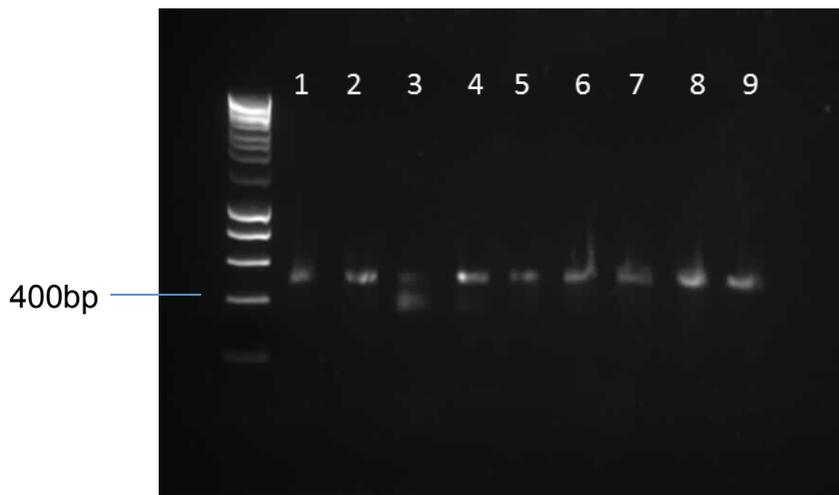
**Figure 3-18 A 1% agarose gel showing successful migration of DNA bands from acrylamide gel slices into agarose**

*Lanes 1 -8 show migration of DNA in extracted DGGE bands into agarose. DNA was ran along side hyper ladder I*

Figure 3-18 shows faint bands in lanes 1-9 corresponding to DNA extracted from bands seen seen in figure Figure 3-17. Although the products created using primer set 2 are approximately 400bp in this image, they appear to be around 600bp. We believe this to be a result of the DNA in lanes A-M having to travel a slightly further distance in solid agarose/ acrylamide than the HyperLadder I sample, hindering their migration. These bands were gel extracted and subjected to another round of PCR with primer 2. The results can be seen in Figure 3-19.
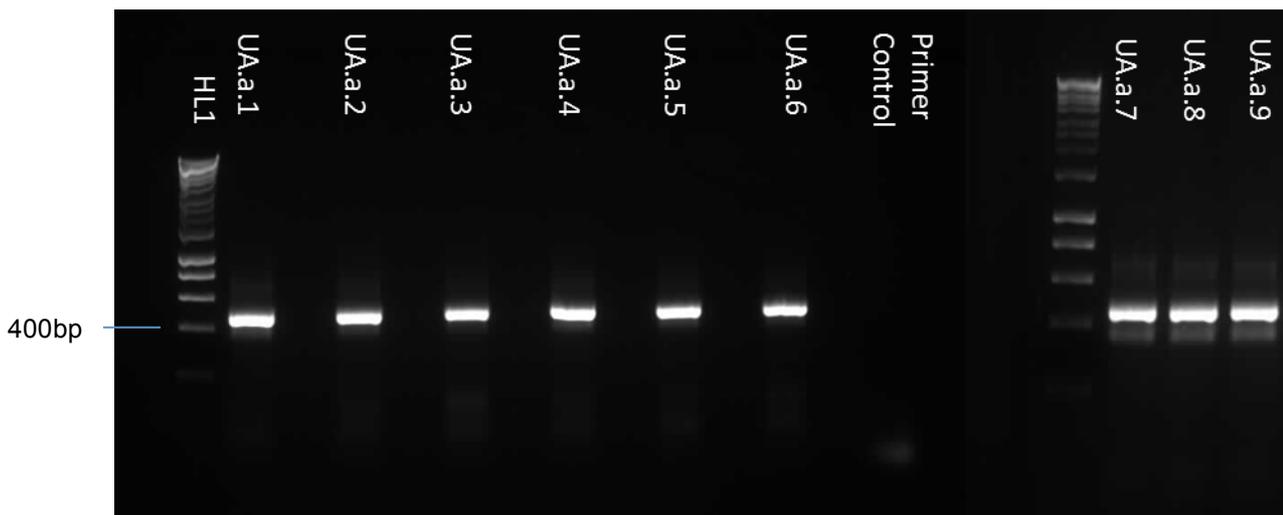


**Figure 3-19 A 1% agarose gel showing successful re-amplification of DGGE bands extracted from the gel in Figure 3-18**

*Lanes labelled UA.a1-8 (Uncultured Arion ater) show successful reamplification of DNA extracted from bands 1-8 seen in Figure 3-18*

Figure 3-19 shows successful re-amplification of DNA isolated from 9 bands cut from the DGGE gel in Figure 3-17 at a size of just over 400bp. This confirms the actual size of the bands seen in

Figure 3-18 to be the predicted ~400bp gene fragments. These PCR products were then subjected to PCR clean up and used as template for sequencing using Bigdye v3.1 as previously described. Sequences were submitted to megaBLAST and queried against the 16s rRNA database. A list of the top hits for each DGGE band can be seen in Table 3-16.

| Band | Genus/Species | E value | % Match | Accession |
|------|---------------|---------|---------|-----------|
| UA.a.1 | *Mycoplasma hyorhinis* | 1.00E-158 | 93% | NR_041845.1 |
| UA.a.2 | *Mycoplasma iners* | 4.00E-158 | 93% | NR_025064.1 |
| UA.a.3 | *Uncultured Citrobacter* | 0 | 99% | AY847172.1 |
| UA.a.4 | *Uncultured Serratia* | 0 | 100% | KC253894.1 |
| UA.a.5 | *Pectobacterium carotovorum* | 0 | 99% | NR_041971.1 |
| UA.a.6 | *Acinetobacter beijerinckii* | 0 | 98% | NR_042234.1 |
| UA.a.7 | *Pantoea sp. 57917* | 0 | 99% | DQ094146.1 |
| UA.a.8 | *Erwinia amylovora* | 0 | 99% | NR_041970.1 |
| UA.a.9 | *Erwinia tasmaniensis* | 0 | 99% | NR_074869.1 |

**Table 3-16 The megaBLAST hits for each partial 16s rRNA PCR product band isolated using DGGE**

Table 3-16 shows the megaBLAST results for each sequence, showing 2 members of the class *Mollicutes (Mycoplasma)* and 7 *Gammaproteobacteria* members. Seven of the 9 hits showed similarity greater than 97% which is the percentage widely accepted as the threshold for accurate identification, the other

2 sequences show similarities of 93%. All 16s rRNA sequences can be found in Appendix 1.

## 3.5 Discussion

In this study bacteria, of the following genera were isolated from the *A. ater* gut and identified using 16s rRNA sequencing: *Serratia, Citrobacter, Aeromonas, Acinetobacter, Salmonella, Kluyvera, Buttiauxella, Enterobacter and Klebsiella*. Each of the isolates was shown to exhibit cellulolytic activity using biochemical growth plate assays. This included observation of the ability to breakdown long chains of cellulose into smaller oligosaccharides through endoglucanase activity and the ability to breakdown down cellobiose, a product of cellulose degradation, through β-glucosidase activity resulting in cleavage and formation of two glucose monoscacharides. These methods have been successfully used to isolate and identify many cellulolytic microbes from gastropods (Antonio *et al.*, 2010), insects (Huang *et al.*, 2012) and mammals (Ruijssenaars and Hartmans, 2001), indicating the robust and reliable nature of the methods.

CMC and esculin hydrate activity growth plate assays allowed us to identify 12 cellulolytic gut microbes, only 4 of which could be identified with great confidence (>97% similarity). This strongly suggests that the *A. ater* gut microbiome contains uncharacterized microbes with uncharacterized cellulolytic systems. The lack of confidence in assignment of species makes it quite difficult to make comparisons with the literature, but with the majority of matches being greater than 92%, comparisons have been made to other studies based on genera level identifications when similarity is less than 97%. This also helps to confirm that the slug gut microbiome is indeed an understudied environment, with two thirds of isolates being undescribed in online databases.

The genera to which these isolated microbes belong have been identified previously in multiple studies of the herbivore gut. Studies of the gut content of the invasive termite *Coptotermes formosanus,* thought to be one of the most efficient lignocelluloses converters on the planet, identified microorganisms from the groups *Serratia, Citrobacter,* and *Enterobacter* which they concluded were involved in symbiotic breakdown of the lignocellulose consumed by the termite (Adams and Boopathy, 2005). The gut enzymes from the microbial consortia of multiple termite species have been extensively studied due to the high activity of the enzymes found. The species *Klebsiella pneumoniae* has been found in the gut of many species of termites (Doolittle *et al.*, 2008), is a known producer of cellulolytic enzymes (Clarke and Tracey, 1956) and also has been observed to improve the degradation of cellulose by other microorganisms in certain conditions by means of its nitrogen fixing capabilities (Ushakova *et al.*, 2003). *Klebsiella pneumoniae* is also a known fermenting bacteria which is has been used in the fermentation of lignocellulose derived sugars into fuel (Banerjee, 1989) with research into the potential to use the species commercially ongoing in 2015 (Li *et al.*, 2015). The microbes *Klebsiella pneumoniae*, *Citrobacter freundii* and *Serratia liquefaciens* have also been identified in the gut of the *Bombyx mori* larvae (silk worm) and their cellulolytic activity was also observed (Anand *et al.*, 2010). Multiple *Enterobacter* species, the species *Salmonella enterica*  and *Serratia marcescens* have also been identified in the gut of herbivorous beetle larvae during their development  (Butera *et al.*, 2012, Azambuja *et al.*, 2004). The identification of all of these microbes in the gut microbiomes of many different plant eating organisms suggests that many species may rely on a similar gut consortium to synergistically degrade lignocellulose.

The use of 16S rRNA PCR is a reliable method for culturable microorganisms but previous studies have shown that a significant portion of invertebrate gut microbiomes are unculturable with estimations reaching as high as 99.9% (Vartoukian et al. 2010). Thus, DGGE was employed in order to identify multiple microbes residing in the gut of *A. ater* that it may not be possible to culture. Many studies have successfully used DGGE to explore the diversity of microorganism populations in environments such as the guts of vertebrates (Li *et al.*, 2012a) and invertebrates (Berlanga *et al.*, 2011) and also in water samples (Gugliandolo *et al.*, 2011).

Separation of DNA bands using DGGE relies on the presence of small sequence variations in amplicons of the same length, which lead to the successive retardation of certain portions of the DNA double helix in certain strands before others. In other words, some sequences can withstand a greater concentration of denaturants than others before retardation and subsequent restriction of movement through the gel. As this separation is based on minute sequence variation, very small variations in running temperature and in the linearity of the denaturant gradient can have a dramatic effect on the separation of DNA fragments along with levels of band resolution. Early attempts seen in Figure 3-13 showed no or very little separation, with streaking being seen in all attempts. The optimization steps seen in Figure 3-13B showed that the 150v used in Figure 3-13A was too great, causing complete streaking of the sample. After the voltage was reduced from 150v to 100v, slight separation of bands could be seen. This suggests that 150v could have been a high enough voltage to cause variation in temperature between the electrodes and the centre of the gel, which would have been detrimental to efficient separation due to retardation of multiple fragments at

the same time simply due to exceeding the heat threshold for double helix integrity in general, rather than at an increment dictated by small sequence variations. Further tests also ruled out the possibility of there being a problem with gel pouring apparatus that might have affected the linearity of the gradient in the gel. This was assessed by testing two different pumping systems and finding that this appeared to have little effect on samples separation Figure 3-13 C+D). To that end, tests were carried out to determine if the denaturant gradient was sufficient to separate the individual genes. A gel was created with a solution containing only 30% denaturant in order to extend the gradient created in the gel. This was ran alongside a gel made with a 45% lower limit denaturant solution. A profound difference was observed, with the 30-60% gradient gel producing a great increase in band separation and a much higher level of band resolution. We therefore deduce that this is due to the more gradual increase in denaturant which causes more gradual retardation based on smaller sequence variations. The conditions in Figure 3-14D were utilized to separate a PCR sample created using the same primers (Primer set 1) but with metagenomic DNA as template, initial tests were unsuccessful but, because we now know that the conditions are adequate to separate DNA fragments of the size produced by primer set 1, it was deduced that the quality of the PCR itself was the issue. To circumvent this issue DGGE primer set 2 was used to amplify a 400 bp region of the 16s rRNA gene from the many microbial genomes present in the metagenomic DNA sample. The amplified PCR products were separated successfully with the only modification to the conditions being an increase in electrophoresis time to 20 hours to account for the increase from around 200bp fragments to the 400bp fragments produced by DGGE primer set 2. The majority of studies in the literature which utilise DGGE go onto place

extracted bands into nuclease free water overnight to liberate the DNA fragments, giving mixed results. Electrophoresis of these bands into 1% agarose and subsequent gel extraction using a specialised kit proved to be a very effective way to extract DNA with a purity and concentration sufficient to be used as template for another round of PCR, which in turn facilitates direct sequencing of the PCR product. This streamlines the traditional approaches and greatly reduces the likelihood of failure to identify individual bands, with 100% success being achieved in this study.

Consistent with the data acquired during the culturable organism identification study, many microbe identifications made using DGGE were from members of the Gammaproteobacteria class, with only 2 being from other groups. Also, in concordance with the culturable study, members of the genera *Citrobacter, Serratia* and *Klebsiella* were found upon querying sequence data produced during the DGGE study with megaBLAST. In addition to these, 2 identifications were made for *Mycoplasma sp. hyorhinis* and *iners,* which have previously been identified in the gut of the termite *Isoptera Rhinotermitidae (Hongoh et al., 2003)*. The microbe *Pectobacterium caotovorum* which is present in the *A. ater* gut is a plant pathogen species that is linked with the development of rot based diseases such as blackleg in potato tubers (Czajkowski *et al.*, 2009) and is also known to produce all three types of enzyme required to break down lignocellulose (An *et al.*, 2005), activity which the *A. ater* species could be taking advantage of to aid degrade lignocellulose. The species *Erwinia amylovora* identified here is also an agricultural plant pathogen, causing fire blight of rosaceous plants such as apple and pear trees (Wei *et al.*, 1992) and is also known to produce multiple CAZymes including a number of endoglucanases (Riekki *et al.*, 2000).

In this study we have confirmed that at least a portion the cellulolytic activity seen in the gut of *A. ater* is due to symbiotic activity of gut microbes and, for the first time, isolated and identified individual cellulolytic microbes present. Interestingly, a metagenomic study of the gut microbiome of the giant African Snail shares all but one of the microbial species identified here (Cardoso *et al.*, 2012a), which suggests that there may be a set of gut microbes on which multiple land gastropods rely to aid their digestion of lignocellulose. The study by Cardoso *et al.* (2012a) also revealed a large repertoire of novel CAZymes associated with the gut microbiome using in metagenomic sequencing. The discovery of cellulolytic microbes in the guts of multiple gastropod species suggests that the gut microbe host interaction could have played an important role in the evolutionary dietary transitions of land gastropods as is thought to have been the case for insects (Hansen and Moran, 2014). The identification of cellulolytic microbes in the guts of both termites and cattle lead to very successful metagenomic studies. Here we have discovered cellulolytic microbes using similar methodology, this and the fact that thousands of novel CAZymes have recently been identified in the closely related African Snail indicate that the gut of *A. ater* is a viable target for a shotgun metagenomic study in search of novel, plant cell wall degrading enzymes, discovery of which may be key to improving contemporary biochemical methods in the biofuel industry.

# Chapter 4. *A.ater* Gut Microbial Ecology Analysis Using Metagenomics

## 4.1 Abstract

Many herbivores are able to gain access to otherwise well-protected carbon sources in plant biomass by exploiting microorganisms in the environment, or those harboured in their digestive system. Although it is well established that gut microbes contribute to cellulose breakdown, the microbial composition of many of these organism's gut environments are still unstudied, mainly due to the only very recent development of next generation sequencing which can allow, for the first time, in-depth study of the entire gut microbiome. We have gathered evidence that suggests that the gut microbiome of *Arion ater* harbours a microbiome rich in microbes that facilitate the breakdown of lignocellulose. Using metagenomic techniques, we have begun to characterize the bacterial diversity of the gut microbiome of this notorious agricultural pest, analysing over 6 Gbp of gut metagenomic community sequences to reveal abundant populations of known lignocellulose-degrading bacteria, along with multiple bacterial plant pathogens. Here we report a gut microbiome that is dominated by members of the P*roteobacteria* phylum a trait observed in multiple herbivorous insects and snails. The vast majority of hits were attributed to a small number of genera with *Enterobacter, Citrobacter, Escherichia, Pseudomonas, Acinetobacter* and an unclassified genus belonging to the family *Sphingbacteriaceae* making up 73% of the microbiome. Despite this observed dominance, 375 species level identifications were made from amongst 253 genera, showing rich microbial ecology within the gut. Identification of microbiome by dominance of groups found

in high abundances in multiple other herbivore gut environments indicates that the A. ater gut could be well equipped for lignocellulose breakdown possibly using a similar core set of gut microbes as other plant eating eukaryotes. Here we also report the detection of multiple plant pathogen species including the species *Erwinia amylovora* and *Pectobacterium carotovorum* which have now been observed over two successive years, once during sampling for DGGE study in 2012 and during sample for a metagenomic study in 2014, indicating that the slug may have a role in overwintering of plant pathogens. This study demonstrates the importance of studying microbial communities in complex organisms, firstly with respect to understanding links between feeding, evolutionary success and microbial lignocellulose digestion and secondly to gain a better understanding of the potential of the eukaryote gut as a vector for economically important plant pathogens.

## 4.2 Introduction

It is now well established that the gut microbiome plays a pivotal role in digestion in many invertebrates and vertebrates, such as termites (Brune, 2014), cockroaches (Bertino-Grimaldi *et al.*, 2013), cattle (Hess *et al.*, 2011) and humans (Qin *et al.*, 2010). Much of the empirical evidence gathered to clarify this long standing hypothesis was collected using metagenomics. Metagenomics is described by Riesenfeld *et al.* (2004) as "the functional and sequence based analysis of the collective microbial genomes contained in an environmental sample". The name metagenomics means beyond genomics, where single genomes are studied, this is can be achieved using two methods; a traditional method of inserting environmental DNA fragments into fosmid vectors and testing clones for acquisition of a function or through shotgun high throughput sequencing analysis of environmental DNA (Gilbert and Dupont, 2011). Metagenomic shotgun sequencing of environmental samples can provide an accurate profile of the composition of an environments microbial community. Sequencing data produced in a metagenomic study can provide taxonomic data through assignment of sequences to the most plausible microbial lineage, in many cases with far greater accuracy that is achievable using 16s ribosomal markers in isolation (Segata *et al.*, 2012). The fact that the technology to carry out large scale shotgun sequencing studies is relatively new means that many environments of interest still remain unstudied. The gut microbiomes of members of the gastropod class are still largely unstudied, despite their ability to digest a wide range of materials efficiently. One recent study has demonstrated the ecological richness of the gut microbiome of the gastropod *Achatina fulica* (giant snail), highlighting its metabolic capabilities, with greater than 2,700 carbohydrate active enzymes being observed

(Cardoso *et al.*, 2012a). In chapter 3 it was demonstrated that the gut microbial consortium of *A. ater* has the functional potential to play an active role in breakdown of the lignocellulose portion of its diet (Joynson *et al.*, 2014), while showing that this activity is stable at a broad range of temperatures and pH levels. This leads us to believe that the gut environment of *A. ater* could harbour microbial consortia of considerable ecological and economic importance.

In this study the microbial ecology of the gut microbiome of *A. ater* was examined. There are three reasons why this research is important. First, knowledge of the gut microbiome composition of *A. ater* offers a means of understanding how this microbial population may facilitate the digestion of such a broad range of foodstuffs. Second, it may offer insight into the survivability and feeding ability of slug species. This is now especially important following the European Union ban on traditional molluscicide pellets, in force from September 2014 (Commission Implementing Regulation 187/2014). This was introduced because of the rapid build-up of traditional molluscicide metabolites in water sources (Kay and Grayson, 2013). Finally, the microbiological profile of the slug gut may also provide a target for future bacterial crop pathogen diagnostics, tracking, and control measures for use in agriculture, as slugs have recently been proposed as vectors for the transmission of bacterial pathogens (Gismervik *et al.*, 2014).

## 4.3 Methods

### 4.3.1 *A. ater* Gut Metagenomic DNA Extraction

Slugs were collected from a suburban area in North Cheshire (53.391463 N, 2.211214 W), two hours after last light. Individuals were cooled to 4°C to reduce spontaneous mucus production during dissection. Slugs were collected from the same area as those used in chapter 3 for gut microbe isolation and the DGGE study. Whole gut tracts were extracted, with care taken to avoid rupturing the gut wall, which would result in loss of gut juices. All dissections were carried out with each slug in a separate sterile petri dish. Ten gut tracts were cut into small pieces using fresh scalpel blades and, along with any gut liquid, were pooled and genomic DNA extracted using the soil extraction protocol of the Meta-G-Nome DNA isolation kit (Epicentre, WI, USA) described in section 3.3.2.1 with slight modifications. Samples were homogenised into an extraction buffer and large gut debris cleared using the centrifugation steps previously described. After centrifugation, because of the large number of guts used, an additional prefiltering step was carried out to prevent the 1.2 μm filter from being blocked. The resulting 50 mL homogenate cleared of debris, was filtered firstly through a 3 μm filter followed by filtration through a 1.2 μm filter using a Millipore vacuum based filtration system. These steps removed eukaryote derived debris from the host, plant biomass and fungi. The filtrate was finally filtered through a 0.45 μm filter in order to capture gut microbes while also allowing smaller debris such as free protein and free DNA present in the gut sample to flow through, leaving only microbial prokaryotes trapped on the filter paper. Microbes were washed from the filter paper and metagenomic DNA extracted as previously described in section 3.3.2.1. Given the importance of accurate quantification in shotgun DNA library

preparation, DNA samples were quantified by running 2 µL of metagenomic DNA alongside a quantified DNA ladder (HyperLadder I) and a quantified fosmid control DNA sample (100 ng/µL). Densitometry was then carried out using Gene Snap Tools (Syngene, UK), where the metagenomic DNA sample concentration was deduced by comparison with the pre-quantified fosmid control sample. Sample purity was also important for shotgun library preparation, which was measured using a nanodrop spectrophotometer, ensuring that the 260:280 ratio was ~1.8 and the 260:230 ratio was >1.5. After quality control checks 1 µg of metagenomic DNA was sent to the Centre for Genomic Research (CGR) at the University of Liverpool for preparation of a 2 x 250bp paired end shotgun DNA library (V2 chemistry) which was sequenced using a Miseq next generation sequencing platform (Illumina®).

### 4.3.2  Shotgun Metagenomic Sequencing Quality Control

Raw paired-end Illumina® sequencing reads were organised into two files, one containing the reads for one end of the DNA that was sequenced and a second file containing the sequence read at the opposite end of each of the DNA fragments sequenced. These files will be referred to as R1 and R2 where their description as inputs is required. Raw sequencing reads (fastq format) were trimmed for Illumina library adapter sequencings using Cutadapt (version 1.2.1) (Martin, 2011), set to default options and -O 3 which trimmed sequences matching adapter sequences for 3 bp or more at the 3' end of the sequence reads. Adapter trimmed reads were then filtered using Sickle (version 1.2) (Joshi and Fass, 2011). Reads shorter than 10 bp or with an average quality score of <20 were removed. The paired read associated with filtered sequences were also removed. Statistics were then generated for the trimmed and quality filtered files using ea-utils

(Aronesty, 2011). Further quality control analysis of the reads was done using FastQC (Andrews) in order to check for sequencing issues that may have occurred at specific points in the reads and to detect any biases that may affect biological conclusions. The sequencing files passed the quality control checks by FastQC and were then utilized in further analysis to elucidate the composition of the gut microbiome.

### 4.3.3  Phylogenetic Analysis of The Gut Metagenome Using MetaPhlAn

MetaPhlAn (Segata *et al.*, 2012) was used to estimate the phylogenetic composition of the gut microbiome. MetaPhlAn allows profiling of microbial communities through alignment of raw metagenomic shotgun library sequences against a database containing clade-specific marker genes. To maintain mapping accuracy, sequence data files R1 and R2 were input into MetaPhlAn in the .fastq format to allow inference of quality data when mapping. R1 and R2 where mapped by MetaPhlAn using Bowtie2 (Langmead and Salzberg, 2012) as follows:

```
$ cat R1.fastq R2.fastq > all.fastq | metaphlan.py all.fastq—
bowtie2db bowtie2db/mpa -bt2_ps sensitive —nproc 4 >
all_results.txt
```

Here, firstly the two read files were concatenated into a single .fastq file, which is then used as the input for the metaphlan.py script. The option –bt2_ps sensitive indicates the level of alignment sensitivity required while –nproc 4 allowed the script to use 4 processors. The .txt file output contained a list of the microbes identified in the sample down to species level along with an estimation of their percentage abundance in the sample.

In order to visualise the phylogenetic data produced, a phylogenetic tree containing data down to genus level was produced using GraPhlAn. The

MetaPhlAn output file was used as the input for the metaphlan2graphlan.py script. This produced a "tree" file, which contained the names and phylogenies of the microorganisms present, and an "annotation" file that contains information on the abundance of microbes at each taxonomic level. At this point the "annotation" file was modified to produce colouration of different taxonomic groups of biological relevance to this study. The tree and annotation files were then used as the input to the script graphlan_annotate.py which was used to coordinate the tree and annotation data into an .xml file which is then used to generate a phylogenetic tree using graphlan.py script. The commands for this are shown below.

```
$ metaphlan2graphlan.py metaphlan_output.txt –tree_file
tree_file.txt–annot_file annotation_file.txt
$ graphlan_annotate.py –annot tree_file.txt annotation_file.txt
annotated_tree.xml
$ graphlan –dpi 300 annotated_tree.xml phylogenetic_tree.png
```

This produced at phylogenetic tree which showed the microorganisms present in the gut metagenomic sample along with their relative abundance, as indicated by the size of the clade marker.

### 4.3.4  Phylogenetic Analysis of The Gut Metagenome Using MG-RAST

The R1 and R2 raw read files were also submitted to the online MG-RAST server (Meyer *et al.,* 2008) for analysis. The MG-RAST pipeline was used to perform further quality control steps followed by protein fragment prediction from the raw reads. These protein fragments were then submitted for similarity based alignment searches against databases including, GenBank KEGG, RefSeq, SEED, SwissProt and TrEMBL. The pipeline also searched the data for 16s rRNA sequences that were used for similarity searches using the 16s rRNA databases Greengenes and SILVA. A combination of the assignments of 16s rRNA genes

and the phylogeny of hits to which protein fragments mapped were then used as a basis for prediction of microbial ecology. In order to upload the sequence, the two sequence files R1 and R2 were interleaved using the shuffleSequences_fastq.pl that merged the sequences into a single file. This file was then uploaded through the online interface and submitted for analysis by the MG-RAST pipeline.

### 4.3.5  Identification and Analysis of Plant Pathogen Associated Sequences

Both methods of phylogenetic analysis undertaken gave results that revealed hits for a number of plant pathogen species. In order to elucidate the likelihood of these hits being false, further analysis was carried out on the reads that were predicted to be of plant pathogen origin. Reads that mapped to the pathogens from the genus *Dickeya* were extracted and aligned to the reference genomes of the bacterial species *D. Dadantii, D. Zeae, D. Solani and D. chrysanthemi*.

To gain further insight to whether the assignments were false or if the plant pathogens were present in the *A.ater* gut, the *Dickeya* associated reads were downloaded from the MG-RAST server in the .fasta format. In order to allow accurate mapping of the sequences, the original .fastq reads, which contain base calling quality data are required. The header from each sequence read present in each of the .fasta files downloaded from MG-RAST was then used to extract the corresponding read (and base calling quality data) from the original R1 and R2.fastq files using the python script fastq_mgrast_intersection.py (Written by Dr Leighton Pritcard, The James Hutton Insitute, UK). The resulting .fastq format sequences were then aligned to *Dickeya* reference genomes.

### 4.3.6 Alignment of Plant Pathogen Associated Reads to Reference Genomes

Reference genomes for the *Dickeya* species were provided by Dr Leighton Pritchard (The James Hutton Institute, UK). First, each reference genome was indexed in order to make them searchable by alignment software; this was done using Burrow-Wheeler Aligner (BWA) (Li and Durbin, 2009) as follows. To index the reference genomes BWA index was used:

```
$ bwa index reference_genome.fna
```

BWA mem function was then used to map the extracted reads from the gut metagenome (denoted by R1_extracted.fastq and R2_extracted.fastq) to the indexed reference genome.

```
$ bwa mem -t 4 reference_genome.fna R1_extracted.fastq
R2_extracted.fastq > alignment.sam
```

This created a Simple AlignMent file (SAM), which must then be converted to a Binary AlignMent format (BAM) that is smaller and therefore more easily visualised. Samtools (Li *et al.*, 2009a) view was used to convert the alignment file from SAM to BAM format.

```
$ samtools view -bS -T reference_genome.fna alignment.sam >
alignment.bam
```

Here Samtools view was used with the options –bS flagged, which indicates that the output should be in BAM format and the input is in SAM format. The -T option indicated the next file in the command was the reference genome from which the input SAM file was created. For visualisation the BAM file must be "sorted"; this places the reads within the file in the order they appear along the reference genome sequence. The BAM file was sorted using Samtools sort.

```
$ samtools sort alignment.bam alignment.sorted.bam
```

An index file for the alignment was then created to allow visualisation software to quickly locate reads that map to a certain region of the reference genome, which was created using the Samtools index function.

```
$ samtools index alignment.sorted.bam
```

Finally, a file containing the mapping statistics for the alignment was created using Samtools flagstat function.

```
$ samtools flagstat alignment.sorted.bam > mappingstats.txt
```

If the mapping statistics appeared to show a successful alignment, the sorted alignment BAM file was then visualised using the TABLET alignment viewer (Milne *et al.*, 2013).

## 4.4 Results

### 4.4.1 *A. ater* gut Metagenomic DNA Extraction

Gut metagenomic DNA was extracted as previously described. To ensure that the genomic DNA was intact and in order to quantify the sample it was ran on a 1% agarose gel alongside a fosmid control. This is shown in Figure 4-1.
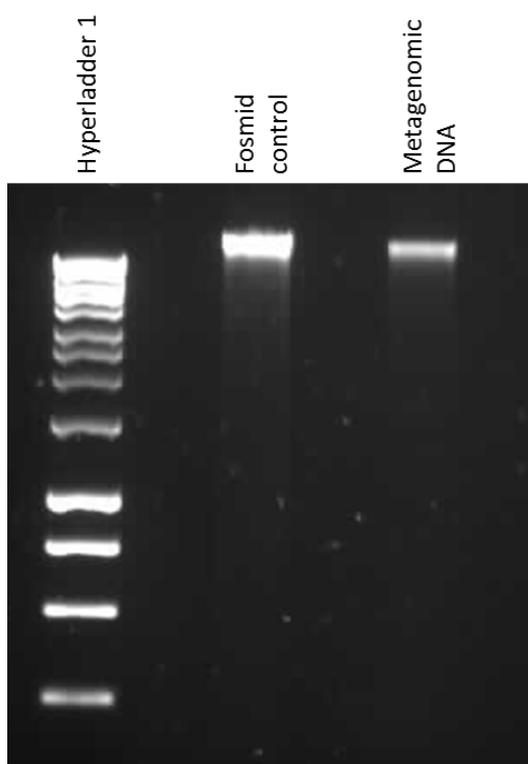


**Figure 4-1 A 1% agarose gel showing 2 µL of gut metagenomic DNA extract and 100 ng of fosmid control DNA**
*Lane 1 shows the HyperLadder I, lane 2 shows 100ng (1 µL) of a ~40 kbp fosmid control DNA sample, and lane 3 shows 2 µL of the gut metagenomic sample*

Figure 4-1 shows a single band with no smearing for the metagenomic sample indicated that the genomic DNA was intact and therefore viable for shotgun library preparation. The band seen for the fosmid control sample was used as the quantified standard for densitometry, from which the concentration of the metagenomic sample was deduced (~30 ng/µL). The 260:280 ratio was 1.83 and the 260:230 ratio was 1.56 for this sample.

### 4.4.2 Shotgun Metagenomic Sequencing Quality Control

The metagenomic DNA sample seen in Figure 4-1 was then used to create a 2x250 paired end shotgun metagenomic DNA library that was sequenced using an Illumina® Miseq a summary of the sequence statistics can be seen in Table 4-1 and the distribution of sequence lengths in Figure 4-2.

| Sequencing statistics | *A. ater* Gut Metagenome |
|---|---|
| Number of trimmed reads | 25,996,846 |
| Raw sequence data (Gbp) | 6.175 |
| Mean sequence length (bp) | 237 ± 34 |
| Mean GC content (%) | 51 ± 8 |

Table 4-1 The statistics of the raw shotgun metagenomic sequencing results

Table 4-1 indicates that the sequencing run was successful, showing that over 6 billion base pairs of data were produced in almost 26 million sequences. The mean sequence length indicates that the majority of the reads were almost sequenced to the maximum length possible of 250 bp. A more detailed distribution can be seen in Figure 4-2. The GC content of 51% demonstrates no obvious skews in the data towards genomes with extreme (high or low) GC contents.
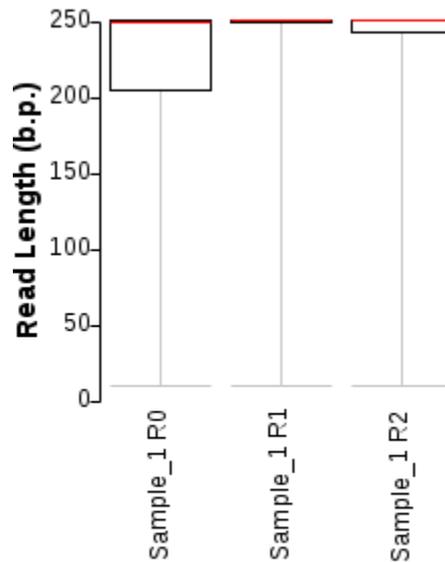
**Figure 4-2 A box plot showing the read length distribution of the raw sequencing reads**

The box plot in Figure 4-2 shows the length distribution of the forward reads (Sample_1R1), the reverse reads (Sample_1R2) and the reads whose paired-end counterpart reads were filtered during quality control (Sample_1R0). The red line on each plot denotes the median sequence length, the boxes show the interquartile range and the whiskers indicate the maximum and minimum read lengths present in each sequence file. The figure indicates that the vast majority of the reads in the forward and reverse sequence file were within a very short range of sequence lengths nearing the sequencing chemistry maximum of 250 bp. The sequence files R1 and R2 were then subjected to further quality control assessments using FastQC. The results of these analyses can be seen in Figure 4-3 -Figure 4-6.
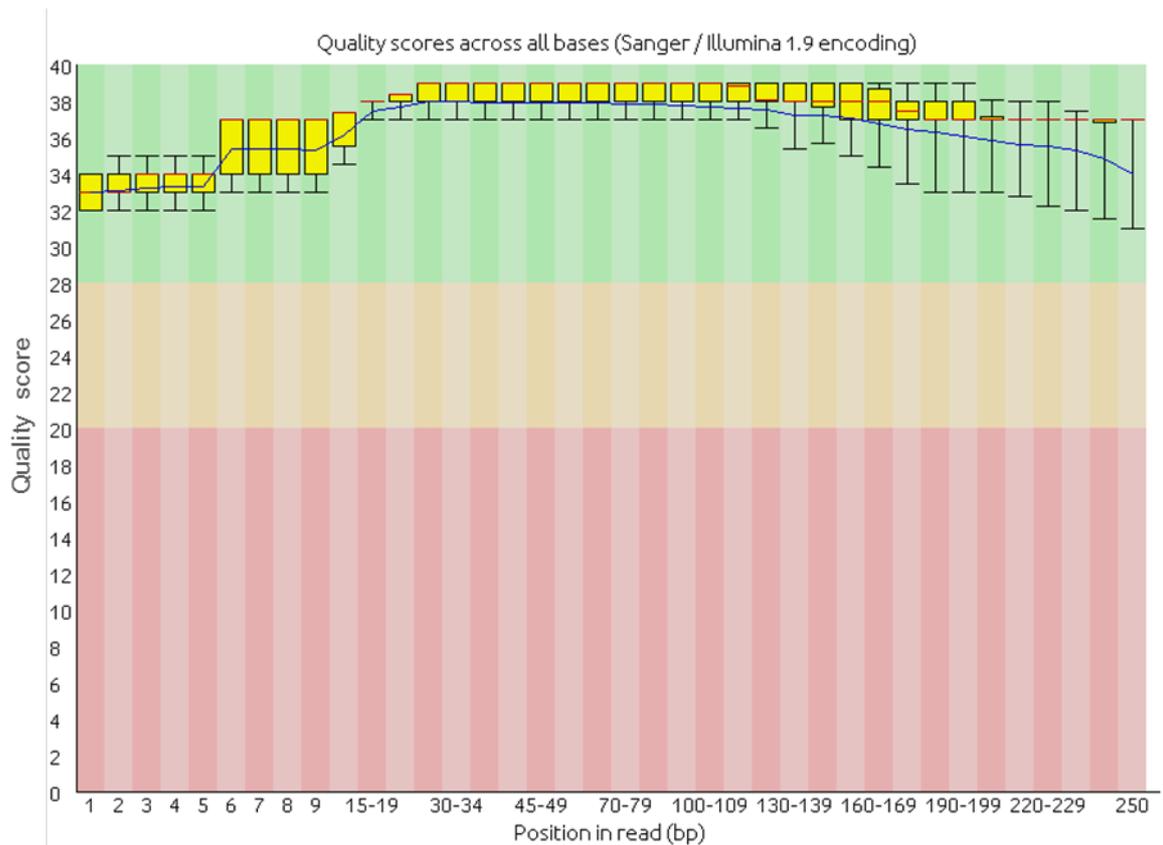
**Figure 4-3 A box plot showing the distribution of quality scores at each position in the raw reads found in the forward sequence file (R1.fastq)**

Figure 4-3 shows a boxplot showing the distribution of quality (Phred) scores (y axis) at different positions along the sequence reads in the forward sequence file (R1.fastq). This permits visual analysis that makes it possible to ensure that there was no sequencing run wide issue at any of the 250 sequencing cycles carried out. The plot shows upper and lower values of 39 and 30 (denoted by whiskers), where a score of 30 indicates a base call accuracy of 99.9% and a score of 40 indicates an accuracy of 99.99%. Within the first 9 base positions the quality scores appear between 32 and 37 with base positions 6-9 showing the largest interquartile ranges. The quality scores appear to be the highest and the most stable between positions 5-130, with the majority between scores of 36-38 with lower limits indicated by whiskers at a score of 37. Between bases 139-250 the difference between the upper and lower scores at each position increase steadily.

Is it known that this is the point at which the sequence quality scores begin to deteriorate using Illumina® as the continuous exposure to laser light at each cycle begins to damage the DNA strands. At all sequence positions, even the lowest quality scores were above the quality control threshold of 20 dictated by FastQC.
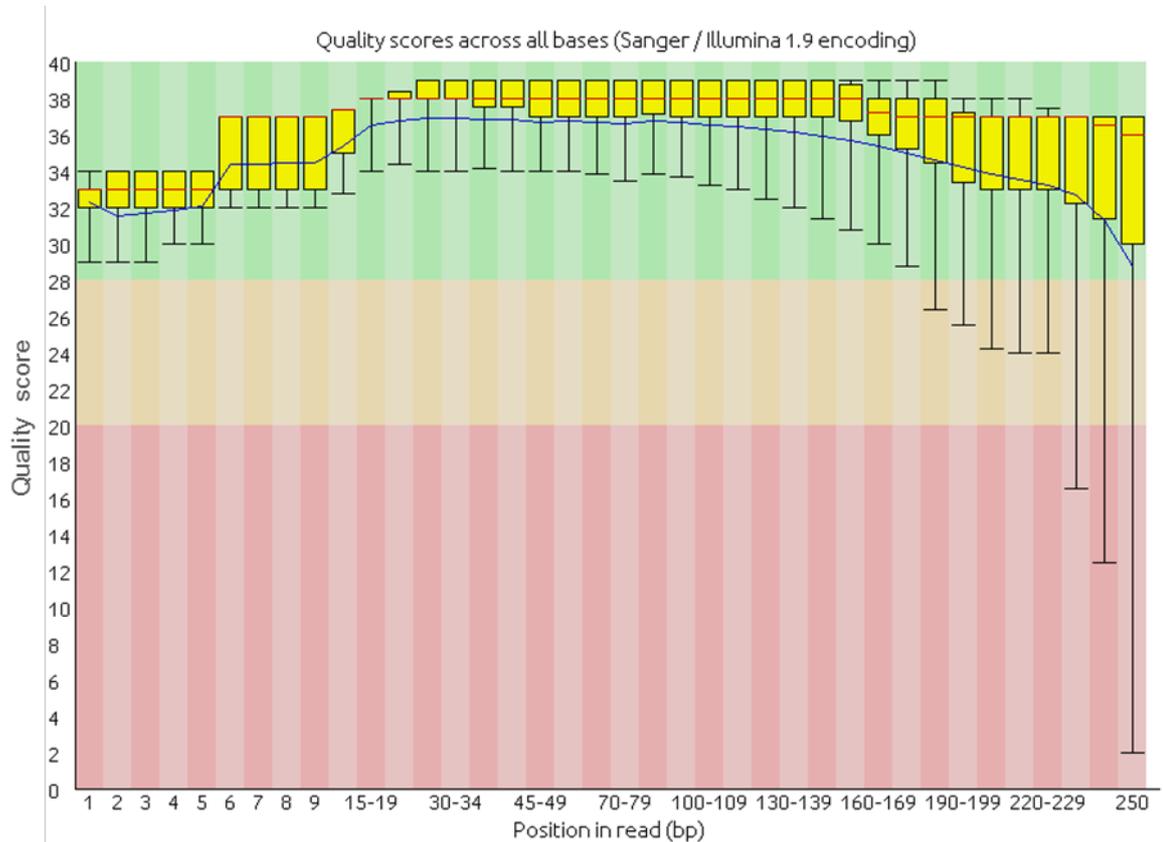


**Figure 4-4 A box plot showing the distribution of quality scores at each position in the raw reads found in the reverse sequence file (R2.fastq)**

Figure 4-4 shows a boxplot showing the distribution of quality (Phred) scores (y axis) at different positions along the sequence reads in the reverse sequence file (R2.fastq). The plot shows upper and lower quality score values of 39 and 2. Within the first 9 base positions there is a range of between 29 and 37 with relatively large interquartile ranges of between 37 and 31 for positions 6-9. Between positions 9 – 160 the interquartile ranges of the scores are stable indicating continual high quality base calling in these regions. However, a steady

decline in the lower limit of quality scores begins at position 100 and continues through to position 250. Although the lowest values at positions 229-250 appear to drop below the ideal quality score threshold (20), the vast majority of the bases called at these positions are still called with greater than 99.9% accuracy.
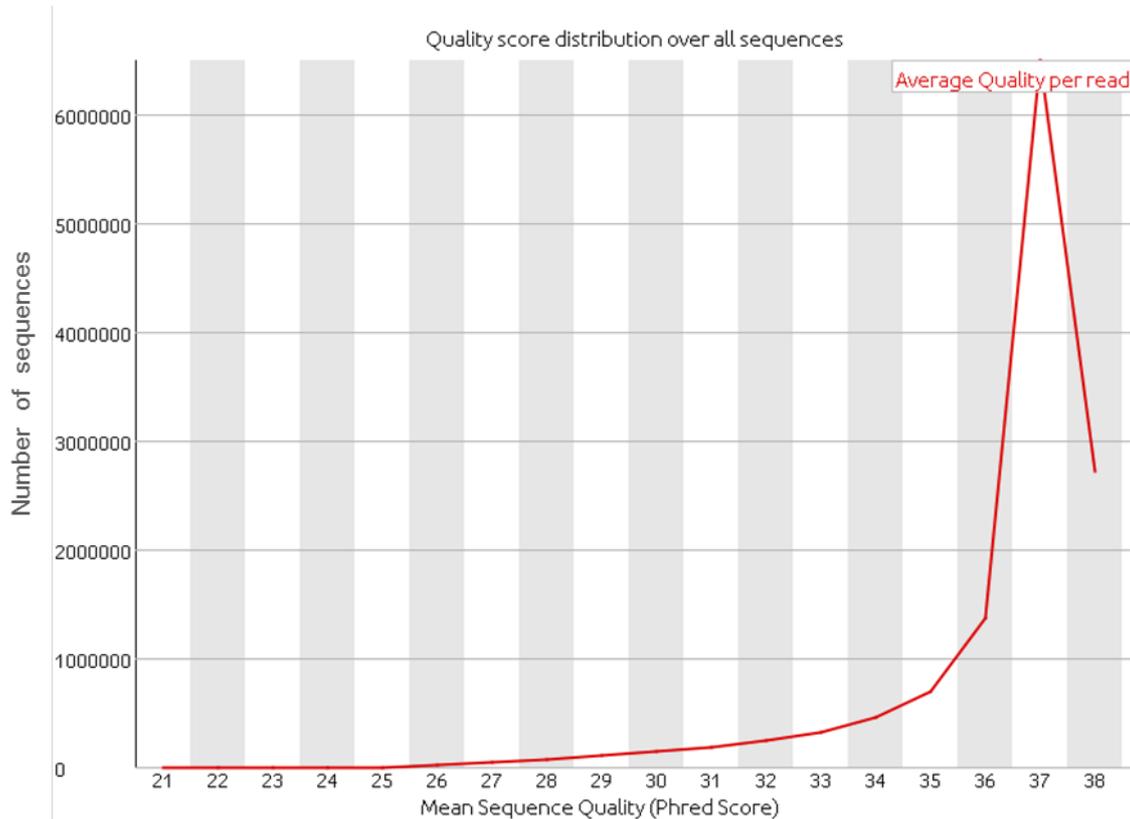


**Figure 4-5 The average quality score for each read in the forward sequence file (R1.fastq)**

Figure 4-5 shows the distribution of the quality score averaged over each entire sequence read for the forward sequences (R1.fastq).The graph shows that the majority of the sequence reads had average sequence quality scores of greater than 34, with the largest amount of sequences having an average quality score of 37.
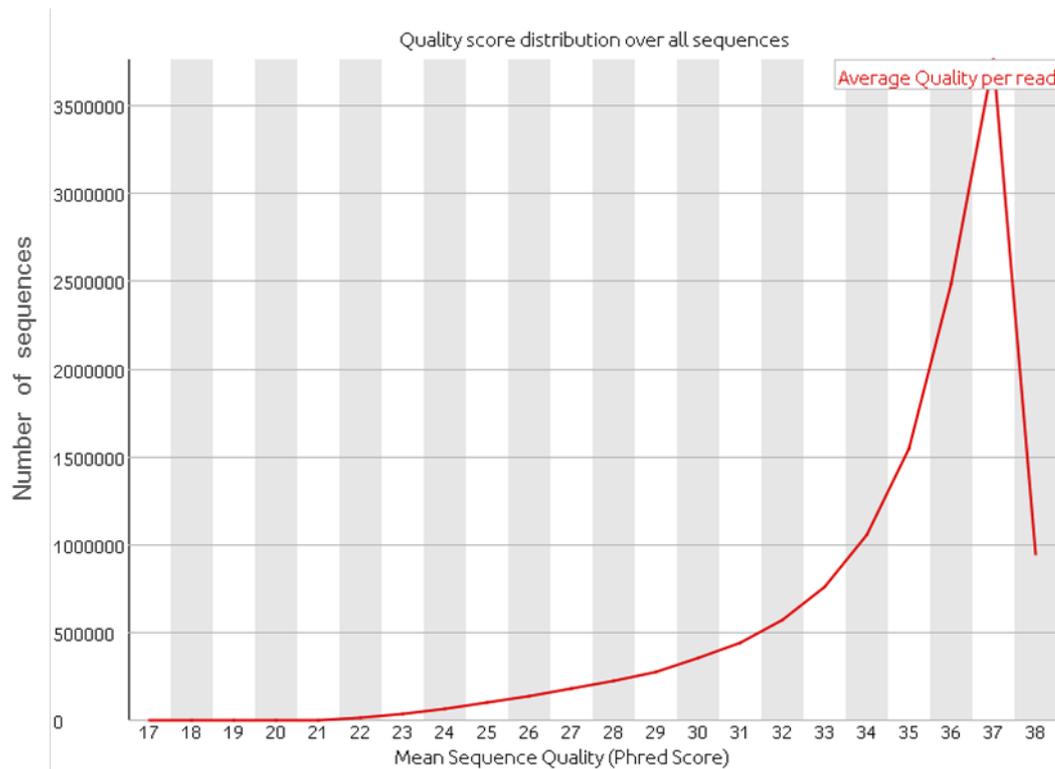
**Figure 4-6 The average quality score for each read in the reverse sequences (R2.fastq)**

Figure 4-6 shows the distribution of the quality score averaged over each entire sequence read for the reverse sequences (R2.fastq). The graph shows that the majority of the sequence reads had average sequence quality scores of greater than 28, with the largest amount of sequences having an average quality score of 37.

### 4.4.3 Phylogenetic Analysis of *A.ater* Gut Metagenome Using MetaPhlAn

After quality control analysis, the forward and reverse sequence files were subjected to phylogenetic analysis using MetaPhlAn. Some of the most abundant groups identified at each taxonomic level can be seen in Table 4-2.

| Classification | Percentage Abundance |
|---|---|
| k__Bacteria | 99.99% |
| k__Archaea | 0.01% |
| | |
| p__Proteobacteria | 88.15% |
| c__Gammaproteobacteria | 82.16% |
| o__Enterobacteriales | 64.56% |
| f__Enterobacteriaceae | 64.56% |
| g__Enterobacter | 26.86% |
| g__Citrobacter | 19.86% |
| g__Escherichia | 3.91% |
| o__Pseudomonadales | 14.25% |
| f__Pseudomonadaceae | 10.56% |
| g__Pseudomonas | 10.54% |
| f__Moraxellaceae | 3.69% |
| g__Acinetobacter | 3.68% |
| p__Bacteroidetes | 10.53% |
| c__Sphingobacteria | 8.57% |
| o__Sphingobacteriales | 8.57% |
| f__Sphingobacteriaceae | 8.56% |
| g__Sphingobacteriaceae_unclassified | 8.1% |
| p__Firmicutes | 0.59% |
| p__Actinobacteria | 0.28% |
| p__Chlamydiae | 0.21% |
| p__Chloroflexi | 0.16% |

**Table 4-2 Phylogenetic classification and microbial abundance of the most dominant microbes from the *A. ater* gut generated using MetPhAn**

Table 4-2 gives an indication of the most abundant groups of bacteria that appear to dominate the slug gut microbiome. The MetaPhlAn analysis showed that 99.99% of the hits from the alignments were allocated to the bacterial kingdom, with only 0.01% being attributed to archaea. Analysis revealed that the gut is dominated by members of the *Proteobacteria* phylum, with 88.15% of hits, and

82.15% of these being aligned to members of the *Gammabroteobacteria* class. Five of the six most abundant genera identified were from the *Gammanproteobacteria* class, 3 of which are members of the *Enterobacteriaceae* family (*Enterobacter, Citrobacter and Escherichia*). The second most abundant phylum seen is *Bacteroidetes* with 10.53% abundance; other phyla detected in smaller abundances include *Firmacutes* (0.59%) and *Actinobacteria* (0.28%). The full phylogenetic analysis results using MetaPhlAn can be seen in Appendix 2.

Figure 4-7 shows a phylogenetic tree containing clade markers for microbes identified at each taxonomic level down to genus. The clade marker size is directly proportional to the overall abundance of the group of microbes it represents.
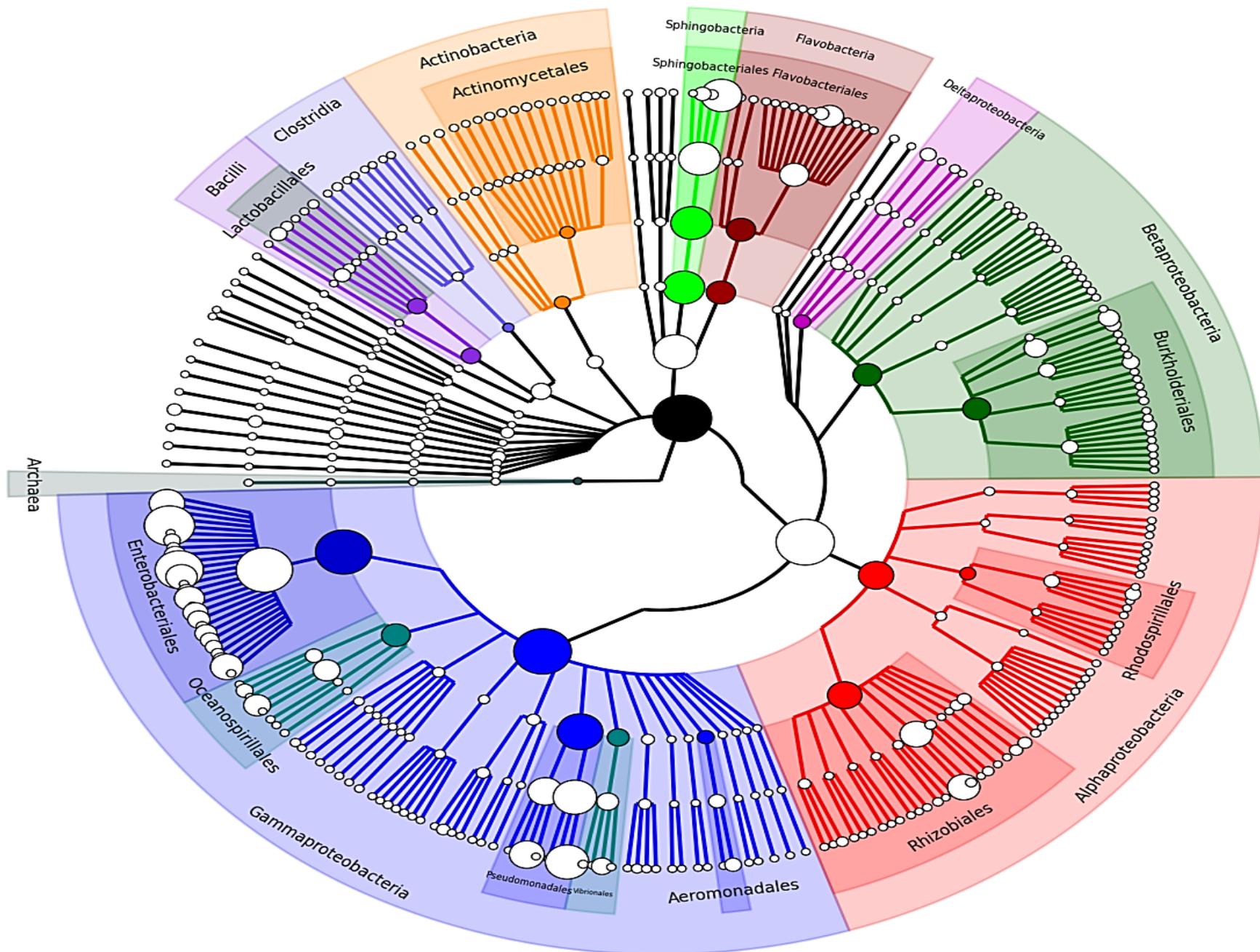
**Figure 4-7 A phylogenetic tree showing the diversity of the _A. ater_ gut microbiome (created using GraPhLan)**

The tree demonstrates the dominance of the *Proteobacteria* phylum, in particular with respect to the *Gammaproteobacteria* class. The figure also indicates the relatively large portion of the microbiome made up by *Sphingobacteria* and *Flavobacteria*, as well as demonstrating the large number of genera detected within the metagenomic sample, with 253 represented in total. MetaPhlAn also made 375 species specific identifications, but this was deemed too many to illustrate in a phylogenetic tree such as that seen in Figure 4-7.

### 4.4.4 Phylogenetic Analysis of The *A.ater* Gut Metagenome Using MG-RAST

For comparison with clade marker gene based identification of gut microbe phylogeny, sequence files were analysed using the MG-RAST pipeline, which uses overall protein fragment assignment to determine microbial abundance. A comparison of the abundances of the most common microbes identified by MetaPhlAn and MG-RAST can be seen in Table 4-3.

| Classification | MetaPhlAn percentage abundance | MG-RAST percentage abundance |
|---|---|---|
| k__Bacteria | 99.9% | 99.40% |
| Viruses | n/a | 0.30% |
| k__Eukaryota | n/a | 0.20% |
| k__Archaea | 0.01% | 0.10% |
| | | |
| p__Proteobacteria | 88.15% | 94.40% |
| c__Gammaproteobacteria | 82.16% | 91.27% |
| o__Enterobacteriales | 64.56% | 85.30% |
| f__Enterobacteriaceae | 64.56% | 85.19% |
| g__Enterobacter | 26.86% | 20.20% |
| g__Citrobacter | 19.86% | 15.10% |
| g__Escherichia | 3.91% | 10.50% |
| o__Pseudomonadales | 14.25% | 4.55% |
| f__Pseudomonadaceae | 10.56% | 3.40% |
| g__Pseudomonas | 10.54% | 3.30% |
| f__Moraxellaceae | 3.69% | 1.12% |
| g__Acinetobacter | 3.68% | 1.05% |
| p__Bacteroidetes | 10.53% | 2.96% |
| c__Sphingobacteria | 8.57% | 1.31% |
| o__Sphingobacteriales | 8.57% | 1.31% |
| f__Sphingobacteriaceae | 8.56% | 1.31% |
| g__Sphingobacteriaceae_unclassified | 8.1% | 8.10% |
| p__Firmicutes | 0.59% | 1.16% |
| p__Actinobacteria | 0.28 | 0.36% |

**Table 4-3 Comparison of the percentage abundances of the most common gut microbes identified using MetaPhlAn and MG-RAST**

Table 4-3 shows that the phylogenetic analysis by the MG-RAST pipeline produces similar patterns of dominance of microbial groups at each taxonomic level as is seen with MetaPhlAn. Indeed, MG-RAST suggests an even greater dominance of *Gammaproteobacteria* with 91.27% of allocations; MG-RAST also shows that the genera *Enterobacter, Citrobacter*, *Acinetobacter* and *Escherichia* are present in the sample in high abundance. Although there is a general consensus about the more dominant groups of microbes present in many cases

the difference in predicted abundance between MG-RAST exceeded 5% with the greatest difference seen for the order *Enterobactierales* where the difference in predicted abundance was greater than 20%.

Having previously used non-NGS based techniques to identify both cultured cellulolytic, microbes and identifications using DGGE, the outputs of the metagenomic phylogenetic analyses were also searched to determine if the microbes identified in previous studies were present in the gut of samples collected from the same location one year later, from which the sequenced metagenomic DNA was isolated. The results of these searches can be seen in Table 4 4.

| Name | Description | Percentage Match | Identified using MetaPhlAn | Identified using MG-RAST |
|---|---|---|---|---|
| Isolate 1 | *Citrobacter braakii* | 85% | Yes | Yes |
| Isolate 2 | *Salmonella enterica* | 91% | Yes | Yes |
| Isolate 3 | *Klebsiella pneumonia* | 96% | Yes | Yes |
| Isolate 4 | *Serratia marcescens* | 91% | Yes | Yes |
| Isolate 5 | *Buttiauxella agrestis* | 79% | Not in reference database | Yes |
| Isolate 6 | *Serratia liquefaciens* | 86% | Yes | Yes |
| Isolate 7 | *Aeromonas hydrophila* | 99% | Yes | Yes |
| Isolate 8 | *Acinetobacter calcoaceticus* | 92% | Yes | Yes |
| Isolate 9 | *Kluyvera intermedia* | 99% | No | Yes |
| Isolate 10 | *Buttiauxella agrestis* | 99% | Not in reference database | Yes |
| Isolate 11 | *Citrobacter freundii* | 99% | Yes | Yes |
| Isolate 12 | *Enterobacter sp. E6-PCAi* | 94% | Yes | Yes |
| UA.a.1 | *Mycoplasma hyorhinis* | 93% | No | Yes |
| UA.a.2 | *Mycoplasma iners* | 93% | Not in reference database | Yes |
| UA.a.3 | *Uncultured Citrobacter* | 99% | Yes | Yes |
| UA.a.4 | *Uncultured Serratia* | 100% | Yes | Yes |
| UA.a.5 | *Pectobacterium carotovorum* | 99% | Yes | Yes |
| UA.a.6 | *Acinetobacter beijerinckii* | 98% | No | Yes |
| UA.a.7 | *Pantoea sp. 57917* | 99% | Yes | Yes |
| UA.a.8 | *Erwinia amylovora* | 99% | Yes | Yes |
| UA.a.9 | *Erwinia tasmaniensis* | 99% | Yes | Yes |

**Table 4-4 Cross referencing previous microbial identifications using traditional methods with the phylogenetic analyses of the gut metagenome.**

**Table 4-4** shows that each of the microbes identified in chapter 3 using culture dependent methods and DGGE were also identified in the metagenomic sample. Three microbes were not identified by MetaPhlAn, *Acinetobacter beijerinckii,*

*Mycoplasma hyorhinis, Kluyvera intermedia* and two genomes of *Mycoplasma iners* and *Buttiauxella agrestis* are not in the reference genome database from which MetaPhlAn assigns phylogeny. All species previously observed were, however, identified in the analysis conducted using the MG-RAST pipeline.

The phylogenetic analyses carried out also obtained multiple hits for plant pathogen species, including many of the plant pathogens recently voted by experts in the field to be within the top 10 most important (economically and ecologically) bacterial pathogen species. A ranked list of the microbial pathogens identified within the *A. ater* gut can be seen in Table 4-5.

| Ranking | Pathogenic species | Micobiome abundance (%) |
|---------|-------------------|-------------------------|
| 1 | **Pseudomonas syringae** | **0.08264** |
| 3 | **Agrobacterium tumefaciens** | **0.06987** |
| 5 | **Xanthomonas campestris** | **0.0144** |
| 7 | **Erwinia amylovora** | **0.03587** |
| 9 | **Dickeya dadantii** | **0.04896** |
| 10 | **Pectobacterium carotovorum** | **0.04215** |

**Table 4-5 Microbiome abundance of plant pathogens present in the *A. ater* gut microbiome, as ranked by a survey of experts in the field**

Table 4-5 shows the plant pathogen species identified by MetaPhAn and their predicted abundance in the gut microbiome. In total, 6 of the "Top 10 plant pathogenic bacteria in molecular plant pathology" (Mansfield *et al.*, 2012) were identified.

In order to gain further insight into the presence of plant pathogens in the gut of *A. ater,* all reads from the metagenomic sample that aligned to members of the

pathogenic *Dickeya* genus were downloaded from the MG-RAST analysis (in the fasta format) and their corresponding reads (in the fastq format) were extracted from the original sequencing files as described in section 4.2.5. These reads were then aligned to reference genomes for *Dickeya solani, Dickeya dadantii, Dickeya chrysanthemi* and *Dickeya zeae.* The resulting alignment statistics can be seen in Table 4-6 and Table 4-7.

| Reference species | Dickeya solani | | | Dickeya dadantii | | |
|---|---|---|---|---|---|---|
| Reference genome ID | MK10 | IPO_2222 | MK16 | NCPPB_3537 | NCPPB_898 | NCPPB_2976 |
| *Dickeya* specific reads mapped (%) | 31.82%(85949 reads) | 31.71 (85668 reads) | 32.1% (86724 reads) | 27.18% (73379 reads) | 25.42% (68640 reads) | 32.84% (88751 reads) |
| Mapped reads properly paired (%) | 86.36 (74227 reads) | 86.45% (74062 reads) | 85.81% (74426 reads) | 84.03% (61664 reads) | 84.24% (57828 reads) | 84.98% (75428 reads) |
| Singleton reads (%) | 4.12% (11125 reads) | 4.09% (11041 reads) | 4.32% (11683 reads) | 3.88% (10486 reads) | 3.72% (10049 reads) | 4.52% (12214 reads) |
| Reads unique to species | 454 | | | 1641 | | |

**Table 4-6 Alignment statistics for alignment of *Dickeya* associated reads against reference genomes for *Dickeya solani* and *Dickeya dadantii* using BWA**

| Reference species | Dickeya chrysanthemi | | | Dickeya zeae | | |
|---|---|---|---|---|---|---|
| Reference genome ID | NCPPB_402 | NCPPB_516 | NCPPB_3533 | MK19 | NCPPB_3531 | CSLRW192 |
| *Dickeya* specific reads mapped (%) | 28.82% (77724 reads) | 30.01% (80975 reads) | 30.60% (82597 reads) | 25.28% (68318 reads) | 25.42% (68677 reads) | 25.73% (69455 reads) |
| Mapped reads properly paired (%) | 86.22% (67015 reads) | 86.04% (69675 reads) | 85.51% (70633 reads) | 82.90% (56641 reads) | 83.08% (57057 reads) | 84.11% (58421 reads) |
| Singleton reads (%) | 3.75% (10122 reads) | 3.90% (10512 reads) | 4.01% (10834 reads) | 3.82% (10328 reads) | 3.84% (10360 reads) | 3.85% (10385 reads) |
| Reads unique to species | 2309 | | | 2854 | | |

**Table 4-7 Alignment statistics for alignment of *Dickeya* associated reads against reference genomes for *Dickeya chrysanthemi* and *Dickeya zeae* using BWA**

Table 4-6 and Table 4-7 show the percentage of the total reads that were initially assigned to the *Dickeya* genus by MG-RAST that mapped back to each of the reference genomes for the 4 *Dickeya* species. The tables also show the percentage of reads that were "properly paired", which means that both the forward and reverse read for each pair ended "read pair" map to the same chromosome, oriented towards each other, and with a plausible insert size (number of base pairs between the two ends of the DNA fragment that was originally sequenced), which is an indicator of accurate alignment. Reads whose paired end counterpart read did not map to the reference genomes are counted as "singleton reads". For *D.chrysanthemi, D.zeae* and *D.solani* species the percentages of reads that mapped back were similar with only a 0.45% difference between the three *D.zeae* genome alignments, 0.72% difference between *D.solani* species and 1.7% difference between *D.chrysanthemi* species. The difference in percentage of reads that mapped back to the *D.dadantii* species genomes was greater, at 7.2% between the NCPPB_898 and NCPPB_2976 genomes. The number of reads that only mapped back to a single species was also calculated in order to ascertain whether allocations down to species level can be accurately assigned. The highest number of "species specific" reads was found for *Dickeya zeae* (2854) and the least for *Dickeya solani* (454).

## 4.5 Discussion

The black slug, *A. ater,* has become one of the most widespread and successful gastropod species in Europe and North America. The success of this (and other) species and its ability to eat a very broad range of foodstuffs has caused the UK agricultural industry alone to spend almost £30 million pounds each year on molluscicide pellets (Agular and Wink, 2005). This makes it a very important species in agro-economical terms but also exemplifies the species' incredible ability to survive and thrive despite society's best efforts to control it.

Research on the digestive system of *A. ater* began in the 1960s, focusing on both carbohydrate breakdown (Evans and Jones, 1962a) and protease activity (Evans and Jones, 1962b). Further research determined rates of cellulose breakdown and characterised the pH and temperature profiles of gut fluids from black slugs of North American origin (James *et al.*, 1997). In a previous study the cellulolytic activity in the gut of the British black slug was determined (Joynson *et al.*, 2014) and in chapter 3 multiple gut microbes that exhibit cellulolytic activity were identifed. This work implicated the gut microbiome in degradation of plant cell wall components (that make up a large portion of the slugs diet) into simple sugars that may be metabolised by the slug or utilized by the microbes themselves as they grow, which in turn could lead to increases in the number of microbes, leading to greater production of other molecules such as short chain fatty acids that may also be utilized by the slug. In this study the *A. ater* gut microbiome was explored by means of metagenomics. Metagenomic DNA was extracted from the slug gut and successfully sequenced using next generation sequencing methods. The sequencing reads produced offer an insight into the microbial ecology of the gut environment

through association of DNA sequence reads with annotated DNA data of known origin available in vast online databases.

This study has revealed the presence of an ecologically rich consortium of bacterial species that have previously been implicated in the digestion of tough vegetation. The bacterial community of the gut microbiome was dominated by a relatively small number of groups. These included the genera *Enterobacter*, *Citrobacter*, *Pseudomonas*, *Escherichia*, *Acinetobacter* and an unclassified genus belonging to the *Sphingobacteriaceae* family. These genera alone accounted for almost three quarters of the sequenced component of the gut metagenome (Table 4-2). This finding was surprising, as it was expected that the broad feeding habits of the slug species would lead to many microbes from a wide spectrum of environments being acquired, which in turn would lead to the creation of a highly diverse gut microbial community. The dominance of a handful of microbial genera may suggest a distinct core community of bacteria that reside in the slug gut or, alternatively, that this composition may reflect the prevailing composition of the material consumed prior to dissection and the microbial consortium that it hosts. The latter circumstance has been seen in cockroach gut microbiomes (Bertino-Grimaldi *et al.*, 2013) where changes in diet were associated with the acquisition or increased abundance of microbial groups that are predicted to have greater ability to break down the new food type.

Although it appears that a very small number of genera make up the majority of the microbiome, a large number of genera were detected in much lower abundances. With over 200 genera account for only ~27% of the microbiome, many of these microbes may simply be transient elements of the gut microbiome that are ingested during proximal feeding or suppressed by nutritional cycling in the gut at a particular

148

time. It is also a possibility that some of these genera that appear in extremely low abundance could be artefacts of false prediction or noise. This can be especially problematic with methods used by MG-RAST which works off a "top hit" allocation methods where sequences are aligned to multiple entire databases and phylogeny linked to the hit from the name of the microbe which is at the top of the list, often not taking into account others that may have the same level of identity with the query sequence. This makes species and genera level assignment difficult using this method, but also allows identification of less well represented species. The methods of MetaPhlAn are thought to provide a more accurate picture of microbial ecology through its more conservative marker gene based identifications, where genes thought to specific to each taxonomic level are identified and used in a database for sequences to be queried against (Segata *et al.*, 2012).

In total, Gammaproteobacteria accounted for the vast majority of the community metagenome, with 82% relative abundance; this included identification of 84 species in this class. Previous studies have shown dominance of the phylum Proteobacteria in the gut microbiomes of various gastropod species, including freshwater planorbid snails (Biomphalaria pfeifferi) and terrestrial snails such as the giant African land snail (Achatina fulica) (Cardoso et al., 2012b). Proteobacteria have also been seen to dominate gut microbiomes of insects, whose diets are largely or entirely comprised of lignocellulose (Dillon and Dillon, 2004, Russell et al., 2009a), which suggests a broad association of this phylum not only with herbivorous insects but also with plant-eating gastropods.  Furthermore, two studies of microbial consortia in fungal gardens used by leaf cutter ants (Atta colombica) to degrade lignocellulose both identified dominance by members of the family Enterobacteriaceae. These studies predict that bacteria of this family would be directly involved in the efficient

breakdown of plant material in these gardens (Aylward et al., 2012, Suen et al., 2010). This study shows similar Enterobacteriaceae family dominance (~65% of the community metagenome). These findings are also consistent with previous culture dependent identification of cellulolytic microbes from the A. ater gut, where all microorganisms identified were from the Gammaproteobacteria class. Here we also reconfirm the presence of all of the microbes previously observed by identification of each species again in the metagenomic sample sequenced here that was isolated from the gut of slugs sampled in the same area a year later (Table 4-4).

These findings suggest that the gut environment of *A. ater* contains a consortium that is well equipped to digest complex plant-derived carbohydrates, including lignocellulose that are thought to make up the majority of the diet of *A. ater* and also suggests that some of the microbes present in the gut are not transient, instead having a constant presence.

Mining of the phylogenetic data associated with the gut microbiome enabled identification of several bacterial plant pathogens. These included six species recently ranked among the top 10 most important species of plant pathogens (Mansfield *et al.*, 2012) (Table 4-5). Many of these pathogens are known to cause necrosis and eventual development of soft rot, blight or blackleg in tuber based crops such as potatoes, but also in ornamental plants and other crops. These include the three relatively closely related *Enterobacteria Dickeya dadantii, Pectobacterium carotovorum* and *Erwinia amylovora (Toth et al., 2011).* If these pathogen species, which are renowned for their suite of plant cell wall degrading enzymes, are commensally present in the slug gut, this would suggest that *A. ater* may act as a perpetual vector species, allowing them to spread from field to field, and persist between growing seasons.

If plant pathogenic bacteria are not commensal in the gut, then they may still be spread opportunistically over a limited geographical range if they are eaten by slugs and eventually excreted. Previous studies of the agricultural impact of black slugs on food crops has shown that they are capable of feeding on a wide variety of potato species (Airey, 1987). Arion species are noted to feed frequently on mature damaged tubers, which allows the slugs to avoid the toxic, alkaloid containing skin (Keiser *et al.*, 2012). These tubers that are damaged are known to be particularly susceptible bacterial pathogens. The slug feeding process, and the excrement left on or near crops by slugs could contribute to the spread of bacterial plant pathogens, while overwintering in the slug gut could provide a possible means of pathogen survival during winter. The role of insects in the transmission and overwintering of plant pathogens is now quite well established, with the squash bug, flea beetle and cucumber beetle all known to spread plant pathogens as well as sustaining populations of the pathogens they harbour during dormant winter months (Nadarasah and Stavrinides, 2011). The fact that the pathogens *Pectobacterium carotovorum* and *Erwinia Amylovora* were identified over two successive years in slugs from the same area provides support for this hypothesis but more work is required to confirm the potential role of slugs in plant pathogen overwintering.

These findings demonstrate a gut microbiome that resembles that of many other lignocellulose feeding organisms using analysis of raw metagenomic sequencing reads. However to understand the metabolic capability of the gut microbiome we must also uncover the repertoire of proteins that the microbes present in the slug gut are able to produce. The metagenomic sequencing results can provide, along with insight into microbial ecology, a wealth of information into the metabolic capabilities

of the microbes present that will enable us to assess the microbial contribution to the degradation of each of the components of lignocellulose.

# Chapter 5. Functional Analysis of The Slug Gut Microbiome: The Use of Metagenomics to Identify Novel Carbohydrate Active Enzymes

**Abstract**

In previous chapters multiple cellulolytic bacteria have been identified in the *A.ater* gut. A metagenomic study also indicated that the gut microbiome comprises bacterial groups that dominate many herbivorous guts and lignocellulose degrading environments, such as leaf cutter ant fungi farms. The aim of this study was to begin to characterize the metabolic capabilities of the gut microbiome through prediction and annotation of open reading frames derived from an assembled metagenome. Here the bioinformatics analyses were validated by amplification and sequencing of a selection of predicted carbohydrate active enzyme genes, after which one was expressed as a recombinant protein and tested as to whether its function matches what was predicted. This study revealed a repertoire of 5,635 carbohydrate active proteins including 2,510 genes corresponding to glycoside hydrolase enzymes, 561 carbohydrate-binding modules and 312 from groups associated with lignin breakdown. This study also describes the creation of large amounts of whole genome amplified (WGA) metagenomic DNA, samples from which 5 novel CAZyme genes were amplified, thus validating the bioinformatics analyses. This included a predicted β-glucosidase gene (9459) that was expressed and its predicted gene function verified. The vast range of enzyme classes identified show that the microbial consortium is capable of degrading all components of lignocellulose, including cellulose, hemicellulose, pectin and lignin. Together, these functions could allow *A. ater* to make extensive use of plant biomass as a source of nutrients. In this study CAZyme catalogue has been created containing thousands of enzymes that can be

amplified in future studies from large stocks of WGA metagenomic DNA, allowing

exploitation of the large number of genes of biotechnological interest discovered.

## 5.1  Introduction

The development of next generation sequencing based metagenomics over the past decade has revolutionised microbiology by facilitating high resolution culture-independent characterization and exploitation of complex microbial communities. Traditional metagenomic analysis methods involved construction of DNA fosmid libraries containing fragments of metagenomic DNA. However discoveries made using these methods are limited because clone acquisition of a specific type of activity could only be identified through agar plate or liquid culture assay of tens of thousands of clones. This greatly limited the number of activity types that the metagenomic libraries could be screened for. This limitation is overcome by using next generation sequencing (NGS) methods. NGS-based metagenomics allows genes of all known functionalities to be identified and function assigned using homology searches against databases containing gene sequences for tens of thousands of gene functions.  NGS-based metagenomics has been employed to study biological activity in a wide array of environments, including hot springs (Jiménez *et al.*, 2012) glacier ice (Simon *et al.*, 2009) soil (Vogel *et al.*, 2009) ocean waters (Rusch *et al.*, 2007) and eukaryote gut environments including humans (Huttenhower *et al.*, 2012). Metagenomics has also been used specifically to search for novel carbohydrate active enzymes for use in the biofuel industry, in so called "bioprospecting" studies (Li *et al.*, 2009b). Many environments have been targeted, such as leaf cutter ant fungi farms (Aylward *et al.*, 2012) and thermophilic "sludge" (Xia *et al.*, 2013), with the majority focusing on the gut microbiomes of lignocellulose feeding eukaryotes. To date studies focussing on CAZyme identification have been carried out on the gut microbiomes of giant pandas (Zhu *et al.*, 2011), buffalo (Duan *et al.*, 2009), cattle (Hess *et al.*, 2011), wallabies (Pope *et al.*, 2010), reindeer (Pope

*et al.*, 2012), termites (Warnecke *et al.*, 2007) and snails (Cardoso *et al.*, 2012a). Here we characterize the metabolic capability, with respect to lignocellulose breakdown, of the gut microbiome of one of the most common land gastropods in Europe and North America, *Arion ater*.

## 5.2 Methods

*A. ater* gut metagenomic DNA was extracted, sequenced and the quality of the resultant sequences assessed, as described in chapter 4. To gain an insight into the metabolic capability of the slug gut, the short sequence reads were assembled into longer sequences called "contigs" from which protein sequences can be predicted.

### 5.2.1 Assembly of Metagenomic DNA Sequencing Reads Using MetaVelvet

The 25,996,846 short sequencing reads that passed quality control were assembled initially using Velvet (V1.2.10) (Zerbino and Birney, 2008). *De novo* genome assembly using Velvet has two stages; the first involves collation of the data set by splitting the sequence reads into "K-mers" where the value of K is equal to the number of DNA bases in each "K-mer word", which is predetermined by the user. This was achieved using the velveth function as follows.

```
$ velveth ./output_directory 51 –fastq –shortPaired –separate
R1.fastq R2.fastq
```

This command specified selection of a K-mer value of 51 and indicated the file format and sequence type of the input files. The "hash table" produced then was used as the input for velvetg using the following options.

```
$ velvetg ./output_directory -exp_cov auto -ins_length 200
```

This command specified the expected coverage value of the assembly to be automatically detected (a requirement of the Metavelvet programme) and that there was an average of a 200 bp "insert" between the paired-end reads. This constructed a De Bruijn graph using the K-mers organised by velveth. This graph was then manipulated to construct longer sequence assemblies. The velvet extension Metavelvet (v1.2.01) (Namiki *et al.*, 2012) was then used to improve the assembly by

making the assumption that multiple genomes are present. The metavelvetg function was run using the following options.

```
$ metavelvetg ./output_directory -ins_length 200 –min_contig 800
```

Metavelvet was set to assume an insert length of 200 and to remove any contig from the resultant contig fasta format file that was smaller than 800bp. Assembly statistics were deduced using GNX tools.

```
$ gnx-tools -min 100 -nx 25,50,75 meta-velvetg.contigs.fa
```

### 5.2.2 Raw Read- Assembly Realignment

As a diagnostic measure, the raw reads from sequence files R1.fastq and R2.fastq were aligned to the assembled contigs created using Metavelvet. To do this BWA was used as described in section 4.3.5.1 but the assembled contigs fasta file was indexed and used as the reference genome. The resulting SAM file was converted to the BAM format, sorted and indexed using Samtools as previously described. The indexed BAM file was then visualised using TABLET alignment viewer.

### 5.2.3 Open Reading Frame Prediction From Assembled Contigs

The programme MetaGeneMark (v.3.26) (Zhu *et al.*, 2010) was used to predict gene sequences from assembled contigs. Firstly, the gmhmmp tool was used with MetaGeneMark_v1.mod to predict open reading frames using the fasta file output from the MetaVelvet assembly as the input file.

```
$ gmhmmp -a -d -f G -m MetaGeneMark_v1.mod -o open_reading_frames.gff
```

This command creates a file containing gene predictions with both nucleotide protein sequences along with information on the contig from which each open reading frame was derived. The nucleotide and amino acid sequences were extracted from the gmhmmp output using the following commands.

```
$ nt_from_gff.pl open_reading_frames.gff > nucleotides.fasta

$ aa_from_gff.pl open_reading_frames.gff > proteins.fasta
```

The protein sequences predicted from the assembled contigs were then used as queries against protein databases in order to annotate the open reading frames.

## 5.2.4 Initial Search For Carbohydrate Active Enzymes Through Pfam Annotation

In order to determine whether it was justified to submit the predicted protein sequences for computationally expensive and time consuming BLAST annotation, the open reading frames were associated with protein families (Pfam). To do this, all predicted amino acid sequences were used as a query in the CAZYmes Analysis Toolkit (CAT) (Park *et al.*, 2010) using the Pfam based annotation tool with an E-value threshold of x$10^{-4}$ against the CAZy database (Lombard *et al.*, 2014).

## 5.2.5 Protein Annotation Using BLASTp and Annotation Exploration Using MEGAN4

In order to assign the function of the predicted proteins in greater detail, the predicted open reading frame amino acid sequences were used as query for a local alignment search using the BLASTp algorithm (Altschul *et al.*, 1990) against the NCBI non-redundant protein database (downloaded March 2014).

```
$ blastp -query proteins.fasta -db ./databasedirectory/nr -out
sampleannotations.blastp -evalue 0.00001 -num_alignments 50 -
num_descriptions 50 -num_threads 12
```

This command sets the algorithm to query each protein against the non-redundant protein database, noting the top 50 hits for each protein sequence. The annotations in the output .blast format file were then input into the MetaGenome ANalyzer 4 (Megan4) (Huson *et al.*, 2011). The predicted genes were sorted into groups based

on the BLAST results and the biochemical pathways annotated in the KEGG database using the KEGG extension in the MEGAN4 software.

## 5.2.6 Amplification of Putative CAZyme Genes

To validate the *de novo* assembly and gene predictions made during bioinformatics analysis, a selection of CAZYme sequences were targeted for amplification, cloning and sequencing to confirm whether the predicted sequences occurred in nature. The small sample size of the metagenomic DNA sample that was sequenced limited the number of PCR amplifications that could be attempted. To circumvent this limitation the metagenomic DNA sample was subjected to whole genome amplification.

### *5.2.6.1 Whole Genome Amplification of Metagenomic DNA*

In order to increase the amount of metagenomic DNA template available for PCR reactions, metagenomic DNA from the sequenced sample was subjected to whole genome amplification (WGA) reactions. 10ng of metagenomic sample DNA was used as template for amplification using the Repli-G mini kit (Qiagen, UK). The kit employs an isothermal genome amplification methodology, called Multiple Displacement Amplification (MDA), where hexamer oligonucleotides act as random primers for amplification of up to 70 kbp fragments by Phi29 polymerase. Briefly, 10 ng of metagenomic DNA was incubated in the alkaline buffers provided in the kit for 3 minutes to denature the double stranded DNA. Phi29 polymerase and reaction buffer were added to the DNA sample, gently mixed by vortexing, and then incubated at 30°C for 16 hours.

Whole genome amplified DNA was then cleaned using the following protocol that utilizes buffers and spin columns of the DNeasy Plant Mini Kit (Qiagen, UK). Firstly, 200 µL of nuclease free water was added to the 50 µL WGA reaction mixture. To

this, 375 µL of buffer AP3 was added and the solution was mixed by pipetting. The entire solution was then transferred to a DNeasy spin column followed by centrifugation for 1 minute at 10,000 x g. The flow through was discarded and 500 µL of buffer AW was added to the column as a cleaning step followed by centrifugation for 1 minute at 10,000 x g. This wash step was then repeated, then after the flow through was discarded, any residual ethanol was removed by further centrifugation at 16,000 x g for 2 minutes. The column was then transferred to a 1.5 mL Eppendorf tube to which 50 µL of buffer AE (heated to 65°C) was added. The column was incubated at room temperature for 5 minutes and then centrifuged for 1 minute at 10,000 x g; the eluent was then labelled as the 1st elution and placed on ice. To ensure maximum recovery of WGA DNA, the elution step was repeated with a further 50 µL which was labelled as the 2nd elution. The DNA concentration of both eluents was assessed using a nanodrop spectrophotometer and structural integrity assessed by running the DNA on a 1% agarose gel. This produced between 4-6 µg whole genome amplified product per 10ng of metagenomic DNA starting material.

### 5.2.6.2  *Predicted CAZyme Gene Selection and Primer Design*

In order to validate the metagenome assembly, six predicted gene sequences for various carbohydrate active enzyme groups were selected for amplification from the WGA metagenomic DNA sample. Genes were selected following inspection of the BLAST alignments in the KEGG extension of MEGAN4. Genes were chosen to represent CAZyme families that are involved in the degradation of cellulose, hemicellulose and lignin in an attempt to strengthen our conclusions (made from gene predictions) that the slug gut microbiome has the ability to breakdown the most abundant and biotechnologically important super structures in the plant cell wall. Due to the nature of metagenomics, the vast majority of full length genes identified were

from more common species, to that end genes were selected from the top three most abundant genera (*Enterobacter*, *Citrobacter* and *Pseudomonas*. See table Table 5-6). Two genes were also selected from the *Raoultella* genus, which was not detected by MetaPhlAn, suggesting that, if truly present, genes annotated to this genus are likely to be novel genes from relatively abundant uknown bacteria. The ID numbers of the protein sequences selected from MEGAN4 were used to extract the corresponding nucleotide sequences from the open reading frame prediction output file. Primers were then designed to amplify the predicted gene sequences. Primer sequences can be seen in Table 5-1. All genes selected were full length except gene 77908.

| Gene ID number/Primer name | Primer sequence | Annealing temperature | Putative function | Predicted size |
|---|---|---|---|---|
| Gene_ID_8282F | ATG TCA CTT ATT CAG AAC CCT G | 53°C | Xylanase | 1680bp |
| Gene_ID_8282R | ATC AAA ACG TGA TTC GCT CGC | | | |
| Gene_ID_71437F | ATG ATG CGT CCA GCC GGT T | 53°C | Cellulase | 1197bp |
| Gene_ID_71437R | TAG CGT GTG ACG GCG CAT | | | |
| Gene_ID_3165F | AGT AAA GAA GCG ATT AAA CGC G | 53°C | FAD-Oxidase (lignin) | 1557bp |
| Gene_ID_3165R | TTG ACC TGC CTG ATG CGA A | | | |
| Gene_ID_77908F | CTC TGG ACG GGC ATG ATG | 57°C | Cellulase | 363bp |
| Gene_ID_77908R | TGA GAA CTT GCG CAT TCC TG | | | |
| Gene_ID_9459F | ATG AGA TAC CGT TTT CCT GAA | 53°C | β-Glucosidase | 1383bp |
| Gene_ID_9459R | ATC GAA TCC ATT ATT GGC GG | | | |
| Gene_ID_13418F | ATG TCC TTG CGT GCT TTA GTC | 57°C | Cellulase | 993bp |
| Gene_ID_13418R | TCA GAC ACC GGT AGC TGC | | | |

**Table 5-1 The primer sets designed for amplification of the 6 selected CAZyme gene sequences**

### 5.2.6.3 *Amplification, Cloning and Sequencing of Selected CAZyme Predicted Genes*

WGA metagenomic DNA was used as template for PCR using the primer sets inTable 5-1. The components and volumes of the PCR reaction mixtures can be seen in Table 5-2 and the PCR cycling conditions inTable 5-3.

| Component | Amount |
|---|---|
| DreamTaq mix | 25 µL |
| Forward Primer | 1 µL(0.2 µM) |
| Reverse Primer | 1 µL (0.2 µM) |
| WGA Metagenomic DNA | 100 ng |
| Nuclease-free water | Up to 50 µL |
| Total volume | 50 µL |

**Table 5-2 The PCR reaction mixture for amplification of predicted protein sequences from the WGA metagenomic DNA template**

| Stage | Temperature | Time (MyTaq®) | Cycles |
|---|---|---|---|
| Initial denaturation | 95°C | 1 minute | 1 |
| Denaturation | 95°C | 30 seconds | |
| Annealing | (see **Table 5-1**) | 30 seconds | 35 |
| Elongation | 72°C | 1 minute | |
| Final | 72°C | 1 minute | 1 |

**Table 5-3 PCR cycling stages, temperatures and times for amplification predicted CAZyme genes**

PCR products were ran on 1% agarose gels. Amplicon bands corresponding to the predicted gene sizes were cut out of the gel and DNA was extracted from the gel slices using the Wizard® SV Gel and PCR Clean-Up System (Promega, UK) and cloned into the pCR®2.1 cloning vector using methods described in section 3.3.1.6.

Vector inserts were then sequenced using the Bigdye v3.1 system method also described in section 3.3.1.6. The resulting sequence files were then submitted as a query to an online BLAST search against the NCBI nucleotide collection database to determine if the putative function for the amplified gene was the same as the putative function of the predicted gene from the metagenomic assembly. The amplicon sequence was then aligned to its corresponding predicted gene sequence in order to confirm amplification of the targeted gene.

### 5.2.7 Expression of β-glucosidase Gene 9459

After successful amplification of multiple predicted CAZyme gene sequences the gene 9459, a predicted β-glucosidase, was cloned into an expression vector and the recombinant protein tested for predicted activity. PCR was carried out using the same reaction mixture shown in Table 5-2 and cycling conditions in Table 5-3 using 10ng of the pCR®2.1 vector containing the gene 9459 as template DNA. The PCR product was ran on a 1% agarose gel and the band corresponding to the ~1500bp 9459 gene was cut from the gel and extracted using the Wizard® SV Gel and PCR Clean-Up System (Promega, UK) (method described in section 3.3.1.6). The PCR product was then cloned into the pBAD TOPO TA Expression vector (Life Technologies, UK). Briefly, 10 ng of extracted PCR product was incubated with the TOPO TA cloning reaction mixture for 20 minutes at room temperature. These closed vectors were transformed into TOP10 competent *E. coli* (Life Technologies, UK) as follows. Firstly, two LB agar plates containing 100 µg/mL ampicillin were made. Plates were equilibrated for 30 minutes at 37°C. One 50 µL vial of One Shot® TOP10 competent cells was thawed on ice for each transformation. 2 µL of the cloning reaction mixture was then placed into a vial of competent cells and mixed gently, without pipetting, and incubated on ice for 30 minutes. The remaining cloning

reaction was stored at -20°C. The competent Cells were then heat shocked at 42°C by submerging the vial in a water bath for 30 seconds, taking care not to agitate the vials at this point. The vials were then placed back on ice and 250 μL of super optimal broth with catabolite repression (S.O.C) medium was added to each vial. Vials were then shaken horizontally in a shaking incubator at 225 rpm, 37°C for 1 hour. After incubation the content of the vial was plated onto two preheated agar plates, 50 μL on one plate and 200 μL on a second plate to ensure at least one plate with colonies with adequate spacing to allow them to be selected. Agar plates were incubated overnight at 37°C. A single colony was selected from each set of transformants and grown in 5 mL of LB broth containing 100 μg/mL ampicillin overnight at 37°C, alongside a control containing uninoculated LB broth with 100 μg/mL ampicillin. To create a frozen stock of the clone, 0.85 mL of culture was mixed with 0.15 mL of glycerol and stored at -80°C.

In order to ensure the insert was in frame in the vector, the plasmid insert sequence was sequenced using the pBAD forward and pBAD reverse sequencing primers using the BigDye v3.1 system method described in section 3.3.1.6.

The gene was then expressed as described in the pBAD topo TA cloning kit manual. Briefly, 10 mL of LB broth containing 100 μg/mL ampicillin was inoculated with 0.1 mL of overnight culture. The culture was then incubated at 37°C, 225 rpm until it reached an $OD_{600}$ of ~0.5. As a time zero control sample, 1 mL of culture was removed and the cells pelleted by centrifugation at 10,000 x g for 1 minute. The pellets were then stored at -20°C. To induce expression of the protein L-arabinose was added (The pBAD promoter contained in this vector is L-arabinose inducible) to the remaining 9 mL of culture to a final concentration of 0.2% followed by incubation at 37°C at 225 rpm for 4 hours. After incubation two 1 mL aliquots were removed and

cells pelleted as previously described and stored at -20°C for use in SDS PAGE and Western blot analysis. In order to identify β-glucosidase activity, 10 μL of the induced culture was plated onto an esculin hydrate (0.1% w/v)-ferric ammonium citrate (0.03% w/v) arabinose (0.2%) growth assay agar plate. The plate was incubated for 24 hours at 37°C.

### 5.2.8 Western Blot Analysis

Firstly 12% SDS PAGE gels were created as described in section 2.3.2 (minus addition of CMC). Bacterial pellets collected before and after expression induction were resuspended in 80μL of 1x SDS PAGE sample buffer and boiled for 5 minutes. Samples were loaded, with 5 μL and 10 μL of both the positive 4 hour induced sample and time 0 sample, alongside 15μL of SeeBlue® Plus2 Pre-Stained Standard (Invitrogen, UK). Electrophoresis was conducted at 100v for 4 hours. Filter paper was moistened in Western blot transfer buffer (25 mM Tris base, 190 mM glycine 20% methanol), the SDS PAGE gel was placed on the filter paper and Amersham Hybond ECL Nitrocellulose membrane (GE Healthcare, UK) (previously moistened by capillary action in transfer buffer) was placed on to the gel (ensuring no air bubbles were present between the membrane and the gel surface) followed by another piece of filter paper. The filter paper-gel pile was then placed into the gel sandwich holder and placed into the SDS PAGE tank with the membrane closest to the negative electrode. The tank was then filled with Western blot transfer buffer (precooled to 4°C) and the proteins transferred from the gel to the nitrocellulose membrane by electrophoresis at 100v for 1 hour. Following transfer, the nitrocellulose membrane was washed with TBST buffer (10 mM tris base, 150 mM NaCl, 0.1% Tween 20, pH 7.5), dried between filter paper and stored at room temperature overnight.

The nitrocellulose membrane was re-wet in TBST the following day, rehydrated using capillary action by dipping one edge in the solution. The membrane was then transferred to a tray containing blocking buffer (TBST, 1% non-fat milk), covered and incubated on a rocking platform for 1 hour. The blocking solution was poured off and the membrane submerged in primary antibody solution (TBST, 0.5% non-fat milk, 1:2000 dilution of mouse anti6xHis tag monoclonal antibody, Sigma Aldrich, UK) and incubated at room temperature for 2 hours. The membrane was then washed with TBST for 5 minutes (repeated 3 times). The membrane was then submerged in secondary antibody solution (TBST, 0.5% non-fat milk, 1:2000 dilution of goat anti mouse horse radish peroxidase (HRP) conjugate monoclonal antibody (Sigma-Aldrich, UK) and incubated at room temperature for 1 hour. The membrane was then washed 3 times using TBST and placed onto piece of Saran wrap. Chemiluminescence detection stages were carried out using the Amersham ECL prime Western blotting detection reagent kit (GE Healthcare, UK), 1.5 mL of detection solutions A and B were mixed and pipetted evenly onto the membrane and incubated for 5 minutes at room temperature. Excess detection reagent was poured off and the edge of the membrane blotted against tissue paper. The membrane was then visualised in a G:Box Chemi transilluminator (Syngene, UK) (exposure time 4 minutes).

## 5.3 Results

### 5.3.1 *De Novo Assembly and Open Reading Frame Prediction*

Raw sequence reads were assembled using Metavelvet and open reading frames were predicted from the resultant sequence contigs produced. Assembly statistics and the number of protein coding genes predicted are shown in Table 5-4.

| | *A. ater* Gut Metagenome |
|---|---|
| **Number of trimmed reads** | 25,996,846 |
| **Raw sequence data (Gbp)** | 6.175 |
| **Number of assembled contigs** | 48,089 |
| **Largest contig (Kbp)** | 56.3 |
| **N50 value (Kbp)** | 1.8 |
| **Protein coding genes** | 108,691 |
| **Total size of metagenome (Mbp)** | 81.74 |

**Table 5-4 The Sequencing and assembly statistics of the gut community metagenome**

The resulting assembled metagenome contained 81.74 million base pairs of sequencing data. The longest assembled sequence present in the metagenome was 56.3 thousand base pairs long and the metagenome had an N50 value of 1.8 thousand base pairs. This statistic is a weighted median where 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value. In total, 108,691 open reading frames were predicted from the assembled contigs using MetaGeneMark.

### 5.3.2 Functional Annotation of Predicted Proteins

These open reading frames were then submitted as a query for protein family annotation. The allocations made to glycoside hydrolase (GH) groups with activities against some plant cell wall polysaccharides can be seen in Table 5-5.

| pFam group | Predominant Activity | Human | Termite | Wallaby | Panda | Snail | Slug |
|---|---|---|---|---|---|---|---|
| **Cellulases** | | | | | | | |
| GH5 | cellulases | 7 | 125 | 27 | 1 | 36 | **15** |
| GH6 | endoglucanases | 0 | 0 | 0 | 0 | 4 | **0** |
| GH7 | endoglucanases | 0 | 0 | 0 | 0 | 0 | **0** |
| GH9 | endoglucanases | 0 | 43 | 5 | 0 | 15 | **11** |
| GH44 | endoglucanases | 0 | 0 | 0 | 0 | 0 | **0** |
| GH45 | endoglucanases | 0 | 6 | 0 | 0 | 0 | **0** |
| GH48 | cellobiohydrolases | 0 | 0 | 0 | 0 | 2 | **0** |
| Total | | 7 | 174 | 32 | 1 | 57 | **26** |
| **Endohemicellulases** | | | | | | | |
| GH8 | endoxylanases | 2 | 21 | 2 | 1 | 46 | **11** |
| GH10 | endo-1,4-β-xylanase | 2 | 102 | 19 | 1 | 25 | **16** |
| GH11 | xylanase | 0 | 19 | 0 | 0 | 1 | **0** |
| GH12 | endoglucanase & xyloglucanase | 0 | 0 | 0 | 0 | 0 | **12** |
| GH26 | β-mannanase & xylanase | 1 | 20 | 8 | 0 | 11 | **0** |
| GH28 | galacturonases | 3 | 15 | 10 | 0 | 69 | **6** |
| GH53 | endo-1,4-β-galactanase | 11 | 20 | 11 | 4 | 9 | **276** |
| Total | | 19 | 197 | 50 | 6 | 161 | **321** |
| **Xyloglucanases** | | | | | | | |
| GH16 | xyloglucanases | 1 | 6 | 6 | 6 | 12 | **117** |
| GH17 | 1,3-β-glucosidases | 0 | 0 | 0 | 0 | 2 | **60** |
| GH81 | 1,3-β-glucanases | 0 | 0 | 0 | 0 | 1 | **0** |
| Total | | 1 | 6 | 6 | 6 | 15 | **177** |
| **Debranching enzymes** | | | | | | | |
| GH51 | α-L-arabinofuranosidases | 15 | 13 | 19 | 2 | 22 | **3** |
| GH62 | α-L-arabinofuranosidases | 0 | 0 | 0 | 0 | 2 | **0** |
| GH67 | α-glucuronidase | 1 | 6 | 1 | 2 | 5 | **1** |
| GH78 | α-L-rhmnosidase | 13 | 7 | 46 | 1 | 73 | **8** |
| Total | | 29 | 26 | 66 | 5 | 102 | **12** |
| **Oligosaccharide degrading enzymes** | | | | | | | |
| GH1 | mainly β-glucosidases | 54 | 27 | 94 | 41 | 294 | **118** |
| GH2 | β-galactosidases | 29 | 32 | 39 | 4 | 66 | **60** |
| GH3 | mainly β-glucosidases | 55 | 109 | 101 | 11 | 219 | **86** |
| GH29 | α-L-fucosidases | 7 | 12 | 5 | 0 | 70 | **11** |
| GH35 | β-galactosidase | 4 | 7 | 8 | 1 | 32 | **14** |
| GH38 | α-mannosidase | 6 | 18 | 3 | 8 | 18 | **39** |
| GH39 | β-xylosidase | 2 | 13 | 3 | 8 | 6 | **279** |
| GH42 | β-galactosidases | 15 | 33 | 17 | 7 | 54 | **6** |
| GH43 | arabinases & xylosidases | 34 | 63 | 72 | 13 | 185 | **28** |
| GH52 | β-xylosidase | 0 | 3 | 0 | 0 | 0 | **0** |
| Total | | 206 | 317 | 342 | 93 | 944 | **641** |

**Table 5-5 Comparison of the glycoside hydrolase (GH) profiles in the gut metagenomes of humans, termites, wallabies, giant pandas, snails and the black slug, showing GH groups that are implicated in breakdown/modification of plant cell wall polysaccharides**
*(Modified from Cardoso et al. (2012a))*

Table 5-5 compares the number of CAZymes identified in groups involved in the breakdown of cellulose and hemicellulose from metagenomic studies of the gut microbiomes of termites, wallabies, pandas, humans, the snail, *Achatina fulica* and our study of the *Arion ater* gut microbiome. The gut environment of the slug appears to contain a similar number of cellulose-degrading enzymes to both the snail and wallaby, with a similarly high number of oligosaccharide degrading enzymes in both molluscs. However, the slug gut environment contains many more enzymes targeting hemicellulose than any of the comparator organisms. In addition to these CAZymes multiple enzymes involved in the enzymatic breakdown of lignin were also identified.

The Pfam assignment search identified 5,635 predicted genes with hits against the Cazy database. This included 2,510 genes corresponding to glycoside hydrolase activity and 561 carbohydrate-binding modules. The majority of the carbohydrate-active genes identified were linked with enzyme groups that break oligosaccharides down into simple sugars (641 enzymes, 20.8%), with fewer targeting cellulose (26 enzymes, 0.85%). This search also identified 312 members of the relatively new CAZyme classes "Auxiliary activities" or AA classes, which describes enzyme classes that act on or consort (e.g. interact) with lignin in their activities (Levasseur *et al.*, 2013). This included 150 members of the class AA3, 2 members of AA2, 11 members of AA4, which are involved in the oxidative degradation of lignin, and 60 members of class AA6, which catalyse reductive degradation of aromatic compounds such as the monolignols that make up lignin superstructure.

The predicted proteins were also queried against the NCBI non-redundant (nr) database (03/02/14) using the BLASTp algorithm. Of 108,691 sequences submitted to the BLAST search, 97,882 were assigned matches in the nr database (~90% of

total predictions). Using the KEGG extension of MEGAN, over 32,000 functional associations were made within the KEGG pathways, of which 8,333 were assigned to carbohydrate metabolism. Using the KEGG extension of MEGAN, multiple proteins were observed that were assigned to phosphotransferase systems (PTS) which allow internalization of many sugars into bacteria. These included 109 proteins that make up the three subunits of the PTS that facilitates specific internalisation of cellobiose. The KEGG diagram showing protein groups assigned to PTS subunits can be seen in Figure 5-1
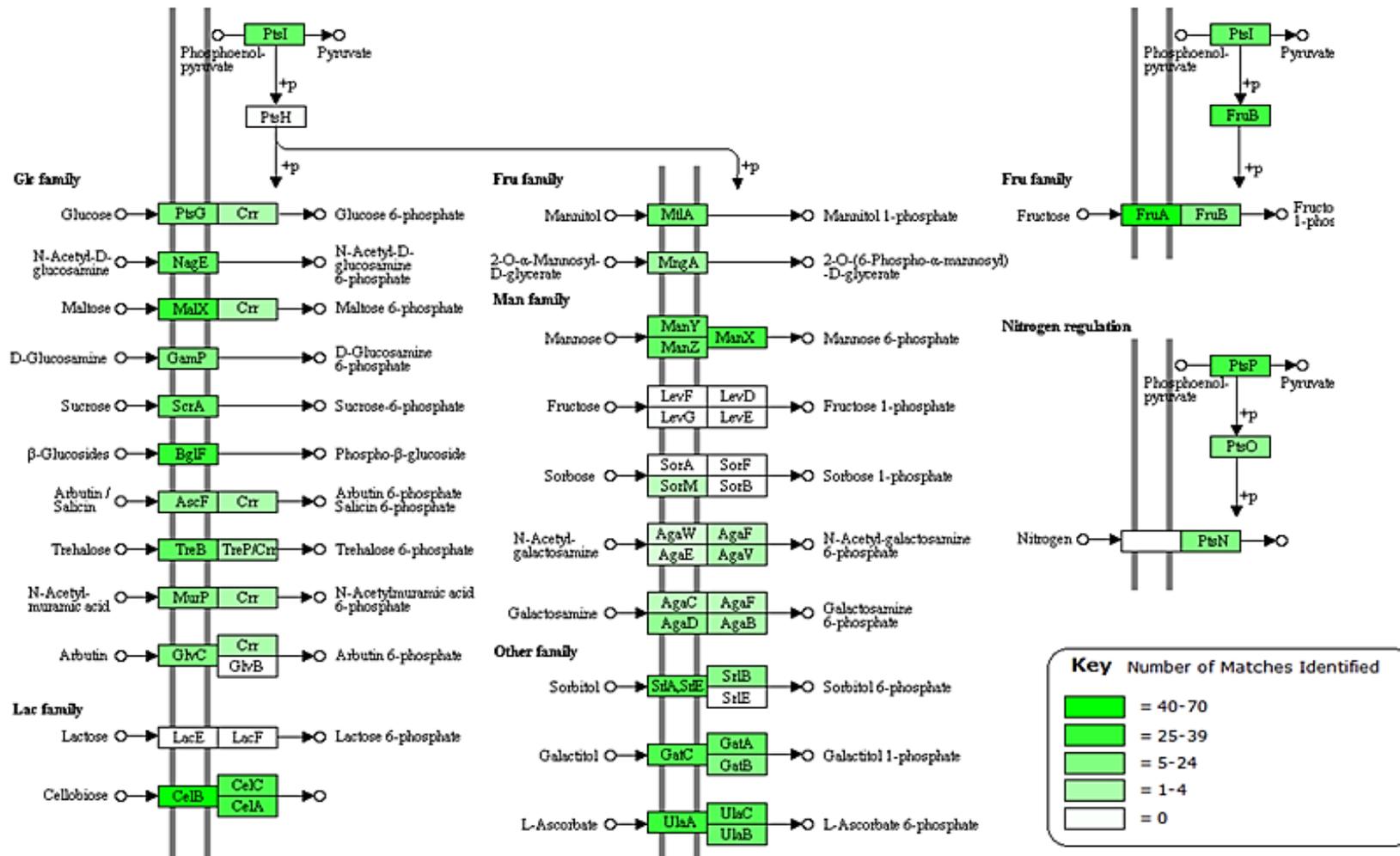
**Figure 5-1 A KEGG diagram showing the phosphotransferase system (PTS), genes identified in the gut metagenome are highlighted in green with colour intensity corresponding to abundance observed (created in MEGAN4)**

### 5.3.3 Whole Genome Amplification of Gut Metagenomic DNA

In order to attain enough metagenomic DNA for a large number of PCR reactions, metagenomic DNA was subjected to whole genome amplification reactions. Figure 5-2 shows the first sample subjected to WGA.
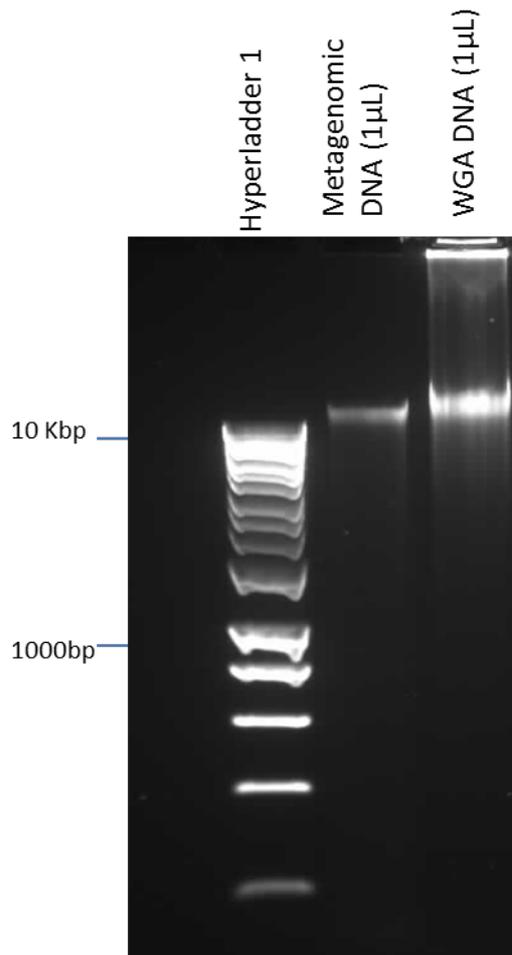


**Figure 5-2 A 1% agarose gel showing successful whole genome amplification of gut metagenomic DNA**

Figure 5-2 shows 1 μL of the original metagenomic sample used as template for WGA alongside 1 μL of WGA sample, created using only 10 ng of metagenomic DNA as template. The gel shows that DNA of lengths comparable to the original sample were successfully amplified using the WGA method. Further WGA runs

were then carried out to create a stock of metagenomic DNA for future PCR amplification applications (Figure 5-3)
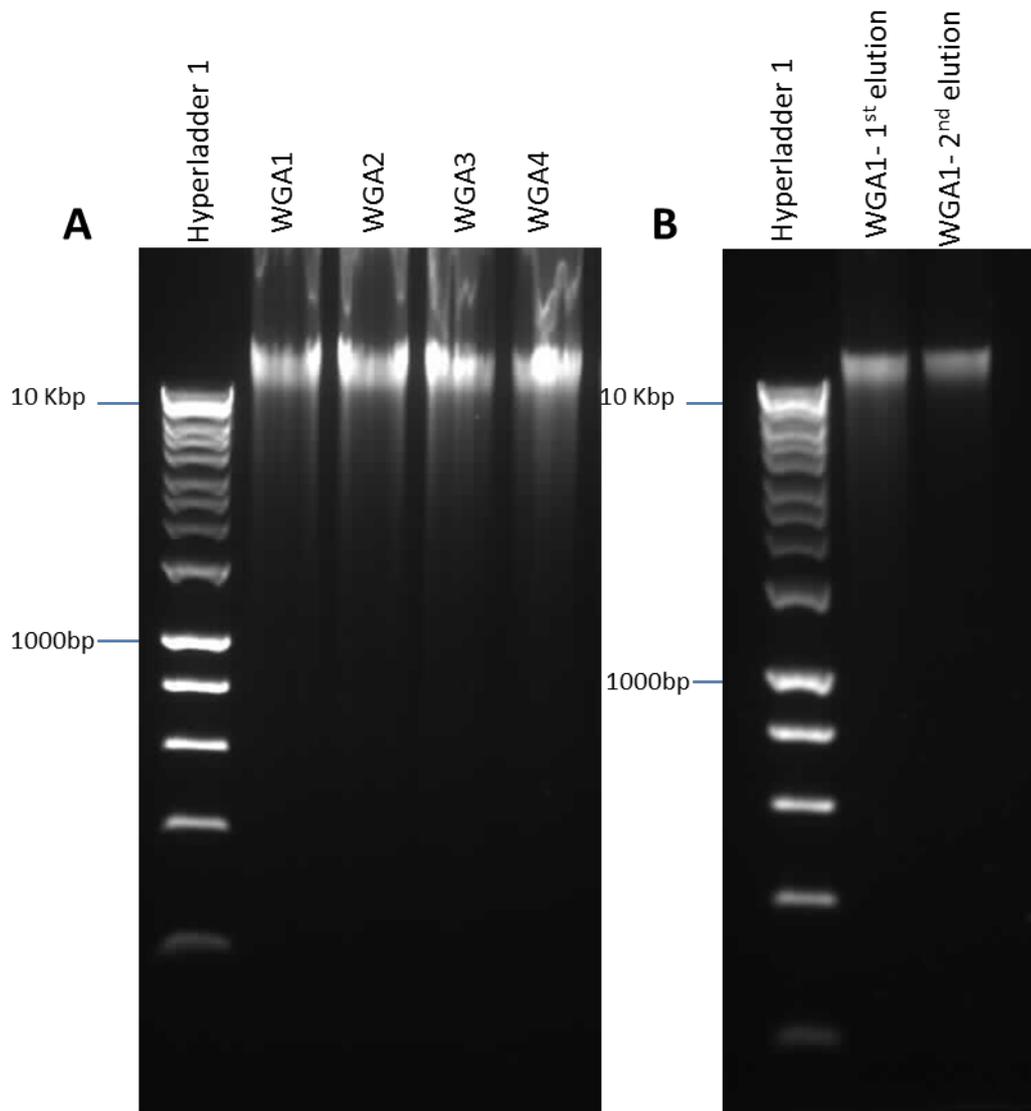


**Figure 5-3 (A) Shows a 1% agarose gel analysis of 4 further WGA reactions (B) shows the sample WGA1 after cleaning steps**
*(A) shows successful WGA in a further 4 samples, where 10 ng of metagenomic DNA was used as template. Gel (B) shows whole genome amplified DNA after clean up and two step elution*

Figure 5-3 confirms that both the WGA reactions were successful and that the clean-up process does not cause degradation of the large DNA fragments. In total, ~4.5 µg of DNA was produced per WGA reaction, each from 10 ng of starting material.

### 5.3.4  PCR Amplification and Sequencing of Predicted Genes From WGA Metagenomic DNA

CAZyme genes of interest were selected for amplification based on the BLAST analysis of the predicted open reading frames. A list of these can be seen in Table 5-6.

| Gene ID | Function/ Closest database match | Percentage match | Accession of hit | CAZy family |
|---|---|---|---|---|
| 9459 | 6-phospho-beta-glucosidase [*Citrobacter freundii*] | 96% | KGZ29936.1 | GH1 |
| 13418 | MULTISPECIES: endoglucanase [Raoultella] | 89% | WP_032690332.1 | GH8 |
| 77908 | Cellulase family 8 [*Enterobacter* sp. 638] | 98% | ABP62583.1 | GH8 |
| 71437 | endo-1,4-D-glucanase [*Pseudomonas fluorescens*] | 83% | AAL71844.1 | GH8 |
| 8282 | beta-1,4-xylosidase [*Raoultella ornithinolytica* B6] | 92% | WP_015585273.1 | GH43 |
| 3165 | FAD-linked oxidase [*Citrobacter freundii*] | 94% | WP_003827490.1 | AA4/AA7 |

**Table 5-6 The predicted genes targeted for PCR amplification, showing predicted function/microbe of origin**

Table 5-6 shows the details of the predicted CAZyme genes that were selected for amplification. The table also shows the putative function of the genes along with their likely taxonomic origin based on the BLAST output. The genes selected include 3 glycoside hydrolase 8 genes from *Enterobacter* sp. 638, *Pseudomonas fluorescens* and an endoglucanase gene found in multiple species from the *Raoultella* genus. A β-glucosidase gene from *Citrobacter freundii* a xylosidase gene from *Raoultella ornithinolytica* and a FAD-linked oxidase gene from *Citrobacter freundii* were also selected. The table also shows the percentage similarity (identity) of the most similar protein sequence previously observed, which ranged from 83% to 98%. The predicted gene sequences can be found in Appendix 3.

PCR amplification reactions were carried out with the primer sets seen in Table 5-1, using WGA metagenomic DNA as template. The results of these PCR reactions can be seen in Figure 5-4 and Figure 5-5.
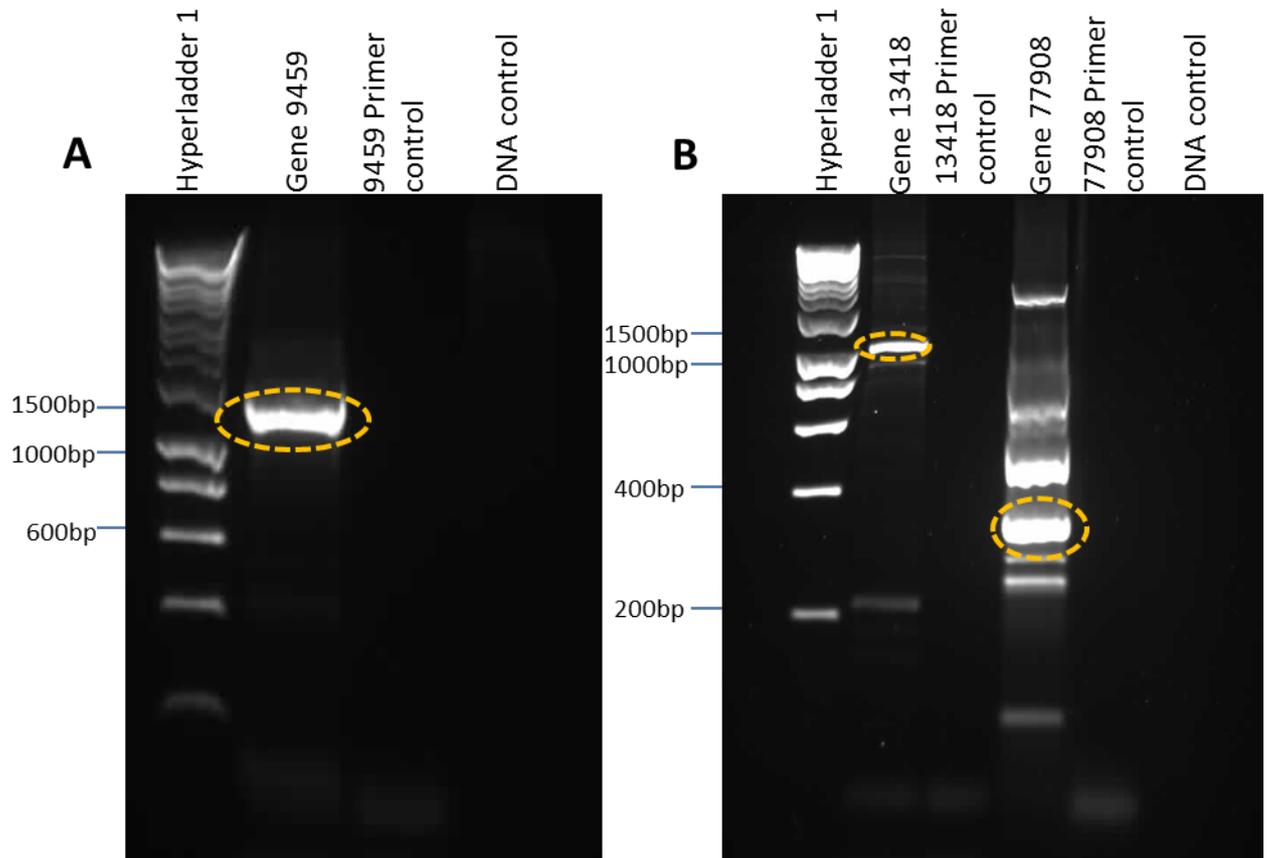


**Figure 5-4 PCR amplification of predicted genes 9459, 13418 and 77908**
*(A) A 1% agarose gel showing successful amplification of predicted gene 9459: (B) A 1% agarose gel showing successful amplification of predicted genes 13418 and 77908*

Figure 5-4 (A) shows the PCR reaction for amplification targeted, gene 9459. In total, two bands could be seen on the gel, a feint band at 400 bp and an intense band at approximately ~1400 bp, which corresponds to the predicted size of gene 9459 (1383bp). Gel (B) shows PCR reactions targeting genes 13418 and 77908. Multiple feint bands could be seen in the PCR sample for gene 13418; a prominent band was also visible at just over 1000bp corresponding to the predicted gene size of 993bp. In the PCR reaction for gene 77908 at least 6 bands

could be seen, ranging from >2500bp to <100bp. Of these, 2 bands were prominent, one at the predicted gene size of 363bp. Bands corresponding to the predicted gene sizes for each PCR reaction (indicated by yellow circles) were excised, gel extracted, cloned and sequenced as described.
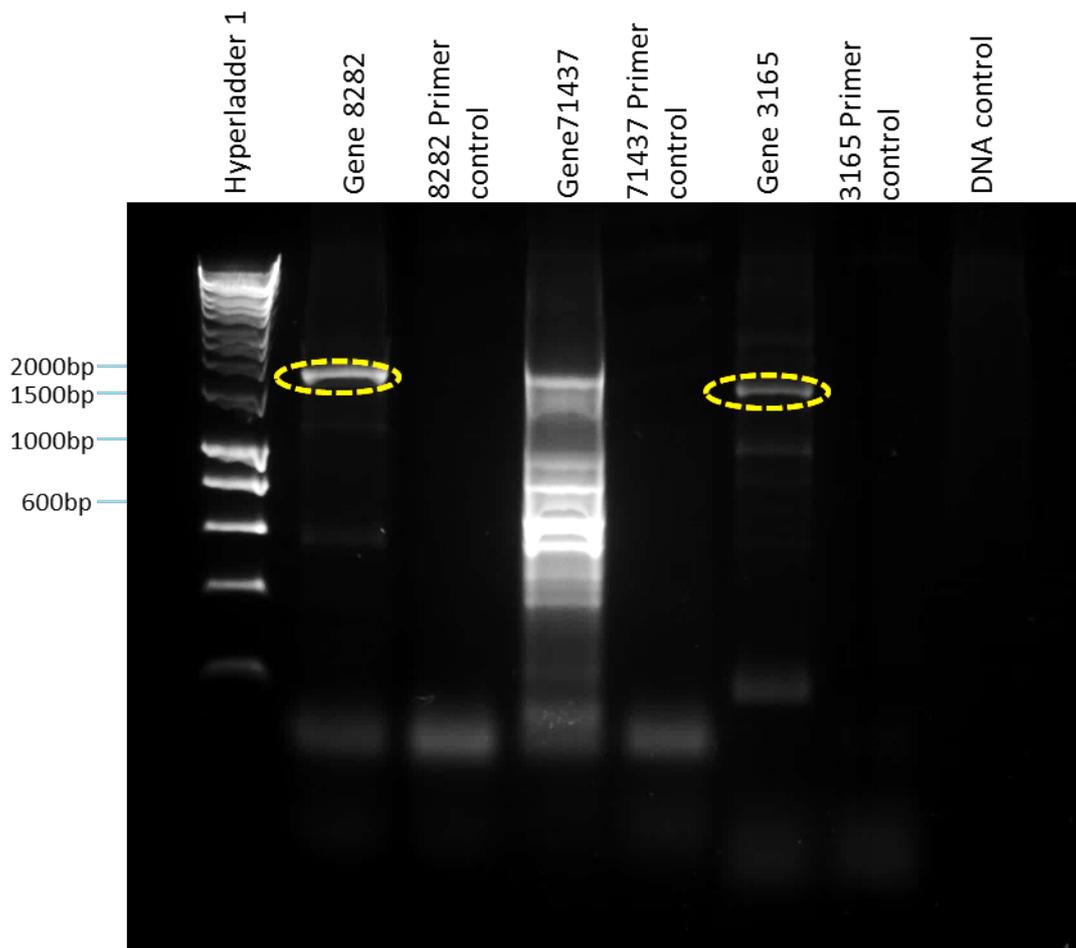


**Figure 5-5 PCR amplification of predicted genes 8282 71437 and 3165**
**A 1% agarose gel showing successful amplification of predicted genes 8282 and gene 3165**

Figure 5-5 shows the PCR reactions targeting genes 8282, 71437 and 3165. Two bands could be seen in PCR sample 8282, with the most prominent between 1500 and 2000bp (~1750bp), which corresponded to the predicted gene size of 1680bp. Multiple prominent bands could be seen in the PCR sample for gene 71437 that formed a streak between 2000bp and 400bp. No obvious band could be identified

at the predicted size of 1197bp. In the PCR reaction for gene 3165, 2 bands were present, with the most prominent being at the predicted size of 1557bp. Each band that corresponded to the predicted gene size in each case (indicated by yellow circles) was gel extracted, cloned and sequenced as described previously. No band extraction was made for sample 71437.

The Sanger sequencing results for each of the cloned amplicons showed the same putative function upon BLAST search as their predicted genes counterparts. Alignment of amplicon sequences to the original predicted genes from the metagenome revealed successful amplification of the target sequences from the WGA metagenomic samples.

In order to determine whether partial open reading frames that were predicted could be used to amplify full length genes from the metagenomic sample, the partial gene 77908 (the c terminus end of a cellulase gene) was used as query for a BLASTn search. The top hit cellulase sequence (CP000653.1|:4244370-4245452bp *Enterobacter* sp. 638, complete genome, 97% similarity) was then used to design a forward primer (sequence: ATGGTCGCGCTGGTTCT) which was used in conjunction with the "Gene_ID_77908R" reverse primer to attempt to amplify a full length gene from the WGA metagenomic sample. The results of this can be seen in Figure 5-6.
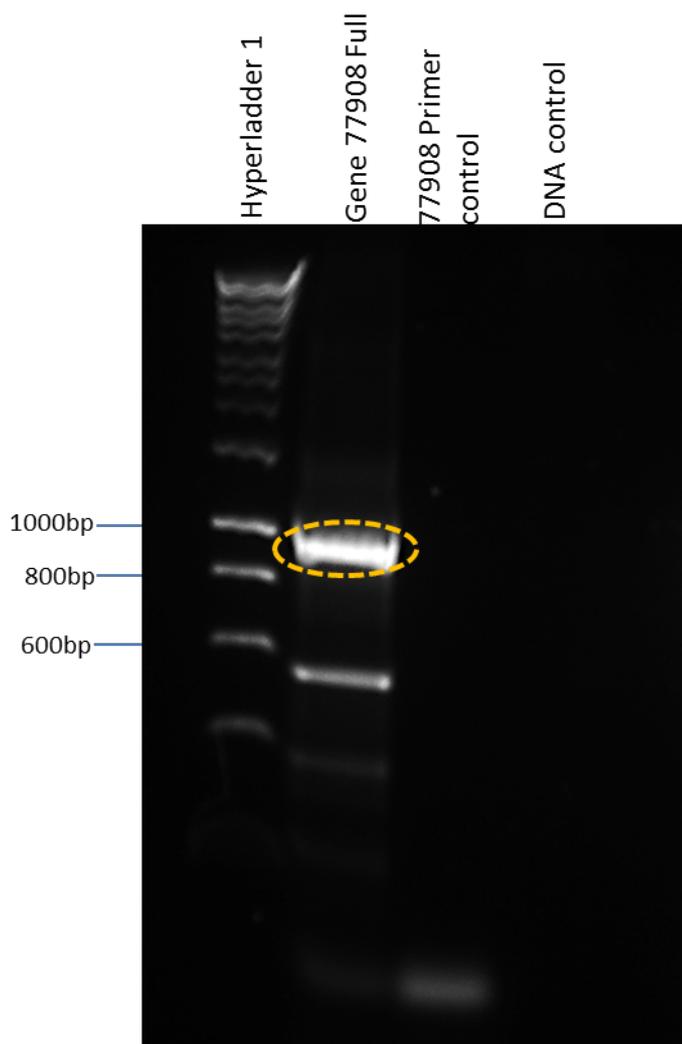
**Figure 5-6 Amplification of predicted gene "77908Full"**
*A 1% agarose gel showing successful amplification of the full version of the previously amplified partial gene 77908*

Figure 5-6 shows the results of the PCR reaction that attempted to amplify the full gene corresponding to the partial predicted gene, 77908. The predicted size of the full gene (deduced from CP000653.1|:4244370-4245452bp *Enterobacter* sp. 638, complete genome) was 1083bp. In total, 3 amplicons could be seen, one at <400bp and another at 550bp with the most prominent and broad band seen just below 1000bp, relatively close to the predicted gene size of 1083bp. The band indicated by a yellow circle in Figure 5-6 was gel extracted, cloned and sequenced in order to determine the identity of the amplicon. A BLAST search of the resultant

sequence revealed the same top hit as gene 77908 and alignment to the partial sequence revealed that the gene had been extended to full length successfully.

### 5.3.5 Expression of Gene 9459, A Putative β-glucosidase Gene

After sequencing of the gene 9459 cloning vector construct revealed successful amplification of the full length predicted gene, it was selected to be taken forward for protein expression. The results of this study can be seen in Figure 5-7.
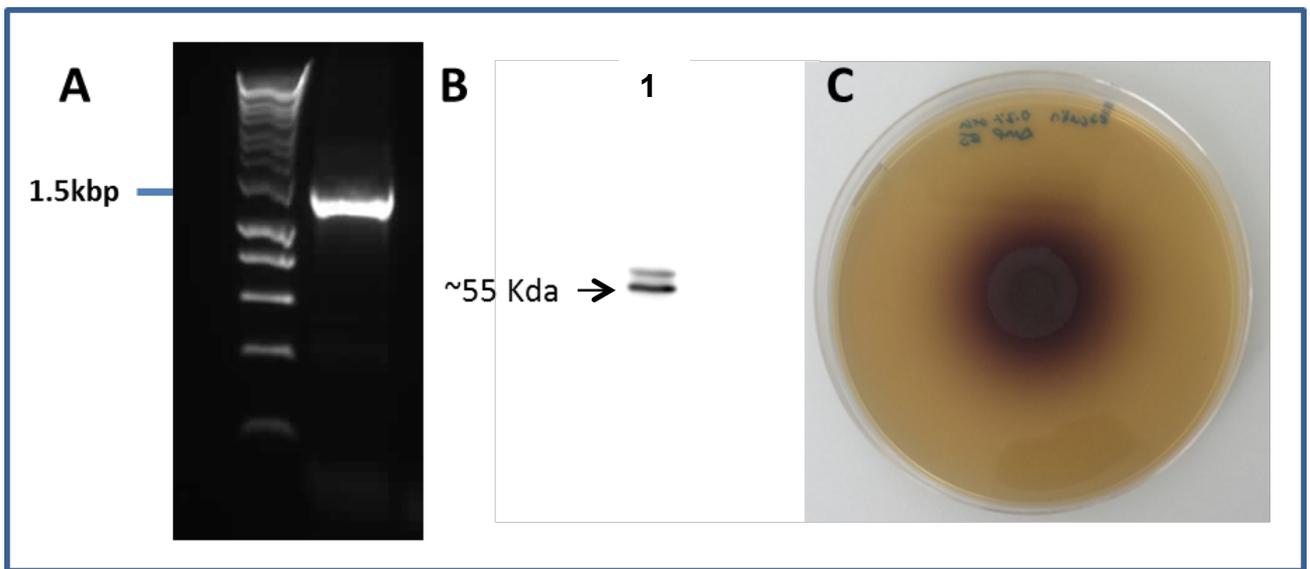


**Figure 5-7 Recombinant expression and activity testing of gene 9459**
*(A) Amplification of gene 9459 from the pCR2.1 cloning vector (B) A Western blot showing successful expression of recombinant protein (lane 1) (C) An esculin hydrate- ferric ammonium citrate activity plate showing the gene 9459 clone β-glucosidase activity*

Figure 5-7 (A) shows the successful amplification of the full 9459 gene using the pCR® 2.1 cloning vector construct as a template for PCR. The amplified gene was gel extracted and inserted into the TOPO TA expression vector. Figure 5-7 shows the western Blot results after expression of the 9459 gene was induced, two bands were identified at around 55 Kda. The predicted mass of the protein coded by gene 9459 is 57.5 Kda (including additional amino acids present at the c and n terminus as an artefact of the expression system used, including the polyhistidine tag and V5 epitope). The appearance of two bands is thought to be due to the

start codon ATG being present in both the vector sequence and the insert sequence close to the ribosome binding site (where the difference in resulting protein mass using each start codon is 1.4 Kda). Plate (C) shows the growth plate assay after spotting of 10 µL of induced culture and incubation for 24 hours. The black precipitate indicates successful cleavage of the cellobiose mimic esculin, as previously described and confirms production of a functional β-glucosidase recombinant protein.

## 5.4 Discussion

In this study the hypothesis that the slug gut microbiome could contribute to digestion and nutrient cycling was tested, specifically with respect to breakdown of complex plant cell wall superstructures that are notoriously difficult for animals to degrade without substantial assistance from microbes (Hansen and Moran, 2014)

The functional potential of the metagenome includes a wide range of gene functions involved in the degradation of plant biomass, including all of the major components of the plant cell wall superstructure, cellulose, hemicellulose and lignin. This finding supports previous work that has implicated the slug gut microbiome in the facilitation of lignocellulose degradation. Many of the CAZymes identified in this study are involved in the breakdown of oligosaccharides and their subsequent metabolism. In total, 641 predicted proteins were allocated to oligosaccharide degrading groups, including 204 β-glucosidases, 80 β-galactosidases and 279 β-xylosidases. Numbers of long chain carbohydrate degrading enzymes were generally much lower in comparison, with only 26 cellulase enzymes being identified in total. This could suggest that this environment is more suited to the breakdown of partially degraded plant material. The dominance of oligosaccharide degrading enzymes appears in all of the other comparator gut environments shown in Table 5-5, including wallabies, termites and also in the gut microbiomes of reindeer and cattle (Pope *et al.*, 2012). This trait has also been identified in other environmental microbiomes such as leaf cutter ant fungus gardens (Aylward *et al.*, 2012). This supports the hypothesis that gut microbes are predominantly involved in the breakdown of partially degraded plant material (be it partially rotten when ingested or chemically pre-processed in a stomach) across the board. However, due to the homology based

nature of metagenomics studies, there is still the possibility that there are many groups of microbial lignocellulose degrading enzymes that are unknown and therefore undetectable using similarity based methods. Enzyme groups that are involved in the degradation of hemicellulose are seen in especially high numbers in our sample when compared with other gut microbiomes, with larger numbers for every stage of degradation. This includes 321 assignments made to groups targeting long chain hemicellulose and 437 to groups that target the oligosaccharides produced during degradation. Further indications that sugars in plant cell walls are utilized by gut microbes come with the identification of numerous sugar transporter proteins. These include a large number of components of the cellobiose-specific phosphotransferase system (PTS) that facilitate the uptake of cellulose degradation products. The KEGG diagram in Figure 5-1 also shows the presence of membrane transport system components specific to Mannose and β-glucosides. Together, the identification of multiple enzymes that break down plant cell walls and the transport systems that facilitate the uptake of the resulting oligosaccharides provide a strong indication that the microbial population has an active role in the extracellular breakdown of plant cell wall components in the *A. ater* gut.

Several predicted genes from this metagenome were successfully amplified the whole genome amplified metagenomic sample. This serves to validate the assembly and the predictions made thereof, showing that it is very likely that the predicted sequences do exist in nature. A full length predicted β-glucosidase gene was successfully expressed and the enzymatic function observed using growth plate assays. To our knowledge we are the first to succeed in amplifying novel, functioning genes from a whole genome amplified metagenomic sample. The use

of whole genome amplified samples enables study of a far greater number of predicted genes by sidestepping the problem of small sample size often seen with environmental samples, which constrain genes of interest to be studied using expensive gene synthesis methods.

The use of metagenomics in the study of environmental DNA has become instrumental in developing our knowledge of microbial communities. Here we use metagenomics to gain an insight into both the phylogeny and the functional capability, of the gut microbiome of the common black slug. This study also supports previous work presented in chapters 2, 3 and 4 that implicates the slug gut microbiome in the degradation of lignocellulose. Here we identified a large repertoire of genes that give the potential for lignocellulose to be not only degraded but also for the resulting sugars up taken by members of the microbiome itself. Moreover predictions have been validated through amplification of selected glycoside hydrolase genes along with observing predicted functional activity of an amplified β-glucosidase gene. This work therefore begins to shed light on how the black slug can process the large quantities of plant biomass it consumes and provides a further example of a gut microbiome that is well equipped to breakdown plant matter.

# Chapter 6. General Discussion and Future Research

## 6.1 Main Findings

Lignocellulosic bioethanol has the potential to become a fully sustainable replacement liquid fuel source for fossil fuels on which we currently rely. However the current high costs of lignocellulosic bioethanol production are limiting any potential global transition. The major limiting factor is that the enzymes currently used are inadequate, either because of their weak activity or their vulnerability to end product inhibition, so they are currently unable to achieve both the hydrolysis of structural polysaccharides, in which simple sugars are locked, and the efficient degradation of the aromatic polymer lignin. The lack of adequate enzymes causes the production process to be long and thus expensive, with feedstock pre-treatments required that involve the use of large quantities of acids or bases to strip the biomass of the protection it gains from lignin. It is thought that discovery of new, highly active, highly stable and end-product-inhibition resistant CAZymes and lignin degrading enzymes could revolutionise the production of lignocellulosic bioethanol by reducing the cost of production to a level more comparable to that of liquid fossil fuel production. To that end, many studies have been carried out in search of novel highly active CAZymes present in nature, with a special focus on the herbivore gut (Warnecke *et al.*, 2007, Hess *et al.*, 2011, Cardoso *et al.*, 2012a, Duan *et al.*, 2009, Pope *et al.*, 2010, Pope *et al.*, 2012, Brune, 2014, Zhu *et al.*, 2011).

Thus, the aim of the research in this thesis was to identify novel carbohydrate active enzymes from nature. Initially a target environment was identified for bioprospecting that was likely to harbour a wealth of previously unobserved

CAZymes using biochemical and microbiology methodologies. Using these methods we identified the gut environment of the slug species *Arion ater* as a candidate for further study following identification of a number of gut cellulolytic bacteria. Metagenomic DNA extraction methods and bioinformatics workflows were then optimised for in depth analysis of the selected gut environment in order to characterise the microbiome and identify novel CAZymes of interest.

We identified, for the first time, individual CAZyme classes present within the *A. ater* gut juices using PAGE zymography and activity gels. This study identified members of both the enzyme classes required to breakdown long chain cellulose (1,4-β-endoglucanse) and also to break down the resultant oligosaccharides into simple sugars (β-glucosidase). This shows that the gut environment has the metabolic capability to breakdown cellulose, and infers that the environment is also likely to harbour enzymes that make cellulose available for hydrolysis through degradation of protective structural molecules such as hemicellulose, pectin and lignin. In this study we also observed no change in the enzymes detected in different portions of the gut, showing that there was little or no compartmentalization of cellulolytic activity, a trait also observed in the grasshopper *Dissosteira carolina* (Willis *et al.*, 2010a). After biochemical characterization of the cellulolytic activity (Joynson *et al.*, 2014) and identification of individual cellulolytic enzyme classes, it became clear that the gut environment harbours CAZymes but the origin of this activity was, at that point, unknown, with no previous research definitively determining the origin of *A. ater* gut cellulolytic activity.

Due to the identification of microbial cellulolytic activity in the gut environments of multiple eukaryotes, including mammals (Duan *et al.*, 2009, Zhu *et al.*, 2011),

insects (Anand *et al.*, 2010, Cho *et al.*, 2010) and other gastropods (Gupta *et al.*, 2012, Cardoso *et al.*, 2012a), we isolated gut microbes to assess whether members of the *A. ater* gut microbiome could contribute to the observed cellulolytic activity. To do this gut microbes were isolated and tested for both β-1,4-endoglucanase and β-glucosidase activity using growth plate assays and were subsequently identified by sequencing of their 16s rRNA genes. This study yielded identification of 12 cellulolytic gut microbes, proving for the first time that the *A. ater* gut microbiome does indeed possess the metabolic capability to contribute to the symbiotic degradation of lignocellulose in the gut. Another important implication of this study was that many of the microbes isolated could not be identified to within the accepted species level identification threshold of >97% identity with any sequence in the NCBI databases. This indicates that the slug gut harbours not only cellulolytic microbes, but cellulolytic microbes that have not previously been observed. Given these findings, a study was carried out in order to identify other members of the gut consortium that may not have been culturable using techniques used previously. In total the DGGE study allowed a further 9 identifications to be made, 7 of which were members of the *Gammaproteobacteria* class. This began to suggest a microbiome dominated by that class, which was eventually corroborated in later studies using metagenomics.

The results presented in chapters 2 and 3, along with the biochemical analysis of slug gut cellulolytic activity (Joynson *et al.*, 2014) showed that high levels of cellulolytic activity were present in the gut, that the enzymes required to bring about the total degradation of cellulose into simple sugars were present and that at least a portion of this cellulolytic activity was of microbial origin. These studies were therefore successful in identifying the *A. ater* gut as a target environment for

in-depth study using high resolution molecular methods, showing that it harbours unstudied microbial cellulolytic systems that could potentially be exploited by the biofuels industry. Similar methodological approaches have been used previously to advance understanding of cellulolytic activity in cow rumens and in termites, where cellulolytic activity was identified as of microbial origin (Teather and Wood, 1982, Wenzel *et al.*, 2002) using culture activity plates followed by highly successful bioprospecting studies carried out using NGS based metagenomics (Hess *et al.*, 2011, Warnecke *et al.*, 2007). The success seen in previous studies following identification of cellulolytic gut microbes gave a good indication that the *A. ater* gut environment was an ideal target for bioprospecting for CAZymes.

A metagenomic study was then undertaken to characterize the gut microbiome as a whole, both in ecological terms (chapter 4) and metabolically (chapter 5). This is the first time the *A. ater* gut microbiome has been subject to such high-resolution analysis, both to provide information on the composition of the gut microbial consortium and their respective metabolic capabilities.

The phylogenetic analyses described in chapter 4 indicated that the gut microbiome was dominated by members of the Gammaproteobacteria class, as had been predicted in chapter 3. We showed that the vast majority, almost three quarters of allocations, were made for only six genera, Enterobacter, Citrobacter, Pseudomonas, Eschericia, Acinetobacter and an unclassified genus belonging to the Sphingobacteriaceae, with all but one being from the Gammaproteobacteria class (Sphingobacteriaceae, that are members of the Sphingobacteria class). Although the microbiome was dominated by a small number of genera (Table 4-2) over 253 genera (Figure 4-7) were represented in the sample and a total of 375 species level identifications made which indicates that the gut microbiome is in

fact ecologically diverse. The level of dominance observed by this small number of genera suggests that the gut microbiome may have a core community of microbes that are permanent members of the gut microbial consortium along with a more transient element, where some of the members of the remaining 247 genera that make up only 27% may be simply ingested during feeding and excreted. However it is also possible that these other non-dominant groups were suppressed at the time of sampling due to nutritional cycling not being in their favour, a trait which has been observed in cockroaches where microbes more able to digest the food type ingested by the host appear in greater numbers (Bertino-Grimaldi et al., 2013). The dominance of the Proteobacteria phylum in the plant eating gastropod gut is becoming ever more evident with studies into freshwater planorbid snails (Biomphalaria pfeifferi) and terrestrial snails such as the giant African land snail (Achatina fulica) (Cardoso et al., 2012b) and now the common black slug Arion ater all showing high abundances of Proteobacteria. This also extends to herbivorous insects (Dillon and Dillon, 2004, Russell et al., 2009a), and to leaf cutter and fungus gardens (Suen et al., 2010, Aylward et al., 2012) suggesting a more general association of this phylum with facilitation of lignocellulose degradation not only with herbivorous insects but also with plant-eating gastropods and other lignocellulose degrading environments. In this study we also re-confirm the presence of microbes identified in chapter 3 (Table 4-4),

giving confidence to the identifications made and suggesting that some of these cellulolytic microorganisms present in the gut may be permanent members of the gut microbiome.

The phylogenetic studies also revealed the presence of multiple plant pathogen species in the gut environment. These include *Agrobacterium tumefaciens,*

*Pseudomonas syringae, Dickeya dadantii, Pectobacterium carotovorum, Xanthomonas campestris* and *Erwinia amylovora* all of which are rated as among the top 10 most important plant pathogens worldwide *(Toth et al., 2011).* Further analysis was also carried out to determine if enough sequencing data was obtained to allow species specific identification of members of the pathogenic *Dickeya* genus, where between 454 and 2854 species specific reads were found for the species *D. solani, D. dadantii, D. chrysanthemi* and *D. zeae* which demonstrates the presence of multiple species that could be pathogenic to plants. Further to this, in chapter 3 it was elucidated that *Pectobacterium carotovorum* was present in a sample taken in 2012 using DGGE and also identified the same species in samples taken from the same area a year later in 2013. Not only is this the first time that plant pathogen species have been identified in the gut of *A. ater*, it is also the first example of replication of such findings over time, that suggest that the slug gut could have a role in overwintering of a pathogen species. This also implies a role in the transmission of the pathogen, a hypothesis that would, however, require more work over a broad sampling range and over multiple seasons to be tested.

Using the metagenomic sequencing data, in depth analysis of the functional capabilities of the bacteria present in the gut was carried out. This study identified 5,635 genes with Pfam domains matching those present in the CAZy database including 2510 glycoside hydrolase genes, 561 carbohydrate binding modules and 312 auxiliary activity linked genes thought to be involved in the degradation of lignin. These identifications included enzymes that degrade cellulose and hemicellulose along with the resultant oligosaccharides (Table 5-5) and redox enzymes involved in the breakdown of aromatic monolignols present in lignin. This

indicates that our hypothesis made after identification of cellulolytic bacteria was correct, that the *A. ater* gut environment was harbouring microbes that have the ability to fully degrade lignocellulose. This is the first time that these enzyme classes have been observed in the gut of the slug and these findings provide a strong indication that the microbes present are contributing to the breakdown of the lignocellulose portion of the *A. ater* diet. The annotated predicted gene sequences were also validated through amplification of multiple predicted CAZyme genes from whole genome amplified metagenomic DNA. Successful amplification of 5 glycoside hydrolase genes and observation that the predicted sequences truly exist in nature gives strength to the findings from the bioinformatics analyses, while the expression of the β-glucosidase gene, 9459, shows that these homology-based annotation methods are capable of identifying novel functioning gene sequences.  To our knowledge, and as of the time of submission of this thesis, this is only the second study using shotgun metagenomic study on the gut microbiome of a gastropod species, the first being that of the snail *Achatina fulica* by Cardoso *et al.* (2012a). Making this the first study to carry out metagenomic analysis on the slug gut microbiome, with all observations made here being novel discoveries. To our knowledge this is also the first time a functional CAZYme has been amplified and expressed that was identified in a gastropod gut environment. Further to this, again, to our knowledge, this is the first time a CAZyme gene has been amplified and expressed producing a functional protein, from a metagenomic sample that has been subjected to whole genome amplification to increase sample size and therefore allow a great number of the predicted proteins to be studied in-depth.

Clarification that whole genome amplification is effective on samples containing multiple genomes in varying abundances could remove one of the major limitations of metagenomic sampling, which is the small sample size often produced, most of which is often destroyed during shotgun library preparation. The fact that we have shown that ~4.5 µg WGA metagenomic sample, of good enough quality for PCR amplification of predicted genes, can be produced from only 10 ng of sample will allow the important predictions from metagenomic studies to be examined to a much greater extent. Potentially thousands of predicted protein sequences could be studied and both samples and predictions shared between groups looking at different protein functionalities.

In summary, this thesis has made multiple significant contributions to knowledge: firstly the identification of novel cellulolytic microbes in the gut of the slug *A. ater;* secondly that the gut microbiome is dominated by *Proteobacteria,* more specifically, *Gammaproteobacteria* and the members of the genera *Enterobacter;* and thirdly, that the slug gut environment harbours multiple plant pathogens including *Dickeya dadantii*, D*ickeya Solani, Erwinia amylovora and Pectobectrium carotovorum.* Here we have also identified thousands of novel carbohydrate active proteins that target every main structural component of the plant cell wall, including over 300 associated with lignin degradation. Finally, we have also demonstrated that we can overcome the sample size limitation associated with metagenomics, producing a method that can be used to help to exploit further, metagenomics studies in the future where a far greater number of predicted genes can be studied.

## 6.2 Future Work

Three main areas of future work can be identified from the findings in this thesis.

### 6.2.1 Exploitation of Metagenome Predicted CAZYmes

The first is the continuation of amplification and expression of the thousands of carbohydrate active enzymes identified during this research project. With over 2,500 glycoside hydrolases predicted from the metagenome, there is still much that can be achieved from the metagenomic study.

The next step will be to express more of these novel proteins and profile their levels of activity and their stability. This offers the potential to discover additional highly active enzymes that could be utilized in the biofuel industry. These new protein sequences could also be used in gene shuffling experiments that are often used in attempts to increase stability or activity of an enzyme (Cherry and Fidantsef, 2003, Shibuya *et al.*, 2000) by shuffling the coding sequences of multiple orthologues of enzymes. Enzymes of particular interest are those assigned to the lignin linked auxiliary activity Pfam groups. The use of enzymes has removed the requirement of high temperatures and use of extreme pH in many industrial processes (Cherry and Fidantsef, 2003) but the lack of efficient enzymes targeting lignin means that expensive acid/alkaline based pre-treatment of lignocellulosic feedstocks are still required. Further study of the 312 AA enzymes identified in the gut metagenome could lead to discovery of highly active enzymes that can efficiently degrade lignin at moderate temperatures and pH levels. Further investigation should also be carried out into the high numbers of hemicellulose degrading enzymes identified in this study, where the slug gut microbiome was shown to harbour many more than the other comparator

193

organisms (Table 5-5). The WGA metagenomic DNA created in this project could facilitate the study of a great number of proteins predicted in this metagenome. With greater than 40 µg of WGA metagenomic DNA produced, at 100 ng of DNA per amplification reaction, in theory as many as 400 identified genes could be studied using the methods developed in this thesis. Furthermore the entire process of bioprospecting, target identification, metagenomic DNA extraction, bioinformatics analysis and eventual expression of enzymes of interest could be replicated for other eukaryote guts or other environments.

## 6.2.2 Investigation Into The Contribution of *A.ater* In Lignocellulose Degradation

Further work could also be carried out to investigate if the slug itself is contributing to the enzymatic breakdown of lignocellulose. Although some activities require for complete breakdown have not yet been observed in animals, some such as cellulases have been identified in many different arthropods and nematodes (Watanabe and Tokuda, 2001) and a cellobiohydrolase has recently been identified in one marine arthropod, Limnoria quadripunctata (the gribble). Two methods could be used to determine whether lignocellulose degrading enzymes of animal origin are present. The first would be to carry out FPLC to separate the proteins present in crude gut samples and identify which of the resulting aliquots of separated protein sample exhibit cellulolytic activity, using agar or agarose plates containing CMC and the staining and destain methods used in this study. The aliquots that exhibit activity could then be separated using SDS PAGE and in gel zymography carried out as described previously. The resultant protein bands from locations that exhibit cellulolytic activity could then be assessed using LC-MS and the peptide sequences determined then submitted for homology searches

against peptide databases to identify whether they are of animal or bacterial origin. A next generation sequencing-based study could also be carried out, such as a genomic study to create a draft assembly of the *A. ater* genome, from which genes could be predicted and identified through homology searches. Transcriptomics could also be used to study gene expression in the gut or digestive gland of the slug, elucidating whether the slug itself also plays a role in lignocellulose degradation.

## 6.2.3 Investigations Into The Role of *A. ater* In Plant Pathogen Survival and Transmission and Into The Ecological Stability of The Gut Microbiome

In this study we also identified a number of bacteria known to be responsible for some of the most financially detrimental plant diseases in Europe. These phylogenetic analyses showed hits for 6 of the top 10 bacterial plant pathogen species in the world. These included *Pectobacterium carotovorum* and *Erwinia amylovora,* that were also identified using DGGE from samples taken a year prior to the metagenomic study, suggesting that the slug gut could be implicated in the over wintering of plant pathogens.   Although the analyses carried out in this study began to clarify whether the phylogenetic results in this project can be used to accurately assign taxonomy of these pathogens down to species level using the sequences associated with members of the *Dickeya* genus, further in depth study is required to confirm the presence of specific pathogenic species in the gut in large numbers. Further study would also be required to determine if the *A. ater* gut microbiome has core set of persistent microbiota or is highly transient. Both of these questions could be answered by a study using 16s population genetics. Metagenomic DNA samples could be taken from *A. ater* samples, obtained over a number of years from different areas of interest such as crop fields suffering from

the pathogens previously identified, garden centres from where many pathogens are thought to be spread, and also from the same suburban area used here for sampling. These samples could then be subjected to PCR, targeting a 500-700 bp region of the 16s rRNA gene, followed by sequencing using a 2x250 or 2x300 bp MiSeq sequencer. The resultant sequences could then quickly lead to microbial identifications using phylogenetic analysis software such as Qiime (Caporaso *et al.*, 2010). A study of this nature would also elucidate whether there is a core set of gut microbes present within the *A. ater* gut or if the microbiome is more transient. Population genetics studies would also show if there was intraspecies variation in gut microbiome composition between slugs living in different environments and even within the same environments.

Using the genome assemblies for pathogenic *Dickeya* species provided by Dr Leighton Pritchard (James Hutton, Dundee, UK), we could also seek to identify species specific marker genes that could allow PCR reactions to be carried out for use in diagnostic tests that could be applied to a very large number of samples to quickly identify the presence of certain pathogen species of interest. The two phylogenetic studies described above would determine whether these pathogen microbes are endemic to the *A. ater* gut and elucidate the potential role of slugs in the transmission and overwintering of plant pathogens, which in turn could lead to development of improved epidemiological control measures. Clarification of the role, if any, that slugs play in the transmission on and survival of pathogens is also very important now because of the European Union ban on traditional agricultural slug pellets that comes into effect in September 2015. These pellets are currently used to limit the impact of slug feeding on food crops by means of population control in fields. However if slug populations increase due to lack of pest control

measures, there may also be an increase in the proportion of crops lost to disease.

# Biblography

Adams, L. & Boopathy, R. 2005. Isolation and characterization of enteric bacteria from the hindgut of Formosan termite. *Bioresource technology,* 96, 1592-8.

Agular, R. & Wink, M. 2005. How do slugs cope with toxic alkaloids? *Chemoecology,* 15, 167-177.

Ahmad, M., Taylor, C. R., Pink, D., Burton, K., Eastwood, D., Bending, G. D. & Bugg, T. D. 2010. Development of novel assays for lignin degradation: comparative analysis of bacterial and fungal lignin degraders. *Mol Biosyst,* 6, 815-21.

Airey, W. J. 1987. Laborartory Studies on Damage to Potato Tubers by Slugs. *Journal of Molluscan Studies,* 53, 97-104.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215, 403-10.

An, C. L., Lim, W. J., Hong, S. Y., Shin, E. C., Kim, M. K., Lee, J. R., Park, S. R., Woo, J. G., *et al* 2005. Structural and biochemical analysis of the asc operon encoding 6-phospho-beta-glucosidase in *Pectobacterium carotovorum subsp. carotovorum* LY34. *Research in Microbiology,* 156, 145-153.

Anand, A. a. P., Vennison, S. J., Sankar, S. G., Prabhu, D. I. G., Vasan, P. T., Raghuraman, T., Geoffrey, C. J. & Vendan, S. E. 2010. Isolation and characterization of bacteria from the gut of *Bombyx mori* that degrade cellulose, xylan, pectin and starch and their impact on digestion. *Journal of Insect Science,* 10.

Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antonio, E. S., Kasai, A., Ueno, M., Kurikawa, Y., Tsuchiya, K., Toyohara, H., Ishihi, Y., Yokoyama, H. & Yamashita, Y. 2010. Consumption of terrestrial organic matter by estuarine molluscs determined by analysis of their stable isotopes and cellulase activity. *Est Coast Shelf Sci,* 86, 401-407.

Aronesty, E. 2011. ea-utils : Command-line tools for processing biological sequencing data

Aylward, F. O., Burnum, K. E., Scott, J. J., Suen, G., Tringe, S. G., Adams, S. M., Barry, K. W., Nicora, C. D., Piehowski, P. D., *et al* 2012. Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens. *The ISME Journal,* 6**,** 1688-1701.

Azambuja, P., Feder, D. & Garcia, E. S. 2004. Isolation of *Serratia marcescens* in the midgut of *Rhodnius prolixus*: impact on the establishment of the parasite Trypanosoma cruzi in the vector. *Experimental Parasitology,* 107**,** 89-96.

Baker, A. A., Helbert, W., Sugiyama, J. & Miles, M. J. 2000. New insight into cellulose structure by atomic force microscopy shows the i(alpha) crystal phase at near-atomic resolution. *Biophysical Journal,* 79**,** 1139-1145.

Ballesteros, M., Oliva, J. M., Negro, M. J., Manzanares, P. & Ballesteros, I. 2004. Ethanol from lignocellulosic materials by a simultaneous saccharification and fermentation process (SFS) with Kluyveromyces marxianus CECT 10875. *Process Biochemistry,* 39**,** 1843-1848.

Banerjee, M. 1989. Kinetics of ethanolic fermentation of D-xylose by *Klebsiella pneumoniae* and its mutants. *Applied and Environmental Microbiology,* 55**,** 1169-1177.

Basso, L. C., De Amorim, H. V., De Oliveira, A. J. & Lopes, M. L. 2008. Yeast selection for fuel ethanol production in Brazil. *FEMS Yeast Research,* 8**,** 1155-1163.

Bauer, W. D., Talmadge, K. W., Keegstra, K. & Albersheim, P. 1973. The Structure of Plant Cell Walls: II. The Hemicellulose of the Walls of Suspension-cultured Sycamore Cells. *Plant Physiology,* 51**,** 174-187.

Bayer, E. A., Belaich, J.-P., Shoham, Y. & Lamed, R. 2004. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annual review of microbiology,* 58**,** 521-54.

Bayer, E. A., Morag, E. & Lamed, R. 1994. The cellulosome — A treasure-trove for biotechnology. *Trends in biotechnology,* 12**,** 379-386.

Berlanga, M., Paster, B. J., Grandcolas, P. & Guerrero, R. 2011. Comparison of the gut microbiota from soldier and worker castes of the termite *Reticulitermes grassei. Group,* 14**,** 83-93.

Bertino-Grimaldi, D., Medeiros, M. N., Vieira, R. P., Cardoso, A. M., Turque, A. S., Silveira, C. B., Albano, R. M., Bressan-Nascimento, S., *et al.* 2013. Bacterial community composition shifts in the gut of *Periplaneta americana* fed on different lignocellulosic materials. *Springerplus, 2*, 609.

Blotkamp, P. J., Takagi, M., Pemberton, M. S. & Emert, G. H. 1978. *Enzymatic hydrolysis of cellulose and simultaneous fermentation to alcohol.*

Boerjan, W., Ralph, J. & Baucher, M. 2003. Lignin biosynthesis. *Annual review of plant biology, 54*, 519-46.

Bradford, M. M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem, 72*, 248-54.

Brown, R. M. J. & Saxena, I. M. 2007. Cellulose: Molecular and Structural Biology. *Library.*

Brune, A. 2014. Symbiotic digestion of lignocellulose in termite guts. *Nature Reviews Microbiology, 12*, 168-180.

Butera, G., Ferraro, C., Colazza, S., Alonzo, G. & Quatrini, P. 2012. The culturable bacterial community of frass produced by larvae of *Rhynchophorus ferrugineus Olivier* (Coleoptera: Curculionidae) in the Canary island date palm. *Letters in Applied Microbiology, 54*, 530-536.

Cadoche, L. & Lopez, G. D. 1989. Assessment of Size-Reduction as a Preliminary Step in the Production of Ethanol from Lignocellulosic Wastes. *Biological Wastes, 30*, 153-157.

Cain, A. J. & Williamson, M. H. 1958. Variation and specific limits in the *Arion ater* aggregate. *Journal of Molluscan Studies, 33*, 72-86.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature methods, 7*, 335-336.

Cardoso, A. M., Cavalcante, J. J., Cantao, M. E., Thompson, C. E., Flatschart, R. B., Glogauer, A., Scapin, S. M., Sade, Y. B., *et al.* 2012a. Metagenomic analysis of the microbiota from the crop of an invasive snail reveals a rich reservoir of novel genes. *PLoS One, 7*, e48505.

Cardoso, A. M., Cavalcante, J. J. V., Vieira, R. P., Lima, J. L., Grieco, M. a. B., Clementino, M. M., Vasconcelos, A. T. R., Garcia, E. S., *et al.* 2012b. Gut Bacterial Communities in the Giant Land Snail *Achatina fulica* and Their Modification by Sugarcane-Based Diet. *Plos One,* 7.

Cazemier, A. E., Den Camp, H. J. M. O., Hackstein, J. H. P. & Vogels, G. D. 1997. Fibre Digestion in Arthropods. *Comparative Biochemistry and Physiology Part A: Physiology,* 118, 101-109.

Cherry, J. R. & Fidantsef, A. L. 2003. Directed evolution of industrial enzymes: an update. *Current Opinion in Biotechnology,* 14, 438-443.

Cho, M.-J., Kim, Y.-H., Shin, K., Kim, Y.-K., Kim, Y.-S. & Kim, T.-J. 2010. Symbiotic adaptation of bacteria in the gut of *Reticulitermes speratus*: low endo-beta-1,4-glucanase activity. *Biochemical and biophysical research communications,* 395, 432-5.

Claassen, P. a. M., Van Lier, J. B., Contreras, A. M. L., Van Niel, E. W. J., Sijtsma, L., Stams, A. J. M., De Vries, S. S. & Weusthuis, R. A. 1999. Utilisation of biomass for the supply of energy carriers. *Applied Microbiology and Biotechnology,* 52, 741-755.

Clarke, P. H. & Tracey, M. V. 1956. The Occurrence of Chitinase in some Bacteria. *Journal of General Microbiology,* 14, 188-196.

Cruden, D. L. & Markovetz, A. J. 1979. Carboxymethyl cellulose decomposition by intestinal bacteria of cockroaches. *Applied and environmental microbiology,* 38, 369-72.

Czajkowski, R., Grabe, G. & Van Der Wolf, J. 2009. Distribution of Dickeya spp. and *Pectobacterium carotovorum* subsp. *carotovorum* in naturally infected seed potatoes. *European Journal of Plant Pathology,* 125, 263-275.

Dahlman, O., Jacobs, A. & Sjoberg, J. 2003. Molecular properties of hemicelluloses located in the surface and inner layers of hardwood and softwood pulps. *Cellulose,* 10, 325-334.

Dillon, R. J. & Dillon, V. M. 2004. The gut bacteria of insects: Nonpathogenic interactions. *Annu Rev Entomol,* 49, 71-92.

Divne, C., Ståhlberg, J., Teeri, T. T. & Jones, T. A. 1998. High-resolution crystal structures reveal how a cellulose chain is bound in the 50 A long tunnel of

cellobiohydrolase I from *Trichoderma reesei. Journal of Molecular Biology,* 275, 309-325.

Doi, R. H. & Kosugi, A. 2004. Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nature reviews. Microbiology,* 2, 541-51.

Dojnov, B., Pavlovic, R., Bozic, N., Margetic, A., Nenadovic, V., Ivanovic, J. & Vujcic, Z. 2013. Expression and distribution of cellulase, amylase and peptidase isoforms along the midgut of *Morimus funereus* L. (*Coleoptera: Cerambycidae*) larvae is dependent on nutrient substrate composition. *Comp Biochem Physiol B Biochem Mol Biol,* 164, 259-267.

Doolittle, M., Raina, A., Lax, A. & Boopathy, R. 2008. Presence of nitrogen fixing *Klebsiella pneumoniae* in the gut of the Formosan subterranean termite (*Coptotermes formosanus*). *Bioresource technology,* 99, 3297-300.

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, 245-256.

Edgerton, M. D. 2009. Increasing crop productivity to meet global needs for feed, food, and fuel. *Plant Physiology,* 149, 7-13.

Edwards, M. C., Henriksen, E. D., Yomano, L. P., Gardner, B. C., Sharma, L. N., Ingram, L. O. & Peterson, J. D. 2011. Addition of Genes for Cellobiase and Pectinolytic Activity in *Escherichia coli* for Fuel Ethanol Production from Pectin-Rich Lignocellulosic Biomass. *Appl Environ Microbiol,* 77, 5184-5191.

Erdei, B., Franko, B., Galbe, M. & Zacchi, G. 2012. Separate hydrolysis and co-fermentation for improved xylose utilization in integrated ethanol production from wheat meal and wheat straw. *Biotechnology for Biofuels,* 5.

Evans, N. J. 1986. An investigation of the status of the terrestrial slugs *Arion ater ater* (L.) *and Arion ater rufus* (L.) (*Mollusca, Gastropoda, Pulmonata*) in Britain. *Zoologica Scripta,* 15, 313-322.

Evans, W. a. L. & Jones, E. G. 1962a. Carbohydrates in the alimentary tract of the slug *Arion ater* L. *Comparative Biochemistry and Physiology,* 5, 149-160.

Evans, W. a. L. & Jones, E. G. 1962b. A note on the proteinase activity in the alimentary tract of the slug *Arion ater* L. *Comparative Biochemistry and Physiology,* 5, 223-225.

Festucci-Buselli, R. A., Otoni, W. C. & Joshi, C. P. 2007. Structure, organization, and functions of cellulose synthase complexes in higher plants. *Brazilian Journal Of Plant Physiology,* 19, 1-13.

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. *Nat Rev Micro,* 6, 121-131.

Fontes, C. M. G. A. & Gilbert, H. J. 2010. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annual review of biochemistry,* 79, 655-81.

Galbe, M. & Zacchi, G. 2002. A review of the production of ethanol from softwood. *Applied Microbiology and Biotechnology,* 59, 618-628.

Ghosh, P. & Ghose, T. K. 2003. Bioethanol in India: recent past and emerging future. *Advances in biochemical engineeringbiotechnology,* 85, 1-27.

Gijzen, H. J., Van Der Drift, C., Barugahare, M. & Op Den Camp, H. J. M. 1994. Effect of Host Diet and Hindgut Microbial Composition on Cellulolytic Activity in the Hindgut of the American Cockroach, *Periplaneta americana*. *Applied and Environmental Microbiology,* 60, 1822-1826.

Gilbert, J. A. & Dupont, C. L. 2011. Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science,* 3, 347-371.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C. & Warren, R. A. 1991. Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families. *Microbiological reviews,* 55, 303-315.

Gismervik, K., Bruheim, T., Rorvik, L. M., Haukeland, S. & Skaar, I. 2014. Invasive slug populations ( Arion vulgaris ) as potential vectors for *Clostridium botulinum. Acta Vet Scand,* 56, 65.

Gossett, J. M., Stuckey, D. C., Owen, W. F. & Mccarty, P. L. 1982. Heat treatment and anaerobic digestion of refuse. *Journal of the Environmental Engineering Division,* 108, 437-454.

Grabber, J. H. 2005. How Do Lignin Composition, Structure, and Cross-Linking Affect Degradability? A Review of Cell Wall Model Studies. *Crop Science,* 45, 820-831.

Graham-Rowe, D. 2011. Agriculture: Beyond food versus fuel. *Nature,* 474, S6-S8.

Gugliandolo, C., Lentini, V. & Maugeri, T. L. 2011. Distribution and diversity of bacteria in a saline meromictic lake as determined by PCR-DGGE of 16S rRNA gene fragments. *Current microbiology,* 62, 159-66.

Gupta, A. & Verma, J. P. 2015. Sustainable bio-ethanol production from agro-residues: A review. *Renewable and Sustainable Energy Reviews,* 41, 550-567.

Gupta, P., Samant, K. & Sahu, A. 2012. Isolation of cellulose-degrading bacteria and determination of their cellulolytic potential. *International journal of microbiology,* 2012.

Gupta, R. & Lee, Y. Y. 2010. Investigation of biomass degradation mechanism in pretreatment of switchgrass by aqueous ammonia and sodium hydroxide. *Bioresource Technology,* 101, 8185-8191.

Hahn-Hagerdal, B., Karhumaa, K., Fonseca, C., Spencer-Martins, I. & Gorwa-Grauslund, M. F. 2007. Towards industrial pentose-fermenting yeast strains. *Applied Microbiology and Biotechnology,* 74, 937-953.

Hansen, A. K. & Moran, N. A. 2014. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology,* 23, 1473-1496.

Harholt, J., Suttangkakul, A. & Vibe Scheller, H. 2010. Biosynthesis of pectin. *Plant physiology,* 153, 384-95.

Hartmann, H., Angelidaki, I. & Ahring, B. K. 2000. Increase of anaerobic degradation of particulate organic matter in full-scale biogas plants by mechanical maceration. *Water Science and Technology,* 41, 145-153.

Havlík, P., Schneider, U. A., Schmid, E., Böttcher, H., Fritz, S., Skalský, R., Aoki, K., Cara, S. D., Kindermann, G. & Kraxner, F. 2010. Global land-use implications of first and second generation biofuel targets. *Energy Policy,* 39, 5690-5702.

Hendriks, A. T. W. M. & Zeeman, G. 2009. Pretreatments to enhance the digestibility of lignocellulosic biomass. *Bioresource Technology,* 100, 10-18.

Henrissat, B. 2000. Glycoside Hydrolases and Glycosyltransferases. Families, Modules, and Implications for Genomics. *Plant Physiology,* 124, 1515-1519.

Henrissat, B. & Davies, G. 1997. Structural and sequence-based classification of glycoside hydrolases. *Current Opinion in Structural Biology,* 7, 637-644.

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., *et al.* 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science,* 331, 463-7.

Hongoh, Y., Ohkuma, M. & Kudo, T. 2003. Molecular analysis of bacterial microbiota in the gut of the termite *Reticulitermes speratus* (Isoptera; Rhinotermitidae). *Fems Micriobiology Ecology,* 44, 231-242.

Horii, F., Yamamoto, H. & Hirai, A. 1997. Microstructural Analysis of Microfibrils of Bacterial Cellulose. *Macromoleular symposium,* 205, 197-205.

Huang, S. W., Sheng, P. & Zhang, H. Y. 2012. Isolation and Identification of Cellulolytic Bacteria from the Gut of *Holotrichia parallela* Larvae (Coleoptera: Scarabaeidae). *Int J Mol Sci,* 13, 2563-2577.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research,* 21, 1552-1560.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature,* 486, 207-214.

Iiyama, K., Lam, T. & Stone, B. A. 1994. Covalent Cross-Links in the Cell Wall. *Plant physiology,* 104, 315-320.

Ingram, L. O., Gomez, P. F., Lai, X., Moniruzzaman, M., Wood, B. E., Yomano, L. P. & York, S. W. 1998. Metabolic engineering of bacteria for ethanol production. *Biotechnology and Bioengineering,* 58, 204-214.

Ireland, M. P. 1979. Distribution of Essential and Toxic Metals in the Terrestrial Gastropod *Arion-Ater*. *Environmental Pollution,* 20, 271-278.

205

James, R., Nguyen, T., Arthur, W., Levine, K. & Williams, D. C. 1997. Hydrolase (beta-glucanase, alpha-glucanase, and protease) activity in *Ariolimax columbianus* (banana slug) and *Arion ater* (garden slug). *Comp Biochem Physiol B Biochem Mol Biol,* 118, 275-283.

Jiménez, D. J., Andreote, F. D., Chaves, D., Montaña, J. S., Osorio-Forero, C., Junca, H., Zambrano, M. M. & Baena, S. 2012. Structural and Functional Insights from the Metagenome of an Acidic Hot Spring Microbial Planktonic Community in the Colombian Andes. *PLoS ONE,* 7, e52069.

Jin, H., Chen, L., Wang, J. & Zhang, W. 2014. Engineering biofuel tolerance in non-native producing microorganisms. *Biotechnology Advances,* 32, 541-548.

Joshi, N. & Fass, J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

Joynson, R., Swamy, A., Bou, P. A., Chapuis, A. & Ferry, N. 2014. Characterization of cellulolytic activity in the gut of the terrestrial land slug *Arion ater*: Biochemical identification of targets for intensive study. *Comp Biochem Physiol B Biochem Mol Biol,* 177-178, 29-35.

Kay, P. & Grayson, R. 2013. Using water industry data to assess the metaldehyde pollution problem. *Water and Environ J.*

Keiser, A., Häberli, M. & Stamp, P. 2012. Quality deficiencies on potato (*Solanum tuberosum* L.) tubers caused by *Rhizoctonia solani*, wireworms (*Agriotes ssp.*) and slugs (*Deroceras reticulatum, Arion hortensis*) in different farming systems. *Field Crops Res,* 128, 147-155.

Kennedy, G. Y. 1959. A porphyrin pigment in the integument of *Arion ater* (L.). *Journal of the Marine Biological Association of the United Kingdom,* 38, 27-32.

Kim, N., Choo, Y. M., Lee, K. S., Hong, S. J., Seol, K. Y., Je, Y. H., Sohn, H. D. & Jin, B. R. 2008. Molecular cloning and characterization of a glycosyl hydrolase family 9 cellulase distributed throughout the digestive tract of the cricket *Teleogryllus emma*. *Comp Biochem Physiol B Biochem Mol Bio,* 150, 368-376.

King, A. J., Cragg, S. M., Li, Y., Dymond, J., Guille, M. J., Bowles, D. J., Bruce, N. C., Graham, I. A. & Mcqueen-Mason, S. J. 2010. Molecular insight into

lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proceedings of the National Academy of Sciences of the United States of America,* 107, 5345-50.

Klein-Marcuschamer, D., Oleskowicz-Popiel, P., Simmons, B. A. & Blanch, H. W. 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnology and Bioengineering,* 109, 1083-1087.

Konig, H., Li, L. & Frohlich, J. 2013. The cellulolytic system of the termite gut. *Appl Microbiol Biotechnol,* 97, 7943-7962.

Kurian, J. K., Nair, G. R., Hussain, A. & Raghavan, G. S. V. 2013. Feedstocks, logistics and pre-treatment processes for sustainable lignocellulosic biorefineries: A comprehensive review. *Renewable & Sustainable Energy Reviews,* 25, 205-219.

Kwon, K. S., Lee, J., Kang, H. G. & Hah, Y. C. 1994. Detection of Beta-Glucosidase Activity in Polyacrylamide Gels with Esculin as Substrate. *Appl Environ Microbiol,* 60, 4584-4586.

Laemmli, U. K. 1970. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature,* 227, 680-685.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth,* 9, 357-359.

Laser, M., Schulman, D., Allen, S. G., Lichwa, J., Antal, M. J. & Lynd, L. R. 2002. A comparison of liquid hot water and steam pretreatments of sugar cane bagasse for bioconversion to ethanol. *Bioresource Technology,* 81, 33-44.

Lebo, S. E., Gargulak, J. D. & Mcnally, T. J. 2000. Kirk-Othmer Encyclopedia of Chemical Technology. 85-91.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels,* 6.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Li, H., Wu, S., Wirth, S., Hao, Y., Wang, W., Zou, H., Li, W. & Wang, G. 2014. Diversity and activity of cellulolytic bacteria, isolated from the gut contents of grass carp (*Ctenopharyngodon idellus)* (*Valenciennes*) fed on Sudan grass (*Sorghum sudanense*) or artificial feedstuffs. *Aquaculture Research.*

Li, L.-L., Mccorkle, S. R., Monchy, S., Taghavi, S. & Van Der Lelie, D. 2009b. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels,* 2.

Li, L., Li, K., Wang, Y., Chen, C., Xu, Y., Zhang, L., Han, B., Gao, C. 2015. Metabolic engineering of *Enterobacter cloacae* for high-yield production of enantiopure (2R,3R)-2,3-butanediol from lignocellulose-derived sugars. *Metabolic Engineering,* 28**,** 19-27.

Li, M., Zhou, M., Adamowicz, E., Basarab, J. A. & Guan, L. L. 2012a. Characterization of bovine ruminal epithelial bacterial communities using 16S rRNA sequencing, PCR-DGGE, and qRT-PCR analysis. *Veterinary microbiology,* 155**,** 72-80.

Li, S., Yang, X., Yang, S., Zhu, M. & Wang, X. 2012b. Technology Prospecting on Enzymes: Application, Marketing and Engineering. *Computational and Structural Biotechnology Journal,* 2**,** e201209017.

Liu, C. & Wyman, C. E. 2003. The Effect of Flow Rate of Compressed Hot Water on Xylan, Lignin, and Total Mass Removal from Corn Stover. *Industrial & Engineering Chemistry Research,* 42**,** 5409-5416.

Liu, Z. L., Slininger, P. J., Dien, B. S., Berhow, M. A., Kurtzman, C. P. & Gorsich, S. W. 2004. Adaptive response of yeasts to furfural and 5-hydroxymethylfurfural and new chemical evidence for HMF conversion to 2,5-bis-hydroxymethlfuran. *Journal of Industrial Microbiology & Biotechnology,* 31**,** 345-352.

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research,* 42**,** D490-D495.

Lynd, L. R., Laser, M. S., Brandsby, D., Dale, B. E., Davison, B., Hamilton, R., Himmel, M., Keller, M. *et al.* 2008. How biotech can transform biofuels. *Nature Biotechnology,* 26, 169-172.

Lynd, L. R., Van Zyl, W. H., Mcbride, J. E. & Laser, M. 2005. Consolidated bioprocessing of cellulosic biomass: an update. *Current Opinion in Biotechnology,* 16, 577-583.

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, 506-577.

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V. *et al.* 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology,* 13, 614-629.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011,* 17.

Mateo, N., Nader, W. & Tamayo, G. 2001. Bioprospecting. *Encyclopedia of biodiversity,* 1, 471-87.

Mechaly, A., Fierobe, H. P., Belaich, A., Belaich, J. P., Lamed, R., Shoham, Y. & Bayer, E. A. 2001. Cohesin-dockerin interaction in cellulosome assembly: a single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *The Journal of biological chemistry,* 276, 9883-8.

Menon, V., Prakash, G., Prabhune, A. & Rao, M. 2010. Biocatalytic approach for the utilization of hemicellulose for ethanol production from agricultural residue using thermostable xylanase and thermotolerant yeast. *Bioresource Technology,* 101, 5366-5373.

Menon, V. & Rao, M. 2012. Trends in bioconversion of lignocellulose: Biofuels, platform chemicals & biorefinery concept. *Progress in Energy and Combustion Science,* In Press, 1-29.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., *et al.* 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics,* 9, 386.

Mikaelyan, A., Strassert, J. F. H., Tokuda, G. & Brune, A. 2014. The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (*Nasutitermes spp.*). *Environmental Microbiology,* 16, 2711-2722.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D. & Marshall, D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics,* 14, 193-202.

Modig, T., Liden, G. & Taherzadeh, M. J. 2002. Inhibition effects of furfural on alcohol dehydrogenase, aldehyde dehydrogenase and pyruvate dehydrogenase. *Biochemical Journal,* 363, 769-776.

Mohnen, D. 2008. Pectin structure and biosynthesis. *Current Opinion in Plant Biology,* 11, 266-277.

Mohr, A. & Raman, S. 2013. Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. *Energy Policy,* 63, 114-122.

Murashima, K., Kosugi, A. & Doi, R. H. 2002. Determination of subunit composition of *Clostridium cellulovorans* cellulosomes that degrade plant cell walls. *Applied and Environmental Microbiology,* 68, 1610-1615.

Nadarasah, G. & Stavrinides, J. 2011. Insects as alternative hosts for phytopathogenic bacteria. *FEMS Microbiology Reviews,* 35, 555-575.

Naik, S. N., Goud, V. V., Rout, P. K. & Dalai, A. K. 2010. Production of first and second generation biofuels: A comprehensive review. *Renewable and Sustainable Energy Reviews,* 14, 578-597.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*.

Ni, J., Takehara, M., Miyazawa, M. & Watanabe, H. 2007. Random exchanges of non-conserved amino acid residues among four parental termite cellulases by family shuffling improved thermostability. *Protein engineering design selection PEDS,* 20, 535-542.

O'neill, M. A., Ishii, T., Albersheim, P. & Darvill, A. G. 2004. Rhamnogalacturonan II: Structure and function of a borate cross-linked cell wall pectic polysaccharide. *Annual Review of Plant Biology,* 55**,** 109-139.

Ohkuma, M. 2008. Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends in Microbiology,* 16**,** 345-352.

Olofsson, K., Bertilsson, M. & Liden, G. 2008. A short review on SSF - an interesting process option for ethanol production from lignocellulosic feedstocks. *Biotechnology for Biofuels,* 1.

Olson, D. G., Mcbride, J. E., Shaw, A. J. & Lynd, L. R. 2012. Recent progress in consolidated bioprocessing. *Current Opinion in Biotechnology,* 23**,** 396-405.

Oppert, C., Klingeman, W. E., Willis, J. D., Oppert, B. & Jurat-Fuentes, J. L. 2010. Prospecting for cellulolytic activity in insect digestive fluids. *Comp Biochem Physiol B Biochem Mol Biol,* 155**,** 145-54.

Orfao, J. J. M., Antunes, F. J. A. & Figueiredo, J. L. 1999. Pyrolysis kinetics of lignocellulosic materials - three independent reactions model. *Fuel,* 78**,** 349-358.

Pandey, M. P. & Kim, C. S. 2011. Lignin Depolymerization and Conversion: A Review of Thermochemical Methods. *Chemical Engineering Technology,* 34**,** 29-41.

Parawira, W. & Tekere, M. 2011. Biotechnological strategies to overcome inhibitors in lignocellulose hydrolysates for ethanol production: review. *Critical Reviews in Biotechnology,* 31**,** 20-31.

Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R. & Uberbacher, E. C. 2010. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology,* 20**,** 1574-1584.

Peng, L. & Chen, Y. 2011. Conversion of paper sludge to ethanol by separate hydrolysis and fermentation (SHF) using *Saccharomyces cerevisiae*. *Biomass and Bioenergy,* 35**,** 1600-1606.

Petiot, E. 2008. On the road to cost-competitive cellulosic ethanol. *Chimica Oggi-Chemistry Today,* 26**,** 20-22.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry,* 25**,** 1605-1612.

Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., Mchardy, A. C., Cheng, J.-F., *et al.* 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,* 107**,** 14793-14798.

Pope, P. B., Mackenzie, A. K., Gregor, I., Smith, W., Sundset, M. A., Mchardy, A. C., Morrison, M. & Eijsink, V. G. 2012. Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. *PLoS One,* 7**,** e38571.

Qin, J. J., Li, R. Q., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., *et al.* 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464**,** 59-U70.

Reijnders, L. & Huijbregts, M. a. J. 2007. Life cycle greenhouse gas emissions, fossil fuel demand and solar energy conversion efficiency in European bioethanol production for automotive purposes. *Journal of Cleaner Production,* 15**,** 1806-1812.

Ridley, B. L., O'neill, M. A. & Mohnen, D. A. 2001. Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. *Phytochemistry,* 57**,** 929-967.

Riekki, R., Palomaki, T., Virtaharju, O., Kokko, H., Romantschuk, M. & Saarilahti, H. T. 2000. Members of the *amylovora* group of *Erwinia* are cellulolytic and possess genes homologous to the type II secretion pathway. *Molecular and General Genetics,* 263**,** 1031-1037.

Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. 2004. Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics,* 38**,** 525-52.

Romaní, A., Pereira, F., Johansson, B. & Domingues, L. 2015. Metabolic engineering of *Saccharomyces cerevisiae* ethanol strains PE-2 and CAT-1 for efficient lignocellulosic fermentation. *Bioresource Technology,* 179**,** 150-158.

Rowson, B., Anderson, R., Turner, J. A. & Symondson, W. O. C. 2014. The Slugs of Britain and Ireland: Undetected and Undescribed Species Increase a Well-Studied, Economically Important Fauna by More Than 20%. *PLoS ONE,* 9, e91907.

Ruijssenaars, H. J. & Hartmans, S. 2001. Plate screening methods for the detection of polysaccharase-producing microorganisms. *Appl Microbiol Biotechnol,* 55, 143-149.

Ruiz, E., Cara, C., Manzanares, P., Ballesteros, M. & Castro, E. 2008. Evaluation of steam explosion pre-treatment for enzymatic hydrolysis of sunflower stalks. *Enzyme and microbial technology,* 42, 160-6.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., *et al.* 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology,* 5, e77.

Russell, J. A., Moreau, C. S., Goldman-Huertas, B., Fujiwara, M., Lohman, D. J. & Pierce, N. E. 2009a. Bacterial gut symbionts are tightly linked with the evolution of herbivory in ants. *Proceedings of the National Academy of Sciences of the United States of America,* 106, 21236-21241.

Russell, J. B., Muck, R. E. & Weimer, P. J. 2009b. Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen. *Fems Microbiology Ecology,* 67, 183-197.

Saha, B. C., Iten, L. B., Cotta, M. A. & Wu, Y. V. 2005. Dilute acid pretreatment, enzymatic saccharification and fermentation of wheat straw to ethanol. *Process Biochemistry,* 40, 3693-3700.

Sánchez, O. J. & Cardona, C. A. 2008. Trends in biotechnological production of fuel ethanol from different feedstocks. *Bioresource Technology,* 99, 5270-5295.

Sarma, S. J., Brar, S. K., Le Bihan, Y., Buelna, G. & Soccol, C. R. 2014. Mitigation of the inhibitory effect of soap by magnesium salt treatment of crude glycerol – A novel approach for enhanced biohydrogen production from the biodiesel industry waste. *Bioresource Technology,* 151, 49-53.

Scheller, H. V. & Ulvskov, P. 2010. Hemicelluloses. *Annu Rev Plant Biol,* 61, 263-89.

Schwarz, W. H., Bronnenmeier, K., Grabnitz, F. & Staudenbauer, W. L. 1987. Activity Staining of Cellulases in Polyacrylamide Gels Containing Mixed Linkage Beta-Glucans. *Analytical Biochemistry,* 164**,** 72-77.

Scrivener, A. & Slaytor, M. 1994. Cellulose digestion in *Panesthia cribrata Saussure*: does fungal cellulase play a role? *Comp Biochem Physiol B Biochem Mol Bio,* 107**,** 309-315.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods,* 9**,** 811-4.

Seric Jelaska, L., Jurasovic, J., Brown, D. S., Vaughan, I. P. & Symondson, W. O. 2014. Molecular field analysis of trophic relationships in soil-dwelling invertebrates to identify mercury, lead and cadmium transmission through forest ecosystems. *Mol Ecol,* 23**,** 3755-66.

Shallom, D. & Shoham, Y. 2003. Microbial hemicellulases. *Current Opinion in Microbiology,* 6**,** 219-228.

Shaw, A. J., Covalla, S. F., Hogsett, D. A. & Herring, C. D. 2011. Marker removal system for *Thermoanaerobacterium saccharolyticum* and development of a markerless ethanologen. *Applied and Environmental Microbiology,* 77**,** 2534-2536.

Shi, W., Xie, S., Chen, X., Sun, S., Zhou, X., Liu, L., Gao, P., Kyrpides, N. C., *et al*. 2013. Comparative Genomic Analysis of the Endosymbionts of Herbivorous Insects Reveals Eco-Environmental Adaptations: Biotechnology Applications. *PLoS Genet,* 9**,** e1003131.

Shi, W. B., Ding, S. Y. & Yuan, J. S. 2011. Comparison of insect gut cellulase and xylanase activity across different insect species with distinct food sources. *Bioenergy Research,* 4**,** 1-10.

Shibuya, H., Kaneko, S. & Hayashi, K. 2000. Enhancement of the thermostability and hydrolytic activity of xylanase by random gene shuffling. *Biochem. J,* 349**,** 651-656.

Shin, H. D., Mcclendon, S., Vo, T. & Chen, R. R. 2010. *Escherichia coli* binary culture engineered for direct fermentation of hemicellulose to a biofuel. *Applied and Environmental Microbiology,* 76**,** 8150-8159.

Simon, C., Wiezer, A., Strittmatter, A. W. & Daniel, R. 2009. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Applied and Environmental Microbiology,* 75, 7519-7526.

Slotsbo, S., Hansen, L. M. & Holmstrup, M. 2011. Low temperature survival in different life stages of the Iberian slug, Arion lusitanicus. *Cryobiology,* 62, 68-73.

Smith, H. 1977. The Molecular Biology of Plant Cells. 14, 6-12.

Srinivasan, S. 2009. The food v. fuel debate: A nuanced view of incentive structures. *Renewable Energy,* 34, 950-954.

Stambuk, B. U., Eleutherio, E. C., Florez-Pardo, L. M., Souto-Maior, A. M. & Bon, E. P. 2008. Brazilian potential for biomass ethanol: Challenge of using hexose and pentose cofermenting yeast strains. *Journal of Scientific and Industrial Research,* 67, 918.

Sticklen, M. B. 2008. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics,* 9, 433-443.

Stinson, J. A. & Ham, R. K. 1995. Effect of lignin on the anaerobic decomposition of cellulose as determined through the use of a biochemical methane potential method. *Environmental science & technology,* 29, 2305-10.

Štursová, M., Žifčáková, L., Leigh, M. B., Burgess, R. & Baldrian, P. 2012. *Cellulose utilization in forest litter and soil: identification of bacterial and fungal decomposers.*

Suen, G., Scott, J. J., Aylward, F. O., Adams, S. M., Tringe, S. G., Pinto-Tomás, A. A., Foster, C. E., Pauly, M., *et al.* 2010. An Insect Herbivore Microbiome with High Plant Biomass-Degrading Capacity. *PLoS Genet,* 6, e1001129.

Sugiyama, J., Okano, T., Yamamoto, H. & Horii, F. 1990. Transformation of Valonia cellulose crystals by an alkaline hydrothermal treatment. *Macromolecules,* 23, 3196-3198.

Sun, J.-Z. & Scharf, M. E. 2010. Exploring and integrating cellulolytic systems of insects to advance biofuel technology. *Insect Science,* 17, 163-165.

Sun, Y. & Cheng, J. 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresour Technol,* 83, 1-11.

Sweet, M. S. & Winandy, J. E. 1999. Influence of degree of polymerization of cellulose and hemicellulose on strength loss in fire-retardant-treated southern pine. *Holzforschung,* 53**,** 311-317.

Talebnia, F., Karakashev, D. & Angelidaki, I. 2010. Production of bioethanol from wheat straw: An overview on pretreatment, hydrolysis and fermentation. *Bioresource Technology,* 101**,** 4744-4753.

Tamaru, Y. & Doi, R. H. 2001. Pectate lyase A, an enzymatic subunit of the *Clostridium cellulovorans* cellulosome. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 4125-9.

Teather, R. M. & 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**,** 777-780.

Teeri, T. T., Koivula, A., Linder, M., Wohlfahrt, G., Divne, C. & Jones, T. A. 1998. *Trichoderma reesei* cellobiohydrolases: why so efficient on crystalline cellulose? *Biochemical Society transactions,* 26**,** 173-8.

Teeri, T. T., Svensson, B., Gilbert, H. J., Feizi, T., Bolan, D. N., Szabo, L., Xie, H., Williamson, M. P., Simpson, P. J., Jamal, S., Boraston, A. B., Kilburn, D. J. & Warren, R. a. J. 2002. Carbohydrate Bioengineering.

Todaka, N., Moriya, S., Saita, K., Hondo, T., Kiuchi, I., Takasu, H., Ohkuma, M., Piero, C., *et al.* 2007. Environmental cDNA analysis of the genes involved in lignocellulose digestion in the symbiotic protist community of *Reticulitermes speratus. Fems Micriobiology Ecology,* 59, 592-599.

Tokuda, G. & Watanabe, H. 2007. Hidden cellulases in termites: revision of an old hypothesis. *Biology Letters,* 3**,** 336-339.

Tolonen, A. C., Chilaka, A. C. & Church, G. M. 2009. Targeted gene inactivation in *Clostridium phytofermentans* shows that cellulose degradation requires the family 9 hydrolase Cphy3367. *Molecular Microbiology,* 74**,** 1300-1313.

Tomas-Pejo, E., Oliva, J. M., Ballesteros, M. & Olsson, L. 2008. Comparison of SHF and SSF processes from steam-exploded wheat straw for ethanol production by xylose-fermenting and robust glucose-fermenting *Saccharomyces cerevisiae* strains. *Biotechnology and Bioengineering,* 100**,** 1122-1131.

216

Toth, I. K., Van Der Wolf, J. M., Saddler, G., Lojkowska, E., Hélias, V., Pirhonen, M., Tsror, L. & Elphinstone, J. G. 2011. *Dickeya* species: an emerging problem for potato production in Europe. *Plant Pathology,* 60, 385-399.

Ushakova, N. A., Belov, L. P., Varshavskiĭ, A. A., Kozlova, A. A., Kolganova, T. V., Bulygina, E. S. & Turova, T. P. 2003. Cellulose decomposition under nitrogen deficiency by bacteria isolated from the intestines of phytophagous vertebrates. *Mikrobiologiia,* 72, 400-6.

Van Tilbeurgh, H., Tomme, P., Claeyssens, M., Bhikhabhai, R. & Pettersson, G. 1986. Limited proteolysis of the cellobiohydrolase I from *Trichoderma reesei. FEBS Letters,* 204, 223-227.

Van Zyl, W. H., Lynd, L. R., Den Haan, R. & Mcbride, J. E. 2007. Consolidated bioprocessing for bioethanol production using *Saccharomyces cereviside. Biofuels,* 108, 205-235.

Vanderhart, D. L. & Atalla, R. H. 1984. Studies of Microstructure in Native Celluloses Using Solid-state 13C NMR. *Macromolecules,* 17, 1465-1472.

Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., Van Elsas, J. D., Bailey, M. J., Nalin, R. & Philippot, L. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Micro,* 7, 252-252.

Wang, X., Yomano, L. P., Lee, J. Y., York, S. W., Zheng, H., Mullinnix, M. T., Shanmugam, K. T. & Ingram, L. O. 2013. Engineering furfural tolerance in *Escherichia coli* improves the fermentation of lignocellulosic sugars into renewable chemicals. *Proceedings of the National Academy of Sciences,* 110, 4021-4026.

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., Mchardy, A. C., *et al.* 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature,* 450, 560-565.

Watanabe, H., Noda, H., Tokuda, G. & Lo, N. 1998. A cellulase gene of termite origin. *Nature,* 394, 330-1.

Watanabe, H. & Tokuda, G. 2001. Animal cellulases. *Cellular and Molecular Life Sciences CMLS,* 58, 1167-1178.

Watanabe, H. & Tokuda, G. 2010. Cellulolytic systems in insects. *Annu Rev Entomol,* 55**,** 609-32.

Wei, Y. D., Lee, K. S., Gui, Z. Z., Yoon, H. J., Kim, I., Zhang, G. Z., Guo, X., Sohn, H. D., *et al.* 2006. Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, *Apriona germari. Comp Biochem Physiol B Biochem Mol Biol,* 145**,** 220-9.

Wei, Z., Laby, R., Zumoff, C., Bauer, D., He, S., Collmer, A. & Beer, S. 1992. Harpin, elicitor of the hypersensitive response produced by the plant pathogen *Erwinia amylovora. Science,* 257**,** 85-88.

Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 1991. 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology,* 173**,** 697-703.

Wenzel, M., Schönig, I., Berchtold, M., Kämpfer, P. & König, H. 2002. Aerobic and facultatively anaerobic cellulolytic bacteria from the gut of the termite *Zootermopsis angusticollis. Journal of Applied Microbiology,* 92**,** 32-40.

Whetten, R. & Sederoff, R. 1995. Lignin Biosynthesis. *Plant Cell,* 7**,** 1001-1013.

Willis, J. D., Klingeman, W. E., Oppert, C., Oppert, B. & Jurat-Fuentes, J. L. 2010a. Characterization of cellulolytic activity from digestive fluids of *Dissosteira carolina* (*Orthoptera: Acrididae*). *Comp Biochem Physiol B Biochem Mol Biol,* 157**,** 267-72.

Willis, J. D., Oppert, C. & Jurat-Fuentes, J. L. 2010b. Methods for discovery and characterization of cellulolytic enzymes from insects. *Insect Science,* 17**,** 184-198.

Wingren, A., Galbe, M. & Zacchi, G. 2003. Techno-economic evaluation of producing ethanol from softwood: Comparison of SSF and SHF and identification of bottlenecks. *Biotechnology Progress,* 19**,** 1109-1117.

Xia, Y., Ju, F., Fang, H. H. P. & Zhang, T. 2013. Mining of Novel Thermo-Stable Cellulolytic Genes from a Thermophilic Cellulose-Degrading Consortium by Metagenomics. *Plos One,* 8.

Xiros, C. & Christakopoulos, P. 2009. Enhanced ethanol production from brewer's spent grain by a *Fusarium oxysporum* consolidated system. *Biotechnology for Biofuels,* 2.

Xu, Q., Singh, A. & Himmel, M. E. 2009. Perspectives and new directions for the production of bioethanol using consolidated bioprocessing of lignocellulose. *Current Opinion in Biotechnology,* 20, 364-371.

Yoneda, Y., Krainz, K., Liebner, F., Potthast, A., Rosenau, T., Karakawa, M. & Nakatsubo, F. 2008. "Furan Endwise Peeling" of Celluloses: Mechanistic Studies and Application Perspectives of a Novel Reaction. *European Journal of Organic Chemistry,* 2008, 475-484.

Zakzeski, J., Bruijnincx, P. C. A., Jongerius, A. L. & Weckhuysen, B. M. 2010. The catalytic valorization of lignin for the production of renewable chemicals. *Chemical reviews,* 110, 3552-99.

Zeikus, J. G. 1981. Lignin Metabolism and the Carbon Cycle. *In:* ALEXANDER, M. (ed.) *Advances in Microbial Ecology.* Springer US.

Zerbino, D. R. & Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18, 821-829.

Zhang, D., Lax, A. R., Raina, A. K. & Bland, J. M. 2009. Differential cellulolytic activity of native-form and C-terminal tagged-form cellulase derived from *Coptotermes formosanus* and expressed in *E. coli. Insect Biochemistry and Molecular Biology,* 39, 516-522.

Zhang, M., Eddy, C., Deanda, K., Finkelstein, M. & Picataggio, S. 1995. Metabolic Engineering of a Pentose Metabolism Pathway in Ethanologenic *Zymomonas mobilis. Science,* 267, 240-243.

Zheng, Y., Pan, Z. & Zhang, R. 2009. Overview of biomass pretreatment for cellulosic ethanol production. 2, 51-68.

Zhou, S. D., Davis, F. C. & Ingram, L. O. 2001. Gene integration and expression and extracellular secretion of *Erwinia chrysanthemi* endoglucanase CelY (celY) and CelZ (celZ) in ethanologenic *Klebsiella oxytoca* P2. *Applied and Environmental Microbiology,* 67, 6-14.

Zhou, S. D. & Ingram, L. O. 2001. Simultaneous saccharification and fermentation of amorphous cellulose to ethanol by recombinant *Klebsiella oxytoca* SZ21 without supplemental cellulase. *Biotechnology Letters,* 23, 1455-1462.

Zhu, L., Wu, Q., Dai, J., Zhang, S. & Wei, F. 2011. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proceedings of the National Academy of Sciences,* 108, 17714-17719.

Zhu, W., Lomsadze, A. & Borodovsky, M. 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research,* 38, e132.

Zhu, Y., Lee, Y. Y. & Elander, R. T. 2004. Dilute-acid pretreatment of corn stover using a high-solids percolation reactor. *Applied Biochemistry And Biotechnology,* 117, 103-114.

Zimmer, M. & Bartholmé, S. 2003. Bacterial endosymbionts in *Asellus aquaticus* (Isopoda) and *Gammarus pulex* (Amphipoda) and their contribution to digestion. *Limonology And Oceanography,* 48, 2208-2213.

Zverlov, V. V., Fuchs, K.-P. & Schwarz, W. H. 2002. Chi18A, the endochitinase in the cellulosome of the thermophilic, cellulolytic Bacterium *Clostridium thermocellum. Applied and Environmental Microbiology,* 68, 3176-3179.

# Appendix 1

**16s rRNA sequences for cultured microbes and DGGE microbe identifications**

**The 16s rRNA sequences for isolated bacteria and those identified using DGGE (also showing the name of the closest matching bacteria in the NCBI databses)**

**Isolate 1      Citrobacter braakii**

NNNNTNNNNCGANTACTATAGGGCGAATTGGGCCCTCTAGATGCATGCTNGAGCGGCCGCCAGTGTGATGGATATC
TGCAGAATTCGGCTTAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGG
TAGCACAGAGAGCTTGCTCTTGGGTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGTCCGATGGAGGGG
GATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGTGGGGGACCTTCGGGCCTCACACCAT
CGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAG
GATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGG
GCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGAGGAGGAAG
GCATTGTGGTTAATAACCGCAGTGATTGACGTTACTCGCAGGAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGG
TAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTG
AAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAACTGGCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTTCCAGG
TGTAGCGGTGAAATGCGTGAAGATCTGGGAGGAATACCCGG

**Isolate 2      Salmonella enterica**

NNNNNNNCTGCGCGGCCTAAACATGCAAGTCGAACGGTAACAGAAAGCAGCTTGCTGCTTTGCTGACGAGTGGCGG
ACGGGTGAGTAATGTCTGGGAAACTGCCCGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACG
TCACAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCAGATGTGCCCATATGGGATTATCTAGTATGTGGGGT
AACGGCTCACCTACGCGACTATCCCTATCTGGTCTGAGAGGATGACCACCCACACTGGAACTGACACACGGTCCACAC
TCCTACGGGAGGCAGCAGTGGGGAATATTGCACAGTGGGCGCACGCCTGATGCACCCGTGCCGCGTGTATGAAAAA
GGCCTTCGGGTTGTAAAGTACTTTCCGCGAGGAAGAAGGTGTTGCGGTTAATAACCGCCGCTTTTGACGTTACTCGCA
GAANAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACAGAGGGTGCAAGCGTTAATCAGAATTACTGGGC
GTAAAGCGCACGCGCGGTCTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTTGACACTGG
CGCGCTTGTGTCTTGTAGAGGGGGGTAGATTTCCGGGTGTAGCGGTGAAATGCATAGAGATCGGAAGAACACCGGT
GGCGAGTGCNCCCCTGTGGACAAACACTGACTCTCAAGTGAAACCGCGTGGATACAAGGGGGGTAGATACCCTGGTA
GTACATGCTGAAAACTATGCGATTTTAACTTGGCCCCGTGAACTTGTAGCCTCGACCACAAACCCTTGACTTGCCCCCT
CCAGAGGNCCCCCCCAGATGGTGACAGTTAATTNAACTTGCGTGATTACACCACCGCCCGGGCGNCCTGACGCTTAAT
TATTGCGAANNAAGGATAGCTCCACATATATATTCGCCACCGGGNNGGANN

**Isolate 3      Klebsiella pneumonia**

NNNNNCTGGCGCAGGCCTAACACATGCAAGTCGAGCGGTAGCACAGAGAGCTTGCTCTCGGGTGACGAGCGGCGG
ACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGT
CGCAAGACCAAAGTGGGGGACCTTCGGGCCTCATGCCATCAGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTA
ACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGAC
TCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAA
GGCCTTCGGGTTGTAAAGCACTTTCAGCGGGGAGGAAGGCGTTGAGGTTAATAACCTTGTCGATTGACGTTACCCGC
AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGG
CGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAACTG
GCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATCCAGGTGTAGCGGTGAAATGCGTANAGATCTGGAGGAATACCGG
TGGCGAAGCGGCCCCCTGGACAAGACTGACGCTCACGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACTCTGG
TAGTCCACGCCGTACACGATGTCAATTTGTACGTTGTGCCCTTGAGGCTTGGCCTCCCGGAGCAAACGANTTAATCTG
CCGNNTGGGGAAGTACGGTCNNAGGTTAAANCTGANAATTAATTAGAGGGGGGNGCNCANACGCNGGGGAGGCANT
GGAGTCACAANTAAGTCANANTANANNANGNTACNNGGTANNANCCNNCAGGANTTN

## Isolate 4     Serratia marcescens

TNNNNNNNCNANTCNATAGGGCGAATTGGGCCCTCTAGATGCATGCTTCGAGCGGCCGCCAGTGTGATGGATATCT
GCAGAATTCGGCTTAAAGATGTGATCCAGCCCCAAGTGCCCCCTGGGTTACCGTGTTACAACTTCACCTCAAGCGGTA
ATCACAAANTGGTAAGCCGCCCTCCTGACGTTAAGCTAACTGCTTATTTTGGTCCCGAATACTGTGGGGTGACGGGCG
ATGTGTACTGGGCCCGGGAACTTATTCCCCGTAACATTCTGATATACGATTACTAGACCTTCCGACTTCATGGAGTCGA
GTTGCCNACTCGGTCCAGACTACGACGTACTTTATGAGGTCCGCTTGCTCTCATCAGTTCGCTTCTCTTTGTATACGCCA
TTGTAGCACGTGTGTAGCCCTACTCCTAACGGCCATGATGACTTGACGTCGTCCCCACCTTCCTCCGGTTTATCACCGG
CTGTCTCCTTTGAGTTCCCGACCAAATCGCTGGCAACAAAGGATAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACA
TTTCACAACACGAGCTGACGACAGCCATGCACACCTGTCTCACAGTTCCCGAANGCACCAATCCATCTCTGGAAAGTTC
TCTGGATGTCAAGAGTANGTAAGGTTCTTCGCGTTGCATCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCC
GTCAATTCATTTGAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGTCGATTTAACGCGTTAGCTCCGGAGCCACGCCT
CAAGGGCACACCTCCAAATCGACATCGTTTACACGTGGACTACCAGGGTATCTAATCCTGTTTGCTCCCACGCTATCGC
ACCTGAACGTCAGTCTTTGTCACAGGAGCACGACTTCTCCACCGGANTTACCTCAAATCTG

## Isolate 5     Buttiauxella agrestis

NNNNNNNGNGGNNGNCTACCATGCAAGTCGAGCGGTAGCACAGGGAACTTGGTCCTGGGTGAGAAGAGGGGGAG
GGGGGAGAAAAGTGTGGGAAAATGCGTGAGGGAGGGGGATAACTATTGGGAACGGTAGATAATATAGCATAAAGT
CTTTTGAGAAAAAAGGGGGAGATTTTGGGCTCTTGTGATCTGATGTGTGCAGATGGGATTATATATTAGGTGAGGTA
ATGGGTCTCCTATGGGACAATATCTATCTGGGGTGAGAGGATAAGAAGACACACTGGAAATGAGACACGGTGTACAC
ACTTATAGGAGGGAGAAGTGGGGAATATTGTANAATGGGGGNAAGAGTGATGTGGAGATGTGGTGTGTATGAAAA
AAGGGTTTGGGTTGTAAAGTATTTTTTGAGAGGAGAAAAGGGTTGTGGTTAATAAAAGCGGTGATTGAGGTTATTCT
CAGAAAAAAAACAGGGTAAATCTGTGTGAGAAGACGCGGTAATACAGAGGGTGCGAGAGTTAATCTGAGTTATAGG
GGGTAAAGAGCACACAGGCGGTTTGTGAAGTCGGATGTGAAATCCCCCCGGCTCCCCCCGTGAACTGCGTTCTAAACT
GGGAAGCTATAGTCTCGTGTAGAGGGGTAGAATTCCCCGTGTATCGGTGAGATGCGTATAGATCTCGAGGAAAACCG
CTGGCGAAAGCGGCCCCCTGGGACAAAAACAGACACTCTCGTGCGAAAACGTGGGGAGAAAACAGGAGAATATACC
CTGGGATTCCCCACCCTAAAAAATGTGTACTTGTAGGTTGTTCCCTTGTGAAGTGGGTTCCCGAGCTCACGCGTTAAAC
TCNCNCCCTGGGGGAGTACAGCCCCCACGTTAAAATCCACAGTATTGACGGGGGCCCGCCACANCGGCTGGAGCATG
TGGTTTTATTTCTATGCACGCGAAAACCTCATCCTCCTCCTTGTNATCTCCAGAAATTTGGCGAANATGACCTCTNTGG
CTCTCTGCGAACTNTGGAGACAAGAGNNGCCNTCGGGTGTNNTCNCCTTCTCG

## Isolate 6     Serratia liquefaciens

NNNNNNNNNNGCNGNCTACCCATGCAAGTCGAGCGGTAGCACAGGAAGCTTGGTCCTGGGTGAGAAGAGGCGG
AGAGGGGAGAAAAGTGTGGGAAACTGCCTGAGGGAGGGGGATAACTACTGTNAACGGTATCTAATACCCCATAACG
TCTTCTGAGCAAAAAGGGGGACCTTTTGGCCTCTTGTCATCACATGTGCCCACATGGGATTATATATTATGTGAGGTAA
TGGGTCACCTATGCGACAATCCCTATCTGGGCTGAGAGGATGACAACCCACACTGGAACTGAGACACGGTCCCCACTC
CTACAGGAGGCGCCAGTGGGGAATATTGTGCAATGGGGGCAAGCCTGATGTACCCATGTGGCGTGTATGAAAAAAG
CGTTTTGGTTGTAAAATATTTTTTCAGAGGAGAAAAGCGTTGTGGTTAATAACCCCGGTGATTGAGATTACTCTCACAA
AAAACACCGGCTAAATCTGTGCGACCACCCGCGGTAATATAGAGGGTGTAAGCGTTAATCTCAATTACTGGGCGTAAA
GCGCACACGCGGTTTGTCAAGTCTCATGTGAAATCCCCCGGCTCTCCCTGGGAACTGCGTTCTAAACTGGGAAGCT
ATAGTCTTGTATAGAGGGGTATAATTCCCCGTGTATCGCTGAAATGCGTATAGATCTCGAGGAATATCGGTGGCGAAA
GCGGCCCCCGGACAAAGACTGACACTCACGTGCGAAAGCGTGTGGAGCACACACGAGTATATACCCTGTGTATTCT
ACGCCNCATACGATGTCTACTTGTAGGTTGTTGCCCCTGTGGAGTGTGTTTCTGGAGCTAACACGTTATATCTACCGCC
TGTGGAGTACGCNCGCCAAGGTGAAAACTCACATGAGTTGANAGGGGGCCCGCACACAGGTGGGAGCATGTGGTTT
TTATTTCTATGTCACGCGAANAAACCTTATCNTACTCTTGAGATCTCCANAATTCTGCNCAGATGTNCTATTGTCNCTTC
GGAACTGTGAAGACCAGTGNNCNATGGCTGGTCTTCCACCTC

## Isolate 7　　Aeromonas hydrophila

NNNNNNNNNGNNNNNCTANACATGCAAGTCGAGCGGCAGCGGGAAAGTAGCTTGCTACTTTTGCCGGCGAGCGGC
GGACGGGTGAGTAATGCCTGGGAAATTGCCCAGTCGAGGGGGATAACAGTTGGAAACGACTGCTAATACCGCATAC
GCCCTACGGGGGAAAGCAGGGGACCTTCGGGCCTTGCGCGATTGGATATGCCCAGGTGGGATTAGCTTGTTGGTGA
GGTAATGGCTCACCAAGGCGACGATCCCTAGCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAGACACGGTCC
AGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCCATGCCGCGTGTGTGA
AGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAGGAGGAAAGGTTGGTAGCTAATAACTGCCAACTGTGACGTTAC
TCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACT
GGGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTGGGAATTGCATTTAAAA
CTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATA
CCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATA
CCCTGGTAGTCCACGCCGTAAACGATGTCGATTTGGAGGCTGTGTCCTTGAGACGTGGCTTCCGGAGCTAACGCGTTA
AATCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCNCGCACAAGCGGTGGAGC
ATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGNNTTGACATGTCTGGAATCCTGTAGAGATACGGGAGT
GCCTTCGGGAATCAGACACAGGTGCTGCATGGNTGTCGTCAGCTCGTGTCGTGAGA

## Isolate 8　　Acinetobacter calcoaceticus

GNNNNGGNGNNGCTTACNATGCAGTCGAGCGGAGTGATGGTGCTTGCTCTATCACTTAGCGGCAGACGGGTGAGTA
ANGCTTATGAATCTGCCTATTAGTGGGGGACAACATTTCCAAAGGAATGCTAATACCGCATACGTCCTACGGGACAAA
GCAGGGGATCTTCAGACCTTGCCCTAATGCATGAACCTAAGTCCGATTAGCTAGTTGGTGGGGTAAAGGCCTACCCAG
GCGACGATCTGTATCGGGTCTGAGAGGAAGATCCGCCACACTGGGACTGAAACACAGCCCAGACTCCTACGGGAGGC
AGCAGTGGGGAATATTGAACAATGGGCGCAAGCCTGATCCAGCCATGCCCCGTGTGTGAATAAGGCCTTAGGGTTGT
AAATCACTTTAACTTAGGAGGAGGCTACTGAAGTTAATACCTTCAAATAGTGGACGTTACTCTCATAATAAAAACCGG
CTAACTCTGTGCCAGCACCCGCGGTAATACAAAGGGTGCAGGCGTTAATCGGATTTACTGGGCGTAAAGCGCGCGTA
GGCGGCTAATTAAGTCAAATGTGAAGTCCCCGACCTTAACTTGGGAATTGCATTCCATACTGGTTAGCTAGAGTGTGG
GAGAGGATGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGATGGCGAAGGCAGCCA
TCTGGCCTAACACTGACGCTGAGCGTGCGAAAGCATGGCGGAGCAAACAGGATTAGATACCCTGGTAGTCCATGCCG
TACACGATGTCTACTAGCCGTTGGGGCCTTTGAGGCTTTAGTGGCGCAGCTAACGCGATAAGTANACCGCCTGGGGG
AGTACGGTCGCACGACTAGATCTCACATGAATTGACTGGGGGCGCGCACACACGGTGGAGCATGTGTGTTGATTTCG
ATGCAACGCGCANAAACTTACCTGGTCTTTGACATANTACGANNTTTCTGAGATGNATTGGTGCCTGCGGNAACTTAC
ATACAGNTGCTGNCATGCTGTCGTCAGCTCGTGTCGTGTAGANGNN

## Isolate 9　　Kluyvera intermedia

NNNNNNNNNNNGGNNGCCTACNCATGCAAGTCGAACGGTAGCAACAGAGAGCTTGCTCTTGGGTGACGAGTGGC
GGACGGGTGAGTAATGTCTGGGAAACTGCCCGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAA
CGTCGCAAGACCAAAGTGGGGGACCTTCGGGCCTCACACCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGG
GTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAA
GAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGAGGAGGAAGGCATTGTGGTTAATAACCTTAGTGATTGACGTTACTC
GCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG
GGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAAC
TGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATAC
CGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATAC
CCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGCGTTAA
GTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCA
TGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTACTCTTGACATCCAGAGAACTTAGCAGAGATGCTTTGGTGCC
TTCGGGAACTCTGAGACAGNGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAAT

## Isolate 10      Buttiauxella agrestis

NNNNNNNNNGNNNGCNAACACATGCAAGTCGAGCGGTAGCACAGGGAGCTTGCTCCTGGGTGACGAGCGGCGGAC
GGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCT
TCGGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCAGATGTGCCCAGATGGGATTAGCTAGTAGGTGAGGTAAT
GGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTC
CTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGG
CCTTCGGGTTGTAAAGTACTTTCAGCGAGGAGGAAGGCATTGTGGTTAATAACCGCAGTGATTGACGTTACTCGCAGA
AGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTA
AAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAACTGGCAG
GCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGG
CGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGT
AGTCCACGCCGTAAACGATGTCNACTTGGAGGTTGTTCCCTTGAGGAGTGGCTTCCGGAGCTAACGCGTTAAGTCGAC
CGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGCCGCACAAGCGGTGGAGCATGTGGTTT
ATTTCGATGCAACGCGAAGAACCTACCTACTCTTGACATCCANAGAATCGCTAGAGATACTTATGCCTTCNGGAACTCT
GAGACAGTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAATTGTTGGTNAGT

## Isolate 11      Citrobacter freundii

NNNNNNNNNNGCAGGCTACACATGCAGTCGAACGGTAGCACAGAGGAGCTTGCTCCTTGGGTGACGAGTGGCGGA
CGGGTGAGTAATGTCTGGGAAACTGCCCGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTC
GCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGGTA
ACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGAC
TCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAA
GGCCTTCGGGTTGTAAAGTACTTTCAGCGAGGAGGAAGGTGTTGTGGTTAATAACCGCAGCGATTGACGTTACTCGC
AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGG
CGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCCGAAACTG
GCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCG
GTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCC
TGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAG
TCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCAT
GTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTACTCTTGACATCCAGAGACTTAGCAGAGATGCTATNGTGCCT
TCGGGACTCTGAGACAGTGCTGCATGNTGTCGTCAGCTCGTGTGTGAATGTGGGTN

## Isolate 12      Enterobacter sp. E6-PCAi

NNNNNTGNNGGNANGCNAACNCATGCAGTCGAGCGGTAGCACGGGGGAGCTTGCTCCCTGGGTGACGAGCGGCG
GACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAAC
GTCGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCAGATGTGCCCAGATGGGATTAGCTAGTAGGTGGG
GTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAA
GAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGAGGAGGAAGGCATTGTGGTTAATAACCACAGTGATTGACGTTACT
CGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG
GGCGTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAAC
TGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATAC
CGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATAC
CCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTTCCCTTGAGGAGTGGCTTCCGGAGCTAACGCGTTAA
GTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTTGACGGGGGCCCGCACAAGCGGTGGAGC
ATGNGGNTTAATTCGATGCAACGCGAAGAANCTTACCTACTCTTGACATCCAGAGAACTTAGCAGAGATGCTTTGGTG
CNTNNNACTCTGANACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTNTGAATNNNN

## UA.a.1 Mycoplasma hyorhinis

NNNNNNNNNANNANGGGCTGACCTACGATTACTAGCGATTCCGACTTCATGTAGTCGAGTTGCAGACTACAATCCGA
ACTGAGATCGATTTTTTGAGGTTTGCTCGTATTCACATAGTTGCTTCCCTTTGTATCGACCATTGTAGCACGTGTGTCGC
CCCACTTGTAAGAGGCATGATGATTTGACGTCGTCCCCACCTTCCTCCTGGTTACCCAGGCAGTCTCCTCAGAGTCCCC
AACTTAATGATGGTAACTAAGGATAAGGGTTGCGCTCGTTGCAGGACTTAACCTAACATCTCACGACACGAGCTGACG
ACAACCATGCACCATCTGTCATTCCGTTAACCTCCACTATATCTCTATAGCATTGCGGAAGATGTCAAAAGTGGGTAAG
GTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCGA

## UA.a.2 Mycoplasma iners

NNNNNNNNNNANNNNGGCTGACCTACGATTACTAGCGATTCCGACTTCATGTAGTCGAGTTGCAGACTACAATCCGA
ACTGAGATCGATTTTTTGAGGTTTGCTCGTATTCGCATAGTTGCTTCCCTTTGTATCGACCATTGTAGCACGTGTGTCGC
CCCACTTGTAAGAGGCATGATGATTTGACGTCGTCCCCACCTTCCTCCTGGTTACCCAGGCAGTCTCCTCAGAGTCCCC
AACTTAATGATGGTAACTAAGGATAAGGGTTGCGCTCGTTGCAGGACTTAACCTAACATCTCACGACACGAGCTGACG
ACAACCATGCACCATCTGTCATTCCGTTAACCTCCGCTATATCTCTATAGCATTGCGGAAGATGTCAAAAGTGGGTAAG
GTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCG

## UA.a.3 Uncultured Citrobacter

TNNNNNNNNATTTCTGATCTACGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGACTACG
ACATACTTTATGAGGTCCGCTTGCTCTCGCGAGGTCGCTTCTCTTTGTATATGCCATTGTAGCACGTGTGTAGCCCTACT
CGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCAGTTTATCACTGGCAGTCTCCTTTGAGTTCCCGGCCG
AACCGCTGGCAACAAAGGATAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGACGACA
GCCATGCAGCACCTGTCTCACAGTTCCCGAAGGCACCAAAGCATCTCTGCTAAATTCTCTGGATGTCAAGAGTAGGTA
AGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGGGGGGGGGG

## UA.a.4 Uncultured Serratia

NNNNGNNNNNNNTNCTGATCTACGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGACT
ACGACATACTTTATGAGGTCCGCTTGCTCTCGCGAGGTCGCTTCTCTTTGTATATGCCATTGTAGCACGTGTGTAGCCC
TACTCGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCAGTTTATCACTGGCAGTCTCCTTTGAGTTCCCGG
CCGAACCGCTGGCAACAAAGGATAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGACG
ACAGCCATGCAGCACCTGTCTCAGAGTTCCCGAAGGCACCAAAGCATCTCTGCTAAGTTCTCTGGATGTCAAGAGTAG
GTAAGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCGA

## UA.a.5 Pectobacterium carotovorum

NNNNNNANNCATTTCCTGATCTACGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAAATCCGGAC
TACGACGTACTTTATGAGGTCCGCTTGCTCTCGCGAGGTCGCTTCTCTTTGTATACGCCATTGTAGCACGTGTGTAGCC
CTACTCGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCGGTTTATCACCGGCAGTCTCCTTTGAGTTCCCG
ACCGAATCGCTGGCAACAAAGGATAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGAC
GACAGCCATGCAGCACCTGTCTCACAGTTCCCGAAGGCACTAAGGTATCTCTACCGAATTCTGTGGATGTCAAGAGTA
GGTAAGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCGG

## UA.a.6 Acinetobacter beijerinckii

NNNNNNCNNNNNTTCTGATCTAGCGATTACTAGCGATTCCGACTTCATGGAGTCGAGTTGCAGACTCCAATCCGGAC
TACGATCGGCTTTTTGAGATTAGCATCCTATCGCTAGGTAGCAACCCTTTGTACCGACCATTGTAGCACGTGTGTAGCC
CTGGTCGTAAGGGCCATGATGACTTGACGTCATCCCCGCCTTCCTCCAGTTTGTCACTGGCAGTATCCTTAAAGTTCCC
GGCTTAACCCGCTGGCAAATAAGGAAAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACAAGCTG
ACAACAGCCATGCAGCACCTGTATGAAAGTTCCCAAAGGCACCAATCCATCTCTGGAAAGTTCTTAGTATGTCAAGAC
CAGGTAAGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGGGA

## UA.a.7 Pantoea sp. 57917

NNNNNNNNNNNTTCTGATCTACGATTACTAGCGATTCCGACTTCACGGAGTCGAGTTGCAGACTCCGATCCGGACTA
CGACTCACTTTATGAGGTCCGCTTGCTCTCGCGAGGTCGCTTCTCTTTGTATGCGCCATTGTAGCACGTGTGTAGCCCT
ACTCGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCAGTTTATCACTGGCAGTCTCCTTTGAGTTCCCGGC
CGAACCGCTGGCAACAAAGGATAAGGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGACGA
CAGCCATGCAGCACCTGTCTCAGAGTTCCCGAAGGCACTAAGCTATCTCTAGCGAATTCCCTGGATGTCAAGAGTAGG
TAAGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCGG

## UA.a.8 Erwinia amylovora

GNNNNNNNNNAACNTTCTGATCTACGATTACTAGCGATTCCGACTTCACGGAGTCGAGTTGCAGACTCCGATCCGGAC
TACGACGCACTTTATGAGGTCCGCTTGCTCTCGCGAGTTCGCTTCTCTTTGTATGCGCCATTGTAGCACGTGTGTAGCC
CTGGCCGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCGGTTTATCACCGGCAGTCTCCTTTGAGTTCCC
GACCGAATCGCTGGCAACAAAGGATAAGGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGA
CGACAGCCATGCAGCACCTGTCTCACGGTTCCCGAAGGCACTAAGGCATCTCTGCCGAATTCCGTGGATGTCAAGGCC
AGGTAAGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGCGN

## UA.a.9 Erwinia tasmaniensis

NNNNNNNNTTCTGATCTACGATTACTAGCGATTCCGACTTCACGGAGTCGAGTTGCAGACTCCGATCCGGACTACGA
CGCACTTTATGAGGTCCGCTGGCTCTCGCGAGTTCGCTTCTCTTTGTATGCGCCATTGTAGCACGTGTGTAGCCCTGGC
CGTAAGGGCCATGATGACTTGACGTCATCCCCACCTTCCTCCGGTTTATCACCGGCAGTCTCCTTTGAGTTCCCGACCG
AATCGCTGGCAACAAAGGATAAGGGGTTGCGCTCGTTGCGGGACTTAACCCAACATTTCACAACACGAGCTGACGACA
GCCATGCAGCACCTGTCTCACGGTTCCCGAAGGCACTAAGGCATCTCTGCCGAATTCCGTGGATGTCAAGGCCAGGTA
AGGTTCTTCGCGTTCCCCCCGTGCCCCCGCCCCGCCCGCCGCGCGCGGCGGGC

# Appendix 2

**Gut microbiome analysis using MetaPhlAn**

The full list of MetaPhlAn taxonomic assignments down to genus level: (All values given are a percentage). Taxonomy is indicated using one letter identifiers: k-Kingdom, p- Phylum, c- Class, o- order, g- genus.

K Bacteria        99.99254
K Archaea        0.00746
|p Proteobacteria        88.15086
|p Bacteroidetes        10.53125
|p Firmicutes   0.59062
|p Actinobacteria        0.27643
|p Chlamydiae  0.21079
|p Chloroflexi   0.16307
|p Acidobacteria        0.02473
|p Cyanobacteria        0.02182
|p Thermi        0.01754
k Archaea|p Euryarchaeota        0.00746
|p Chlorobi        0.00197
|p Synergistetes        0.00168
|p Verrucomicrobia        0.00098
|p Lentisphaerae        0.0008
|p Proteobacteria|c Gammaproteobacteria        82.15795
|p Bacteroidetes|c Sphingobacteria        8.56669
|p Proteobacteria|c Alphaproteobacteria        3.87046
|p Bacteroidetes|c Flavobacteria        1.88595
|p Proteobacteria|c Betaproteobacteria1.75051
|p Firmicutes|c Bacilli   0.51692
|p Proteobacteria|c Deltaproteobacteria        0.34768
|p Actinobacteria|c Actinobacteria        0.27643
|p Chlamydiae|c Chlamydiae        0.21079
|p Chloroflexi|c Thermomicrobia        0.15984
|p Bacteroidetes|c Bacteroidia  0.07827
|p Firmicutes|c Clostridia        0.07302
|p Acidobacteria|c Acidobacteria        0.02473
|p Cyanobacteria|c Cyanophyceae        0.02182
|p Proteobacteria|c Epsilonproteobacteria        0.02148
|p Thermi|c Deinococci 0.01754
k Archaea|p Euryarchaeota|c Halobacteria        0.00746
|p Chloroflexi|c Anaerolineae    0.00323
|p Proteobacteria|c Magnetococci        0.00276
|p Chlorobi|c Chlorobia 0.00197
|p Synergistetes|c Synergistia    0.00168
|p Lentisphaerae|c Lentisphaerae_uncl 0.0008
|p Firmicutes|c Negativicutes    0.00068
|p Verrucomicrobia|c Spartobacteria    0.00056
|p Verrucomicrobia|c Opitutae 0.00042
|p Bacteroidetes|c Bacteroidetes_uncl  0.00033
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales 64.56101
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales        14.25299
|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales        8.56669
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales 3.45936
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales        1.88595
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales 1.76953

229

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales	1.63877
|p Proteobacteria|c Gammaproteobacteria|o Vibrionales	0.72331
|p Firmicutes|c Bacilli|o Lactobacillales	0.51667
|p Proteobacteria|c Gammaproteobacteria|o Aeromonadales	0.33464
|p Proteobacteria|c Deltaproteobacteria|o Desulfovibrionales	0.28351
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales	0.2744
|p Actinobacteria|c Actinobacteria|o Actinomycetales	0.23297
|p Chlamydiae|c Chlamydiae|o Chlamydiales	0.21079
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales	0.17352
|p Chloroflexi|c Thermomicrobia|o Thermomicrobia_unclassified	0.15984
|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales	0.12891
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales	0.1136
|p Bacteroidetes|c Bacteroidia|o Bacteroidales	0.07827
|p Proteobacteria|c Betaproteobacteria|o Neisseriales	0.07781
|p Firmicutes|c Clostridia|o Clostridiales	0.07302
|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales	0.07194
|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales	0.06907
|p Proteobacteria|c Deltaproteobacteria|o Desulfobacterales	0.06146
|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales	0.05229
|p Actinobacteria|c Actinobacteria|o Bifidobacteriales	0.03246
|p Acidobacteria|c Acidobacteria|o Acidobacteriales	0.02473
|p Proteobacteria|c Epsilonproteobacteria|o Epsilonproteobacteria_uncl	0.02148
|p Thermi|c Deinococci|o Deinococcales	0.01754
|p Proteobacteria|c Betaproteobacteria|o Methylophilales	0.01396
|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales	0.01247
|p Cyanobacteria|c Cyanophyceae|o Oscillatoriales	0.01111
|p Actinobacteria|c Actinobacteria|o Coriobacteriales	0.01053
|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales	0.01038
|p Proteobacteria|c Gammaproteobacteria|o Cardiobacteriales	0.00871
|p Proteobacteria|c Gammaproteobacteria|o Legionellales	0.0084
k Archaea|p Euryarchaeota|c Halobacteria|o Halobacteriales	0.00746
|p Proteobacteria|c Betaproteobacteria|o Nitrosomonadales	0.00745
|p Proteobacteria|c Gammaproteobacteria|o Methylococcales	0.00734
|p Cyanobacteria|c Cyanophyceae|o Chroococcales	0.00707
|p Proteobacteria|c Gammaproteobacteria|o Acidithiobacillales	0.00398
|p Cyanobacteria|c Cyanophyceae|o Synechococcales	0.00364
|p Chloroflexi|c Anaerolineae|o Anaerolineales	0.00323
|p Proteobacteria|c Magnetococci|o Magnetococci_uncl	0.00276
|p Proteobacteria|c Deltaproteobacteria|o Myxococcales	0.00271
|p Proteobacteria|c Betaproteobacteria|o Gallionellales	0.00214
|p Chlorobi|c Chlorobia|o Chlorobiales	0.00197
|p Synergistetes|c Synergistia|o Synergistales	0.00168
|p Proteobacteria|c Gammaproteobacteria|o Gammaproteobacteria_uncl	0.0016
|p Proteobacteria|c Gammaproteobacteria|o Thiotrichales	0.00133
|p Lentisphaerae|c Lentisphaerae_uncl|o Victivallales	0.0008
|p Firmicutes|c Negativicutes|o Selenomonadales	0.00068
|p Verrucomicrobia|c Spartobacteria|o Spartobacteria_uncl	0.00056
|p Actinobacteria|c Actinobacteria|o Solirubrobacterales	0.00047
|p Verrucomicrobia|c Opitutae|o Opitutales	0.00042
|p Bacteroidetes|c Bacteroidetes_uncl|o Bacteroidetes_uncl	0.00033
|p Firmicutes|c Bacilli|o Bacillales	0.00025

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae   64.56101
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Pseudomonadaceae   10.55916
|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales|f Sphingobacteriaceae   8.56669
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Moraxellaceae   3.69383
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Brucellaceae   2.82159
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae   1.84898
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Oceanospirillaceae   1.4035
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Alcaligenaceae   0.91024
|p Proteobacteria|c Gammaproteobacteria|o Vibrionales|f Vibrionaceae   0.72331
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae   0.35062
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Halomonadaceae   0.34318
|p Proteobacteria|c Gammaproteobacteria|o Aeromonadales|f Aeromonadaceae   0.33464
|p Firmicutes|c Bacilli|o Lactobacillales|f Streptococcaceae   0.31884
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae   0.30945
|p Proteobacteria|c Deltaproteobacteria|o Desulfovibrionales|f Desulfovibrionaceae   0.28351
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae   0.25987
|p Chlamydiae|c Chlamydiae|o Chlamydiales|f Chlamydiaceae   0.21079
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Aurantimonadaceae   0.17427
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Ectothiorhodospiraceae   0.16413
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Methylobacteriaceae   0.14663
|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales|f Xanthomonadaceae   0.12891
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Rhizobiaceae   0.11685
|p Firmicutes|c Bacilli|o Lactobacillales|f Enterococcaceae   0.11364
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Micrococcaceae   0.09285
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Methylocystaceae   0.08117
|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae   0.07781
|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Bacteroidaceae   0.07123
|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae   0.07113
|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae   0.06907
|p Firmicutes|c Bacilli|o Lactobacillales|f Lactobacillaceae   0.06755
|p Proteobacteria|c Deltaproteobacteria|o Desulfobacterales|f Desulfobulbaceae   0.06146
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae   0.06144
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bradyrhizobiaceae   0.05379
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Alteromonadaceae   0.05355
|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_uncl   0.05232
|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales|f Caulobacteraceae   0.05229
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Frankiaceae   0.04603
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Brevibacteriaceae   0.03677
|p Actinobacteria|c Actinobacteria|o Bifidobacteriales|f Bifidobacteriaceae   0.03246
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Mycobacteriaceae   0.03062
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Xanthobacteraceae   0.02488
|p Acidobacteria|c Acidobacteria|o Acidobacteriales|f Acidobacteriaceae   0.02473
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriales_uncl   0.02442
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Beijerinckiaceae   0.02256
|p Proteobacteria|c Epsilonproteobacteria|o Epsilonproteobacteria_uncl|f Epsilonproteobacteria_uncl   0.02148
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Idiomarinaceae   0.02121
|p Thermi|c Deinococci|o Deinococcales|f Deinococcaceae   0.01754
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Pseudoalteromonadaceae   0.01749
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Alcanivoracaceae   0.01615

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Nocardioidaceae        0.01455
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae 0.01453
|p Proteobacteria|c Betaproteobacteria|o Methylophilales|f Methylophilaceae 0.01396
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Blattabacteriaceae        0.01256
|p Cyanobacteria|c Cyanophyceae|o Oscillatoriales|f Oscillatoriaceae    0.01111
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Phyllobacteriaceae    0.0109
|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_Family_XI_Incertae_Sedis       0.01072
|p Actinobacteria|c Actinobacteria|o Coriobacteriales|f Coriobacteriaceae        0.01053
|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae        0.01038
|p Firmicutes|c Bacilli|o Lactobacillales|f Leuconostocaceae        0.01018
|p Proteobacteria|c Gammaproteobacteria|o Cardiobacteriales|f Cardiobacteriaceae    0.00871
|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae    0.00846
|p Proteobacteria|c Gammaproteobacteria|o Legionellales|f Coxiellaceae        0.0084
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Ferrimonadaceae        0.00799
k Archaea|p Euryarchaeota|c Halobacteria|o Halobacteriales|f Halobacteriales_unclassified
        0.00746
|p Proteobacteria|c Betaproteobacteria|o Nitrosomonadales|f Nitrosomonadaceae        0.00745
|p Proteobacteria|c Gammaproteobacteria|o Methylococcales|f Methylococcaceae        0.00734
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Shewanellaceae        0.00731
|p Cyanobacteria|c Cyanophyceae|o Chroococcales|f Cyanobacteriaceae        0.00707
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiales_uncl        0.00703
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Hahellaceae    0.0067
|p Firmicutes|c Bacilli|o Lactobacillales|f Carnobacteriaceae        0.00645
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Chromatiaceae        0.0061
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bartonellaceae        0.00559
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Dermabacteraceae        0.00535
|p Firmicutes|c Clostridia|o Clostridiales|f Ruminococcaceae        0.00455
|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Rikenellaceae 0.00429
|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Erythrobacteraceae    0.00401
|p Proteobacteria|c Gammaproteobacteria|o Acidithiobacillales|f Acidithiobacillaceae    0.00398
|p Cyanobacteria|c Cyanophyceae|o Synechococcales|f Synechococcaceae        0.00364
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Psychromonadaceae    0.00358
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Halothiobacillaceae        0.00329
|p Chloroflexi|c Anaerolineae|o Anaerolineales|f Anaerolineaceae        0.00323
|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiaceae        0.0031
|p Proteobacteria|c Magnetococci|o Magnetococci_uncl|f Magnetococci_uncl  0.00276
|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Porphyromonadaceae 0.00275
|p Proteobacteria|c Deltaproteobacteria|o Myxococcales|f Myxococcaceae        0.00271
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Microbacteriaceae        0.00267
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Alteromonadales_uncl
        0.00247
|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_Family_XVII_Incertae_Sedis        0.00234
|p Proteobacteria|c Betaproteobacteria|o Gallionellales|f Gallionellaceae        0.00214
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Propionibacteriaceae    0.00207
|p Chlorobi|c Chlorobia|o Chlorobiales|f Chlorobiaceae 0.00197
|p Synergistetes|c Synergistia|o Synergistales|f Synergistaceae 0.00168
|p Proteobacteria|c Gammaproteobacteria|o Gammaproteobacteria_uncl|f
Gammaproteobacteria_uncl        0.0016
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Corynebacteriaceae        0.00134
|p Proteobacteria|c Gammaproteobacteria|o Thiotrichales|f Piscirickettsiaceae 0.00133
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Hyphomicrobiaceae    0.00113

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Hyphomonadaceae	0.00081
|p Lentisphaerae|c Lentisphaerae_uncl|o Victivallales|f Victivallaceae	0.0008
|p Firmicutes|c Negativicutes|o Selenomonadales|f Acidaminococcaceae	0.00068
|p Verrucomicrobia|c Spartobacteria|o Spartobacteria_uncl|f Spartobacteria_uncl	0.00056
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Pseudonocardiaceae	0.00049
|p Actinobacteria|c Actinobacteria|o Solirubrobacterales|f Conexibacteraceae	0.00047
|p Verrucomicrobia|c Opitutae|o Opitutales|f Opitutaceae	0.00042
|p Bacteroidetes|c Bacteroidetes_uncl|o Bacteroidetes_uncl|f Bacteroidetes_uncl	0.00033
|p Firmicutes|c Bacilli|o Bacillales|f Bacillaceae	0.00025
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Jonesiaceae	0.00022
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Enterobacter	26.85735
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Citrobacter	19.85988
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Pseudomonadaceae|g Pseudomonas	10.5419
|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales|f Sphingobacteriaceae|g Sphingobacteriaceae_unclassified	8.10302
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Escherichia	3.91216
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Moraxellaceae|g Acinetobacter	3.68675
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Brucellaceae|g Brucella	2.74665
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Pantoea	2.74419
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Yersinia	2.42967
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Klebsiella	2.03562
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Flavobacterium	1.46915
|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Oceanospirillaceae|g Marinomonas	1.40298
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Salmonella	1.09474
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Providencia	0.81916
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Serratia	0.75575
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Cronobacter	0.75043
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Erwinia	0.73635
|p Proteobacteria|c Gammaproteobacteria|o Vibrionales|f Vibrionaceae|g Photobacterium	0.64749
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Alcaligenaceae|g Bordetella	0.62897
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Shigella	0.49298
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Rahnella	0.4475

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Edwardsiella    0.35464

|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales|f Sphingobacteriaceae|g Pedobacter    0.34371

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae|g Burkholderia    0.33527

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Halomonadaceae|g Halomonadaceae_unclassified   0.33122

|p Proteobacteria|c Gammaproteobacteria|o Aeromonadales|f Aeromonadaceae|g Aeromonas    0.3196

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Pectobacterium0.31891

|p Firmicutes|c Bacilli|o Lactobacillales|f Streptococcaceae|g Lactococcus        0.31884

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Dickeya    0.30927

|p Proteobacteria|c Deltaproteobacteria|o Desulfovibrionales|f Desulfovibrionaceae|g Desulfovibrio    0.28351

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Alcaligenaceae|g Achromobacter    0.28127

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Cellulophaga0.27103

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae|g Acetobacteraceae_unclassified  0.25127

|p Chlamydiae|c Chlamydiae|o Chlamydiales|f Chlamydiaceae|g Chlamydiaceae_unclassified    0.21079

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Comamonas    0.18609

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Aurantimonadaceae|g Aurantimonadaceae_unclassified        0.17427

|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Ectothiorhodospiraceae|g Thioalkalivibrio 0.15468

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Xenorhabdus    0.14848

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Methylobacteriaceae|g Methylobacterium        0.14663

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Sodalis    0.13997

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Proteus    0.13364

|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales|f Sphingobacteriaceae|g Sphingobacterium        0.11734

|p Firmicutes|c Bacilli|o Lactobacillales|f Enterococcaceae|g Enterococcus        0.11364

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Photorhabdus   0.10979

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Rhizobiaceae|g Agrobacterium    0.10508

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Micrococcaceae|g Arthrobacter    0.08816

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Methylocystaceae|g Methylocystaceae_unclassified  0.08038

|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales|f Xanthomonadaceae|g Stenotrophomonas        0.07812

234

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Brucellaceae|g Ochrobactrum 0.07493
|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Bacteroidaceae|g Bacteroides 0.07123
|p Firmicutes|c Bacilli|o Lactobacillales|f Lactobacillaceae|g Lactobacillus        0.06755
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Chryseobacterium
        0.06246
|p Proteobacteria|c Deltaproteobacteria|o Desulfobacterales|f Desulfobulbaceae|g
Desulfobulbaceae_unclassified  0.06027
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Acidovorax
        0.05603
|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_uncl|g Blautia   0.05232
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bradyrhizobiaceae|g
Rhodopseudomonas     0.05019
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Alteromonadaceae|g
Marinobacter    0.04931
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g
Candidatus_Regiella      0.04747
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Frankiaceae|g Frankia  0.04603
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Variovorax
        0.03847
|p Proteobacteria|c Gammaproteobacteria|o Vibrionales|f Vibrionaceae|g Grimontia    0.03778
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Brevibacteriaceae|g Brevibacterium
        0.03677
|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Ruegeria
        0.03401
|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales|f Xanthomonadaceae|g
Xanthomonas    0.03337
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae|g
Candidatus_Zinderia      0.03317
|p Actinobacteria|c Actinobacteria|o Bifidobacteriales|f Bifidobacteriaceae|g Bifidobacterium
        0.03246
|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Aggregatibacter
        0.03216
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Mycobacteriaceae|g Mycobacterium
        0.03062
|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Kingella        0.02758
|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales|f Caulobacteraceae|g Caulobacter
        0.02555
|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales|f Caulobacteraceae|g
Brevundimonas 0.02477
|p Acidobacteria|c Acidobacteria|o Acidobacteriales|f Acidobacteriaceae|g Granulicella 0.02473
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriales_uncl|g Candidatus_Sulcia
        0.02442
|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g
Candidatus_Carsonella 0.02313
|p Proteobacteria|c Gammaproteobacteria|o Vibrionales|f Vibrionaceae|g Vibrio        0.023
|p Proteobacteria|c Epsilonproteobacteria|o Epsilonproteobacteria_uncl|f
Epsilonproteobacteria_uncl|g Epsilonproteobacteria_uncl_unclassified  0.02148
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Idiomarinaceae|g Idiomarina
        0.02121
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Xanthobacteraceae|g
Xanthobacteraceae_unclassified        0.0194

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Baumannia     0.0193

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Beijerinckiaceae|g Beijerinckiaceae_unclassified    0.01863

|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Neisseriaceae_unclassified    0.01818

|p Thermi|c Deinococci|o Deinococcales|f Deinococcaceae|g Deinococcus    0.01754

|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Pseudoalteromonadaceae|g Pseudoalteromonas    0.01749

|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Neisseria    0.01513

|p Proteobacteria|c Gammaproteobacteria|o Vibrionales|f Vibrionaceae|g Aliivibrio    0.01504

|p Proteobacteria|c Gammaproteobacteria|o Aeromonadales|f Aeromonadaceae|g Tolumonas    0.01504

|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Pseudomonadaceae|g Azotobacter    0.01496

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Nocardioidaceae|g Aeromicrobium    0.01455

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae|g Oxalobacter    0.01449

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Actinobacillus    0.01431

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Riemerella    0.01385

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Delftia    0.0132

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Blattabacteriaceae|g Blattabacterium    0.01256

|p Cyanobacteria|c Cyanophyceae|o Oscillatoriales|f Oscillatoriaceae|g Oscillatoriaceae_unclassified    0.01111

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Alcanivoracaceae|g Alcanivoracaceae_unclassified    0.01092

|p Actinobacteria|c Actinobacteria|o Coriobacteriales|f Coriobacteriaceae|g Slackia    0.01053

|p Firmicutes|c Bacilli|o Lactobacillales|f Leuconostocaceae|g Leuconostoc    0.01018

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Haemophilus    0.0101

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Buchnera    0.00977

|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales|f Xanthomonadaceae|g Pseudoxanthomonas    0.00954

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Rhizobiaceae|g Rhizobium    0.00953

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae|g Janthinobacterium    0.00923

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Paracoccus    0.00915

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Candidatus_Hamiltonella    0.00898

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae|g Lautropia    0.00873

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Ketogulonicigenium    0.00839

|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Laribacter    0.0082

|p Proteobacteria|c Gammaproteobacteria|o Legionellales|f Coxiellaceae|g Rickettsiella 0.00816

|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Ferrimonadaceae|g Ferrimonas 0.00799

|p Proteobacteria|c Gammaproteobacteria|o Xanthomonadales|f Xanthomonadaceae|g Xylella 0.00788

|p Proteobacteria|c Betaproteobacteria|o Nitrosomonadales|f Nitrosomonadaceae|g Nitrosomonas 0.00745

|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Shewanellaceae|g Shewanella 0.00731

|p Cyanobacteria|c Cyanophyceae|o Chroococcales|f Cyanobacteriaceae|g Cyanobacteriaceae_unclassified 0.00707

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiales_uncl|g Thiomonas 0.00703

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Polaribacter 0.007

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Hahellaceae|g Hahella 0.0067

|p Proteobacteria|c Betaproteobacteria|o Methylophilales|f Methylophilaceae|g Methylobacillus 0.00646

|p Firmicutes|c Bacilli|o Lactobacillales|f Carnobacteriaceae|g Granulicatella 0.00645

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Phyllobacteriaceae|g Mesorhizobium 0.00628

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Zunongwangia 0.00627

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Halomonadaceae|g Chromohalobacter 0.00605

|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_Family_XI_Incertae_Sedis|g Clostridiales_Family_XI_Incertae_Sedis_unclassified 0.00603

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Halomonadaceae|g Halomonas 0.00591

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Basfia 0.00587

|p Proteobacteria|c Gammaproteobacteria|o Cardiobacteriales|f Cardiobacteriaceae|g Cardiobacterium 0.00585

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Labrenzia 0.00579

|p Proteobacteria|c Betaproteobacteria|o Methylophilales|f Methylophilaceae|g Methylovorus 0.00572

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bartonellaceae|g Bartonella 0.00559

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Capnocytophaga 0.00556

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Oceanicola 0.00535

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Dermabacteraceae|g Brachybacterium 0.00535

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Xanthobacteraceae|g Starkeya 0.00523

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Alcanivoracaceae|g Alcanivorax 0.00523

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Alicycliphilus 0.00513

|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Lutiella 0.00501

|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_Family_XI_Incertae_Sedis|g Peptoniphilus    0.00469
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Chromatiaceae|g Allochromatium    0.00455
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Verminephrobacter    0.00454
|p Proteobacteria|c Gammaproteobacteria|o Methylococcales|f Methylococcaceae|g Methylobacter  0.00452
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae|g Magnetospirillum    0.00443
|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Rikenellaceae|g Alistipes    0.00429
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Alteromonadaceae|g Alteromonas    0.00424
|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Weeksella    0.00424
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae|g Gluconacetobacter    0.00419
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Ectothiorhodospiraceae|g Halorhodospira 0.00416
|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Phyllobacteriaceae|g Hoeflea 0.00415
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae|g Rhodocista    0.00412
|p Proteobacteria|c Gammaproteobacteria|o Acidithiobacillales|f Acidithiobacillaceae|g Acidithiobacillus    0.00398
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae|g Cupriavidus    0.00389
|p Cyanobacteria|c Cyanophyceae|o Synechococcales|f Synechococcaceae|g Cyanobium    0.00364
|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Psychromonadaceae|g Psychromonas  0.00358
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Leptothrix    0.00357
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Ectothiorhodospiraceae|g Alkalilimnicola  0.00349
|p Firmicutes|c Clostridia|o Clostridiales|f Ruminococcaceae|g Ruminococcus    0.00346
|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Chromobacterium    0.00345
|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae|g Azospirillum    0.00342
|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Halothiobacillaceae|g Halothiobacillus0.00329
|p Chloroflexi|c Anaerolineae|o Anaerolineales|f Anaerolineaceae|g Anaerolinea    0.00323
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Moraxellaceae|g Psychrobacter  0.00323
|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Moraxellaceae|g Enhydrobacter  0.00309
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Micrococcaceae|g Rothia    0.00298
|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae|g Dechloromonas    0.00289
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae|g Herbaspirillum    0.00289

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Beijerinckiaceae|g Methylocella	0.00286

|p Proteobacteria|c Gammaproteobacteria|o Cardiobacteriales|f Cardiobacteriaceae|g Dichelobacter	0.00286

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Kordia	0.00285

|p Proteobacteria|c Gammaproteobacteria|o Methylococcales|f Methylococcaceae|g Methylococcus	0.00282

|p Proteobacteria|c Magnetococci|o Magnetococci_uncl|f Magnetococci_uncl|g Magnetococcus	0.00276

|p Bacteroidetes|c Bacteroidia|o Bacteroidales|f Porphyromonadaceae|g Parabacteroides	0.00275

|p Proteobacteria|c Deltaproteobacteria|o Myxococcales|f Myxococcaceae|g Anaeromyxobacter	0.00271

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Mannheimia	0.0027

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Microbacteriaceae|g Clavibacter	0.00267

|p Bacteroidetes|c Sphingobacteria|o Sphingobacteriales|f Sphingobacteriaceae|g Mucilaginibacter	0.00264

|p Proteobacteria|c Gammaproteobacteria|o Alteromonadales|f Alteromonadales_uncl|g Teredinibacter	0.00247

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae|g Novosphingobium	0.00244

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Maribacter	0.00241

|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiales_Family_XVII_Incertae_Sedis|g Thermaerobacter	0.00234

|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae|g Azoarcus	0.00233

|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Pseudomonadaceae|g Cellvibrio	0.0023

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Roseovarius	0.00224

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Rhizobiaceae|g Ensifer	0.00224

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae|g Sphingobium	0.00224

|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiaceae|g Pseudoflavonifractor	0.00224

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Methylibium	0.0022

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Ahrensia	0.00219

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae|g Sphingomonas	0.00218

|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae|g Candidatus_Accumulibacter	0.00215

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Erythrobacteraceae|g Citromicrobium	0.00215

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Histophilus	0.00211

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae|g Acetobacter	0.00209

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Propionibacteriaceae|g Propionibacterium	0.00207

239

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae|g Rhodospirillum 0.00203

|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae|g Thauera 0.00202

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae|g Polynucleobacter 0.00202

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bradyrhizobiaceae|g Oligotropha 0.00194

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Ulvibacter 0.00192

|p Proteobacteria|c Gammaproteobacteria|o Enterobacteriales|f Enterobacteriaceae|g Candidatus_Blochmannia 0.00189

|p Proteobacteria|c Gammaproteobacteria|o Pasteurellales|f Pasteurellaceae|g Pasteurella 0.00183

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Rhodobacter 0.00183

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae|g Granulibacter 0.00182

|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Ectothiorhodospiraceae|g Nitrococcus 0.0018

|p Proteobacteria|c Betaproteobacteria|o Methylophilales|f Methylophilaceae|g Methylotenera 0.00178

|p Chlorobi|c Chlorobia|o Chlorobiales|f Chlorobiaceae|g Chlorobium 0.00173

|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales|f Caulobacteraceae|g Asticcacaulis 0.00173

|p Synergistetes|c Synergistia|o Synergistales|f Synergistaceae|g Pyramidobacter 0.00168

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Oxalobacteraceae|g Herminiimonas 0.00167

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Bradyrhizobiaceae|g Nitrobacter 0.00166

|p Proteobacteria|c Gammaproteobacteria|o Gammaproteobacteria_uncl|f Gammaproteobacteria_uncl|g Congregibacter 0.0016

|p Proteobacteria|c Gammaproteobacteria|o Chromatiales|f Chromatiaceae|g Nitrosococcus 0.00155

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Micrococcaceae|g Kocuria 0.00145

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Gramella 0.0014

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae|g Zymomonas 0.00135

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Corynebacteriaceae|g Corynebacterium 0.00134

|p Proteobacteria|c Gammaproteobacteria|o Thiotrichales|f Piscirickettsiaceae|g Methylophaga 0.00133

|p Proteobacteria|c Deltaproteobacteria|o Desulfobacterales|f Desulfobulbaceae|g Desulfurivibrio 0.00119

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Erythrobacteraceae|g Erythrobacter 0.00119

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Hyphomicrobiaceae|g Hyphomicrobium 0.00113

|p Proteobacteria|c Betaproteobacteria|o Gallionellales|f Gallionellaceae|g Gallionella 0.0011

|p Firmicutes|c Clostridia|o Clostridiales|f Ruminococcaceae|g Faecalibacterium 0.00109

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Beijerinckiaceae|g Beijerinckia 0.00108

|p Proteobacteria|c Betaproteobacteria|o Gallionellales|f Gallionellaceae|g Sideroxydans 0.00104

|p Proteobacteria|c Betaproteobacteria|o Rhodocyclales|f Rhodocyclaceae|g Aromatoleum 0.00098

|p Firmicutes|c Clostridia|o Clostridiales|f Clostridiaceae|g Clostridium 0.00086

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Roseibium 0.00086

|p Bacteroidetes|c Flavobacteria|o Flavobacteriales|f Flavobacteriaceae|g Psychroflexus 0.00082

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Hyphomonadaceae|g Hyphomonas 0.00081

|p Lentisphaerae|c Lentisphaerae_uncl|o Victivallales|f Victivallaceae|g Victivallis 0.0008

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Methylocystaceae|g Methylosinus 0.00079

|p Proteobacteria|c Gammaproteobacteria|o Pseudomonadales|f Moraxellaceae|g Moraxella 0.00076

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Pelagibaca 0.00075

|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Burkholderiaceae|g Ralstonia 0.00072

|p Firmicutes|c Negativicutes|o Selenomonadales|f Acidaminococcaceae|g Phascolarctobacterium 0.00068

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Erythrobacteraceae|g Erythrobacteraceae_unclassified 0.00067

|p Verrucomicrobia|c Spartobacteria|o Spartobacteria_uncl|f Spartobacteria_uncl|g Chthoniobacter 0.00056

|p Proteobacteria|c Alphaproteobacteria|o Rhodobacterales|f Rhodobacteraceae|g Sulfitobacter 0.00056

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Rhodospirillaceae|g Nisaea 0.00052

|p Proteobacteria|c Gammaproteobacteria|o Oceanospirillales|f Oceanospirillaceae|g Neptuniibacter 0.00051

|p Proteobacteria|c Alphaproteobacteria|o Rhodospirillales|f Acetobacteraceae|g Gluconobacter 0.0005

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Pseudonocardiaceae|g Amycolatopsis 0.00049

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Phyllobacteriaceae|g Parvibaculum 0.00047

|p Actinobacteria|c Actinobacteria|o Solirubrobacterales|f Conexibacteraceae|g Conexibacter 0.00047

|p Verrucomicrobia|c Opitutae|o Opitutales|f Opitutaceae|g Opitutus 0.00042

|p Bacteroidetes|c Bacteroidetes_uncl|o Bacteroidetes_uncl|f Bacteroidetes_uncl|g Candidatus_Amoebophilus 0.00033

|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Micrococcaceae|g Micrococcus 0.00026

|p Proteobacteria|c Alphaproteobacteria|o Caulobacterales|f Caulobacteraceae|g Phenylobacterium 0.00026

|p Proteobacteria|c Alphaproteobacteria|o Sphingomonadales|f Sphingomonadaceae|g Sphingopyxis 0.00025

|p Proteobacteria|c Gammaproteobacteria|o Legionellales|f Coxiellaceae|g Coxiella 0.00025

|p Proteobacteria|c Alphaproteobacteria|o Rhizobiales|f Xanthobacteraceae|g Azorhizobium 0.00025

|p Firmicutes|c Bacilli|o Bacillales|f Bacillaceae|g Bacillus        0.00025
|p Chlorobi|c Chlorobia|o Chlorobiales|f Chlorobiaceae|g Chlorobaculum        0.00024
|p Proteobacteria|c Betaproteobacteria|o Neisseriales|f Neisseriaceae|g Simonsiella     0.00024
|p Proteobacteria|c Betaproteobacteria|o Burkholderiales|f Comamonadaceae|g Albidiferax
        0.00023
|p Actinobacteria|c Actinobacteria|o Actinomycetales|f Jonesiaceae|g Jonesia   0.00022

# Appendix 3

**Protein and nucleotide sequences for amplified predicted CAZyme genes**

**>gene_id_9459_amino_acid**

MRYRFPENFWWGSACSALQTEGDSLNGGKSQTTWDVWFDRQPGRFHQGVGPADTSGFYQHWKQDIALLK
QLQHNSFRTSLSWSRLIPDGTGEVNPEAVDFYNNVIDELLAQGITPFITLFHFDMPMVMQEKGGWENREVVE
AFGRYAQTCFTLFGNRVKHWFTFNEPIVPVEGGYLYDFHYPNVVDFKRAATVAYHTVLAHSTAVRAYRAGNYD
GEIGVVLNLTPSYPRSQNPADVKAAHHADLLFNRSFLDPVLKGEYPADLVELLKQYDQLPACQPGDSQLIAEGKI
DLLGINYYQPRRVKCRDSAVNPDAPFMPESLFDYYEMPGRKMNPYRGWEIYEPGIYDIITNLRDNYGNPRCFIS
ENGMGVENEQRFIQDDQINDNYRIEFVSEHLKWLHKGINEGCHCLGYHMWTFIDNWSWLNGYKNRYGFVQ
LDLATQKRTVKKSGEWFAKTAANNGFD

**>gene_id_9459_nucleotide**

ATGAGATACCGTTTTCCTGAAAACTTCTGGTGGGGCAGTGCCTGCTCAGCGCTGCAAACCGAAGGGGATA
GCCTGAACGGCGGTAAAAGCCAGACCACATGGGACGTGTGGTTCGACCGTCAGCCAGGTCGTTTCCATCA
GGGGGGTTGGCCCGGCAGACACCTCCGGTTTTTATCAACACTGGAAACAGGACATCGCACTGCTGAAACAG
CTACAGCACAACAGCTTTCGTACATCCTTGAGCTGGTCGCGTCTGATCCCGGACGGCACCGGCGAGGTCAA
TCCCGAAGCGGTCGATTTCTATAACAACGTCATCGACGAGCTGCTGGCGCAGGGTATCACGCCCTTTATCA
CCCTGTTTCACTTCGATATGCCAATGGTCATGCAGGAGAAAGGCGGCTGGGAAAATCGTGAAGTTGTGGA
AGCCTTTGGCCGCTACGCCCAGACGTGTTTTACCTTGTTCGGCAATCGGGTGAAGCACTGGTTCACCTTCAA
CGAACCGATTGTCCCGGTAGAAGGCGGCTATCTGTATGACTTCCACTACCCGAACGTGGTGGATTTCAAAC
GTGCGGCAACCGTGGCTTACCACACCGTACTGGCGCATTCGACGGCGGTCCGTGCGTATCGCGCTGGCAA
CTATGACGGTGAAATCGGCGTGGTGCTGAATTTAACGCCGTCTTATCCGCGCTCGCAAAATCCGGCGGATG
TGAAAGCCGCGCACCATGCGGATCTGCTGTTTAACCGCAGCTTCCTCGATCCGGTATTAAAAGGCGAATAT
CCGGCGGACCTGGTGGAACTGTTGAAACAGTACGATCAACTACCGGCCTGCCAGCCCGGCGACAGCCAGC
TCATCGCCGAGGGTAAAATCGACCTGTTGGGCATTAACTACTACCAACCACGCCGGGTGAAATGCCGCGAT
AGCGCCGTCAATCCTGACGCGCCGTTTATGCCAGAGTCGCTGTTTGATTATTACGAAATGCCGGGACGTAA
AATGAATCCGTACCGTGGCTGGGAAATCTATGAGCCGGGCATTTACGATATTATTACGAATCTGCGTGATA
ATTATGGCAACCCGCGCTGTTTTATTTCTGAAAACGGCATGGGTGTAGAAAACGAGCAGCGTTTTATTCAG
GACGATCAGATTAACGACAATTACCGTATTGAGTTTGTCTCGGAACATCTTAAATGGCTGCATAAAGGCAT
TAACGAGGGTTGTCATTGCCTTGGTTACCATATGTGGACCTTTATTGATAATTGGTCATGGCTTAACGGATA
TAAAAACCGCTACGGTTTTGTCCAGTTAGATTTAGCCACGCAAAAACGTACAGTGAAAAAGAGCGGCGAG
TGGTTTGCGAAAACTGCCGCCAATAATGGATTCGATTAA

**>gene_id_13418_amino_acid**

MSLRALVAVMVTTTMMVLPRAWADTAWESYKSRFMMADGRIVDTGNGSVSHTEGQGFAMLLAVAKDDR
PAFDKLWQWTDKTLRNKDNGLFYWRYNPVAPDPIADKNDATDGDSLIAWALLRAQQQWGDKSYGIASDAIT
ASLLKSTVITFAGHQVMLPGAKGFNRNDHVNLNPSYFIFPAWQAFAARTHLTAWRKLQSDGKALLGKMAWG
KAQLPSDWVALRADGRMEPAKEWPPRMSYDAIRIPLYLSWADPQSALLTPWKSWFQSYPRLQTPAWVNVN
TNDVAPWFMTGGLLAVRDLTTGEAQDDPQLSAQDDYYSASLKMLVWLAKNDRRQAATGV

**>gene_id_13418_nucleotide**

ATGTCCTTGCGTGCTTTAGTCGCAGTCATGGTTACGACGACGATGATGGTGCTTCCCCGCGCGTGGGCAGA
TACCGCCTGGGAAAGCTATAAATCCCGTTTTATGATGGCCGATGGGCGAATCGTTGATACCGGCAACGGC
AGCGTTTCCCATACCGAAGGGCAGGGGTTCGCCATGCTGCTGGCGGTCGCCAAAGACGATCGCCCTGCCT
TTGATAAGCTGTGGCAGTGGACGGATAAAACGCTGCGTAATAAAGACAACGGGCTATTTTACTGGCGCTA
CAACCCCGTCGCGCCGGACCCTATTGCGGATAAAAACGATGCGACCGACGGCGATTCCCTGATCGCGTGG
GCGCTGCTACGCGCCCAACAGCAGTGGGGTGATAAATCCTACGGCATCGCTTCAGATGCCATTACCGCGTC
GCTGCTGAAGTCTACGGTCATCACCTTTGCCGGTCATCAGGTGATGCTGCCCGGTGCGAAGGGCTTCAACC
GCAATGATCATGTCAACCTCAATCCCTCCTACTTTATTTTCCCCGCATGGCAGGCCTTTGCCGCGCGTACGC
ACCTGACGGCGTGGCGTAAGCTGCAGAGCGACGGGAAGGCGCTGCTGGGTAAAATGGCGTGGGGTAAG
GCGCAGCTGCCCAGCGATTGGGTGGCGCTGAGAGCGGACGGCAGAATGGAGCCGGCAAAGGAGTGGCC
GCCCCGGATGAGCTACGACGCGATTCGCATCCCGCTTTATCTCTCCTGGGCCGATCCGCAAAGCGCCCTGC
TGACGCCGTGGAAAAGCTGGTTTCAGAGCTATCCGCGGCTGCAGACTCCGGCGTGGGTCAACGTCAATAC
CAACGACGTGGCCCCGTGGTTTATGACCGGCGGCCTGCTGGCCGTTCGCGACCTGACCACCGGAGAAGCA
CAGGACGATCCGCAGCTTAGCGCGCAGGATGACTATTACTCTGCCAGCCTGAAGATGCTGGTCTGGCTGG
CGAAGAACGACCGCCGCCAGGCAGCTACCGGTGTCTGA

**>gene_id_77908_amino_acid**

LWTGMMHDRDPQKARLLARFKPMATLTTKNGVPPEKVDVTSGKPTGDGPVGFSASLLPFLQDRDAQAVQR
QRVADHFPGNDAYYSYVLTLFGQGWDQHRFRFTAKGELHPDWGQECASSH

**>gene_id_77908_nucleotide**

CTCTGGACGGGCATGATGCACGACCGCGATCCGCAAAAAGCCCGACTGCTGGCACGTTTTAAACCGATGG
CGACGCTCACAACAAAAAATGGCGTCCCGCCGGAGAAGGTCGATGTCACAAGCGGTAAACCCACGGGCG
ATGGCCCGGTAGGCTTTTCCGCCTCGCTGCTGCCGTTTTTACAGGACCGCGATGCACAAGCGGTGCAACGC
CAGCGCGTCGCCGACCATTTCCCCGGCAATGATGCCTATTACAGCTACGTGCTGACCCTGTTCGGACAAGG
ATGGGATCAGCACCGTTTTCGCTTTACCGCAAAGGGTGAATTACACCCTGACTGGGGCCAGGAATGCGCA
AGTTCTCATTAA

**>gene_id_71437_amino_acid**

MMRPAGWVALGIALFAGAAQAQTCDAQWPLWQNYAKRFVQDDGRVLNSSLNPSESNSEGQSYAMFFALV
GNDRARFDKLWTWTKANMAGNDISRTLPGWLWGKTQSGEWGLIDANSASDADLWVAYALLEAARVWNV
PQYRADAQLVLANVEKTLIVRVPGLGKMLLPGPVGYSYPDGLWRFNPSYQVLAQLRRFHKERPNGGWNEVAE
SNAKMLADPKSNPHGIAANWVGYRATGANTGVFVVDPYSDDLGSYDAIRTYLWAGMTAKGDPLAAPMLKAL
GGFSRATAASPNGLPPEKIHVLTGVAEKNNGYSPLGFSASALVFFQARGETALAQLQKHKLDDVLGKALVASAA
DGDQPVYYDYMLSLFSQGFTDQKYRFEQDGTVKLFWEGACAVTR

**>gene_id_71437_nucleotide**

ATGATGCGTCCAGCCGGTTGGGTTGCGCTGGGCATCGCGCTCTTCGCCGGCGCAGCCCAGGCGCAGACGT
GCGACGCGCAGTGGCCACTGTGGCAGAACTACGCGAAGCGGTTTGTGCAGGATGACGGGCGGGTGCTGA
ATTCGTCCCTGAACCCCAGCGAAAGCAATTCTGAAGGACAGTCCTACGCGATGTTTTTTGCGCTGGTGGGC
AACGACCGCGCGCGGTTTGACAAGCTCTGGACCTGGACCAAGGCCAATATGGCCGGCAACGACATCAGCC
GCACCCTGCCGGGGTTGGTTGTGGGGCAAGACCCAGAGCGGCGAATGGGGCCTGATCGACGCCAATTCCG
CCAGCGATGCTGACCTGTGGGTCGCCTACGCGCTGCTGGAAGCGGCACGTGTATGGAATGTGCCGCAGTA
CCGCGCCGATGCGCAGTTGGTGTTGGCCAATGTCGAAAAGACCTTGATCGTACGCGTGCCGGGCCTGGGC
AAGATGCTGTTGCCGGGGCCGGTGGGTTACAGCTACCCAGACGGTTTGTGGCGTTTCAACCCCAGCTACC
AAGTGCTGGCGCAACTGCGACGCTTCCACAAAGAACGCCCCAATGGCGGCTGGAATGAGGTGGCTGAGA
GCAACGCCAAAATGCTCGCCGACCCCAAGAGCAACCCCCATGGCATCGCCGCCAACTGGGTGGGTTACCG
CGCCACGGGTGCAAACACCGGGGTGTTTGTGGTCGATCCGTATTCCGATGACCTGGGCAGCTACGACGCC
ATCCGCACTTACCTGTGGGCTGGAATGACCGCCAAAGGGGACCCGCTGGCGGCGCCGATGCTCAAGGCGT
TGGGTGGTTTTTCGCGTGCTACGGCGGCGTCTCCCAATGGCTTGCCACCGGAGAAGATTCACGTGCTCACC
GGTGTGGCCGAAAAGAACAACGGCTATTCGCCGCTGGGGTTCTCGGCATCTGCTCTGGTGTTCTTCCAGGC
ACGCGGCGAAACCGCCCTGGCTCAACTGCAAAAACACAAGCTGGATGATGTGCTGGGCAAGGCTCTGGTC
GCATCAGCGGCCGACGGCGATCAGCCGGTGTATTACGACTACATGCTCAGCCTGTTCAGCCAAGGCTTTAC
CGATCAAAAGTACCGCTTCGAACAGGACGGTACGGTCAAATTATTCTGGGAGGGCGCATGCGCCGTCACA
CGCTAG

**>gene_id_8282_amino_acid**

MSLIQNPVLPGFNPDPSIIRVEDTYYIANSTFEWFPGVRLHESKDLQNWTLLPSPLSTTALLDMKGNPSSGGIW
APALSYADGKFWLVYTDVKITEGAFKDMTNYLTTATDIRGPWTDPIKLNGVGFDASLFHDEDGRKYLVQQTW
DHREYHHPFDGITLTELDTRTLKLKPETARTIYRGTAVALVEGPHLYKLNGYYYLFAAQGGTVFTHQEVVARSRTL
EANSFETQPGEVFLTNVDTPDSYIQKQGHGALVSTPSGEWYYASLCARPWNRAGESAYDPRGWSTLGRETSIQ
KVYWDEDGWPRIAGGHGGKTFVEGPADAIYTESAKDHSQHDDFKTATLDINWNTLRVPFTEKMGTTGDGRL
TLTGQGSLANTHNLSLIARRWQAFYFDAQVKVKFNPFNYQQMAGLTNYYNDRHWSFVFITWNEINGSVIEIGE
NNRGKYTSYLKDNAIKIPEGTEFVWFRTKVRKETYTYEYSFDSITFTEIPVKLDAAILSDDYVLQSYGGFFTGAFVG
LAAVDYSGYDASAEFYNFVYQELGDKKTGDNAWSWDASESRFD


**>gene_id_8282_nucleotide**

ATGTCACTTATTCAGAACCCTGTACTCCCTGGTTTTAACCCAGACCCAAGCATTATCCGTGTAGAAGATACC
TACTATATCGCAAACTCAACGTTTGAGTGGTTCCCTGGCGTTCGCTTGCATGAATCAAAAGACCTGCAAAAC
TGGACTCTTCTGCCCAGTCCTCTGTCCACAACAGCGCTTTTAGATATGAAAGGCAACCCCTCTTCTGGCGGG
ATTTGGGCTCCGGCGCTCTCATACGCCGATGGCAAATTCTGGTTGGTTTATACAGATGTGAAAATCACAGA
AGGCGCTTTTAAGGACATGACAAATTATCTGACCACCGCAACAGATATACGTGGTCCGTGGACCGATCCCA
TCAAACTAAATGGCGTCGGCTTTGATGCGTCACTGTTCCATGATGAAGACGGGCGTAAATATTTGGTTCAA
CAAACCTGGGACCATCGCGAATACCATCATCCTTTTGATGGTATCACCTTAACGGAGCTGGATACCCGGAC
CTTAAAATTAAAACCAGAGACAGCGCGTACGATTATCGCGGCACTGCCGTTGCACTTGTTGAGGGGCCAC
ATCTCTACAAGCTGAACGGGTACTACTATCTTTTCGCCGCTCAGGGCGGGACTGTATTTACCCACCAGGAG
GTGGTTGCGCGATCCAGAACCTTAGAGGCCAACAGCTTTGAAACACAACGGGAGAGGTATTCTTAACTA
ACGTCGATACGCCAGACAGCTATATCCAGAAGCAGGGGCATGGGGCGTTGGTGTCCACCCCCAGCGGTGA
GTGGTATTATGCCTCGCTCTGCGCACGTCCGTGGAATCGTGCAGGTGAGTCAGCCTACGATCCTCGTGGTT
GGTCAACCCTTGGCCGGGAAACGTCTATTCAGAAAGTGTACTGGGATGAAGATGGCTGGCCACGTATTGC
GGGAGGCCATGGTGGAAAAACCTTTGTCGAGGGGCCGGCCGATGCCATTTATACCGAAAGCGCAAAAGA
CCATAGCCAGCACGATGATTTTAAAACGGCGACACTGGATATTAACTGGAATACGCTTCGTGTCCCCTTTAC
AGAAAAAATGGGTACCACGGGCGATGGAAGGCTGACGTTAACGGGACAGGGTTCTTTAGCGAATACCCA
TAACCTTTCGCTGATTGCCCGACGCTGGCAAGCCTTTTATTTTGATGCTCAGGTTAAAGTCAAATTTAATCC
ATTTAATTACCAACAAATGGCCGGGTTGACGAATTATTATAATGACCGCCACTGGAGTTTCGTTTTCATTAC
CTGGAATGAAATCAACGGCTCTGTCATCGAAATAGGTGAGAATAACCGTGGAAAATATACCTCTTATTTGA
AAGATAACGCCATCAAGATCCCAGAGGGCACAGAATTTGTCTGGTTCCGCACGAAAGTTCGGAAGGAAAC
TTATACGTATGAATACAGCTTCGATAGCATCACTTTCACAGAGATCCCTGTTAAATTAGATGCCGCTATTCTT
TCTGATGATTATGTCCTGCAAAGCTACGGCGGCTTCTTTACCGGAGCGTTCGTTGGCCTGGCGGCAGTCGA
CTACTCTGGTTACGATGCCAGCGCAGAGTTTTATAACTTTGTTTATCAGGAGCTTGGCGATAAGAAAACTG
GCGATAATGCCTGGAGCTGGGACGCGAGCGAATCACGTTTTGATTAA

**>gene_id_3165_amino_acid**

MISKEAIKRGYNRGNYTVGAHTRPAWATDIKSSGGAKGMQIDPVAMSEATFDRLRVIADEVLTSESDVITWTR
DWWAGSMIAETQGAPATSKGAIVRVSTVEQIQDVMRLANALAIPVTVSAGRSNVTGAALPLRGGIVLDVCEL
NKLIDFDPQSQIVDVEAGMFGDIFEQTIQKEYGMTMGHWPSSFGISTVGGWVACRGAGQLSTRYGKIEDMVF
GMDVVLADGRLITVGGYARSATGPDLQQMFIGSEGTLGVIVRVRLKLHRLPDYGRAIAWGFSSFASGLEACREI
LQHGANPAALRLYDNLESGVQFGLPDTNVLLIADEGEPEIVDAVLAISERVCQRSGQQLDGEAIFERWLDTRYLT
GKSAEGFKKSPGLVADTLEMVGRWRDLSDVYDDVVAAINAVPGTLAGSAHQSHAYIDGACLYFSLRGDVAIEQ
RAAWYRAAWDAANAVLIKHNTSLSHHHGVGLLRSPYMQASLGESLSVLADIKYALDPKNILNPGKLGLSVDLPS
HQAGQ

**>gene_id_3165_nucleotide**

ATGATCAGTAAAGAAGCGATTAAACGCGGTTACAACCGTGGAAATTATACGGTAGGTGCACACACGCGCC
CGGCATGGGCGACGGATATCAAAAGCAGCGGTGGCGCGAAAGGGATGCAGATTGATCCGGTGGCGATG
AGCGAGGCAACGTTTGACCGGCTGCGGGTCATTGCTGATGAGGTGCTGACGTCGGAGTCTGATGTGATTA
CCTGGACCCGCGACTGGTGGGCCGGATCGATGATTGCGGAAACCCAGGGCGCACCGGCCACGTCAAAAG
GGGCGATTGTACGTGTCTCTACGGTAGAACAAATTCAGGATGTGATGCGTCTGGCGAATGCGTTGGCAAT
CCCGGTCACGGTTTCCGCCGGGCGCAGTAACGTCACTGGCGCCGCATTGCCGTTGCGGGGGGGCATTGTG
CTGGATGTTTGCGAACTCAACAAACTTATCGACTTCGATCCACAAAGTCAGATTGTTGACGTTGAAGCCGG
CATGTTCGGCGATATTTTTGAACAAACTATCCAAAAAGAGTACGGCATGACGATGGGCCACTGGCCGTCAT
CGTTTGGTATCAGTACCGTGGGCGGTTGGGTGGCCTGCCGTGGCGCGGGACAGTTGTCGACCCGCTACGG
CAAAATTGAAGATATGGTTTTTGGGATGGATGTTGTCCTTGCCGATGGCCGCTTAATCACCGTGGGTGGCT
ATGCGCGTAGCGCAACGGGGCCTGATCTACAGCAGATGTTTATTGGTTCTGAAGGCACATTGGGGGGTGAT
TGTCCGTGTGCGTCTGAAACTGCATCGTTTACCTGACTATGGGCGTGCCATCGCCTGGGGTTTTTCCAGCTT
CGCCAGTGGACTGGAGGCATGCCGTGAAATTTTACAGCATGGCGCGAACCCCGCAGCGTTGCGCCTGTAC
GATAACCTGGAAAGCGGCGTGCAGTTTGGTTTACCAGATACCAATGTCCTGTTGATTGCGGATGAGGGGG
AGCCAGAGATTGTCGATGCGGTGCTGGCTATCAGCGAACGTGTGTGCCAACGCAGCGGACAACAACTCGA
CGGCGAGGCTATTTTTGAACGCTGGCTGGACACCCGATATCTGACCGGAAAAAGTGCTGAAGGCTTTAAA
AAGAGCCCGGGTCTGGTGGCGGATACGCTGGAAATGGTGGGACGTTGGCGCGATCTTAGCGACGTGTAT
GATGATGTGGTGGCTGCGATTAATGCCGTACCCGGCACGCTGGCCGGTTCGGCGCATCAGTCTCACGCCT
ACATCGACGGCGCATGCCTCTATTTCTCATTGCGCGGTGATGTGGCGATTGAGCAACGTGCGGCCTGGTAT
CGTGCGGCATGGGATGCAGCGAATGCGGTGCTAATTAAGCACAATACCTCCCTTAGCCACCATCATGGTGT
TGGTCTCCTGCGTTCGCCCTATATGCAGGCTTCTTTAGGCGAATCGCTTTCCGTTCTGGCGGATATCAAATA
TGCGCTGGATCCGAAAAACATTCTTAACCCAGGCAAGCTGGGGTTAAGCGTCGATCTTCCTTCGCATCAGG
CAGGTCAATAA

248

**>gene_id_77908full_aminoacid**

MVALVLAAANARAACSWPAWEQFKQDYISDGGRVIDPSDARKISTSEGQSYALFFALAANDRKAFDLLLTWTS
DNLAQGSLSQHLPAWLWGKKDADTWAVIDKNSASDADIWIAWSLLEAGRLWKAPQYTATGKALLKRIASEEI
KVPGLGLMLLPGNVGFTEEKAWRFNPSYLPPQLANYFTRFGAPWTTLRETNLRLLLETAPKGFAPNWVQYQQ
KKGWQLQPEKTFIGSYDAIRVYLWTGMMHDRDPQKARLLARFKPMATLTTKNGVPPEKVDVASGKPTGDGP
VGFSASLLPFLQDRDAQAVQRQRVADHFPGNDAYYSYVLTLFGQGWDQHRFRFTAKGELHPDWGQECASSH


**>gene_id_77908full_nucleotide**

ATGGTCGCGCTGGTTCTGGCGGCAGCGAATGCGCGTGCGGCCTGTAGCTGGCCCGCGTGGGAGCAGTTTA
AACAGGACTACATCAGCGATGGCGGGCGCGTGATTGATCCCAGTGACGCGCGGAAAATCAGCACTTCGGA
AGGGCAAAGCTATGCGCTGTTCTTTGCCCTGGCCGCCAACGATCGCAAAGCGTTCGATTTACTGCTGACCT
GGACGAGCGACAATCTCGCCCAGGGCTCCCTGAGTCAGCATCTGCCTGCCTGGTTGTGGGGGAAAAAGGA
TGCGGATACCTGGGCGGTGATCGACAAAAACTCTGCGTCTGATGCGGATATCTGGATTGCCTGGTCGTTGC
TGGAAGCGGGGCGTTTGTGGAAAGCGCCGCAATACACCGCCACCGGCAAAGCACTGCTAAAACGCATCGC
CAGCGAAGAAGTGATCAAAGTGCCGGGTTTAGGGCTGATGCTCCTGCCCGGCAACGTCGGTTTTACCGAG
GAGAAAGCCTGGCGCTTTAACCCCAGCTATCTCCCGCCGCAGCTGGCGAACTATTTCACCCGCTTTGGCGC
GCCGTGGACCACGCTTCGCGAGACGAATCTGCGTTTACTGCTGGAAACCGCGCCAAAAGGATTTGCGCCC
AACTGGGTGCAGTATCAGCAAAAAAAAGGCTGGCAATTGCAGCCAGAAAAAACCTTTATCGGCAGTTACG
ACGCGATTCGCGTGTATCTCTGGACGGGCATGATGCACGACCGCGATCCGCAAAAAGCCCGACTGCTGGC
ACGTTTTAAACCGATGGCGACGCTCACAACAAAAAATGGCGTCCCGCCGGAGAAGGTCGATGTCGCAAGC
GGTAAACCCACAGGCGATGGCCCGGTCGGTTTCTCCGCCTCGCTGCTGCCTTTTTTACAGGACCGTGATGC
ACAAGCGGTGCAACGCCAGCGCGTCGCCGACCATTTTCCCGGCAATGACGCCTATTACAGCTACGTGCTGA
CCCTGTTCGGACAAGGATGGGATCAGCATCGTTTTCGCTTCACCGCAAAGGGTGAATTACACCCTGACTGG
GGCCAGGAATGCGCAAGTTCTCATTAA

249

# Appendix 4

**Paper: Characterization of cellulolytic activity in the gut of the terrestrial land slug Arion ater: Biochemical identification of targets for intensive study**

# Characterization of cellulolytic activity in the gut of the terrestrial land slug *Arion ater*: Biochemical identification of targets for intensive study

Ryan Joynson, Arvind Swamy, Paz Aranega Bou, Ambre Chapuis, Natalie Ferry *

*School of Environment and Life Science, University of Salford, M5 4WT, United Kingdom*

## ABSTRACT

The level of cellulolytic activity in different areas of the gut of the terrestrial slug *Arion ater* was assayed at different temperatures and pH values. To do this, crude gut proteins were isolated and assayed using modified dinitrosalicylic acid reducing sugar assay. Crude protein samples were also separated and cellulolytic activity identified using in gel CMC zymography and esculin hydrate activity gel assays. pH and temperature profiling revealed optimum cellulolytic activity between pH 5.0 and 6.0 for different gut regions and retention of up to 90% of activity at temperatures up to 50 °C. Zymograms and activity gels revealed multiple endoglucanase and β-glucosidase enzymes. To further investigate the source of this cellulolytic activity bacterial isolates from the gut were tested for endoglucanase and β-glucosidase activity using growth plate assays. 12 cellulolytic microbes were identified using 16S rDNA gene sequencing. These include members of the genera *Buttiauxella*, *Enterobacter*, *Citrobacter*, *Serratia* and *Klebsiella*. Gut metagenomic DNA was then subjected to PCR, targeting a 400 bp region of the 16SrDNA gene which was subsequently separated and individuals identified using DGGE. This identified members of the genera *Citrobacter*, *Serratia*, *Pectobacterium*, *Acinetobacter*, *Mycoplasma*, *Pantoea* and *Erwinia*. In summary, multiple glycoside hydrolase enzymes active over a broad range of temperature and pH values in a relatively under studied organism were detected, indicating that the gut of *A. ater* is a viable target for intensive study to identify novel carbohydrate active enzymes that may be used in the biofuel industry.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

Lignocellulose derived from plant cell walls is one of the most abundant organic materials on the planet. The most abundant carbohydrate component it contains is cellulose, made solely of 1 β(1 → 4) linked D-glucose units. Three enzymes act sequentially to degrade cellulose into simple sugars, endo-β-1,4-glucanases (endoglucanases; EC. 3.2.1.4), exo-β-1,4-cellobiohydrolases (exocellulases; EC. 3.2.1.91), and β-glucosidases (EC.3.2.1.21). The glucose monosaccharides produced can then be fermented to produce bioethanol. Use of lignocellulose as a bioethanol feedstock has the potential to overcome many of the economic and environmental consequences of using food crops but lignocellulose has an inherent resistance to degradation due to the complexity of the plant cell wall superstructure; current methods require expensive pre-treatments making its use economically unattractive (Ibrahim et al., 2011; Cao et al., 2012). The most promising method for the production of bioethanol from lignocellulose is the simultaneous saccharification and fermentation (SSF) method. This method incorporates lignocellulose degrading enzyme cocktails and fermenting microorganisms or fermenting bacteria metabolically engineered to produce high numbers of lignocellulose degrading enzymes, which are used to produce ethanol from lignocellulose feedstocks. These enzyme cocktails produce monosaccharides which are fermented into ethanol by bacteria such as *Escherichia coli* recombinant strains (Cotta, 2012). Many of these modified strains have been engineered to express highly active cellulase enzymes found in other species. A study by Edwards et al. (2011) showed the benefits of introducing a highly active cellobiase enzyme found in *Klebsiella oxytoca* to *E. coli* strain KO11, which resulted in a 30% increase in ethanol production. Furthermore, cellulase enzymes are also of great importance in the textile industry, in the food industry and as components of detergents, resulting in a high global demand.

To that end there is considerable interest in the potential for microbial enzymes (cellulases, hemicellulases and lignases) to bring about the biological breakdown of lignocellulose. Of particular interest is the scope for degradation by the symbiont microbiota in wood/plant feeding invertebrates. Mutualisms between microbes and insects have been widely studied and are found in almost every case, they facilitate the exploitation of many different food sources by host insects, including plant cell walls which are difficult and sometimes impossible for most animals to digest (Watanabe and Tokuda, 2010). However the enzymatic contributions of microbes to insect herbivory are still

---

unclear. Some herbivorous insects possess genes encoding plant cell wall degrading enzymes including a termite which produces its own cellulase (Watanabe et al., 1998), but the overall structural complexity of the plant cell wall superstructure requires a multitude of enzyme classes which gut microbes contribute to. It is therefore thought that the interactions of host and microbe have had a direct impact on the evolutionary transitions in diet in many herbivorous eukaryotes, including insects (Hansen and Moran, 2014). Enzymatic activity has been studied extensively in the digestive fluid of various insects including members of the orders Isoptera (Konig et al., 2013), Coleoptera (Dojnov et al., 2013) and Orthoptera (Shi et al., 2011), all of which have a high lignocellulose diet. However, this focus on arthropods has been at the expense of other groups such as gastropods. Specifically, there has not yet been a definitive characterization of the origin of cellulolytic activity in the gut of the common garden slug, *Arion ater*, a significant pest throughout Europe. The diet of the slug is extremely varied depending on location and food availability, including fungi, earthworms, leaves, and plant stems along with dead plant material with a preference for young leaf/stem plants. *A. ater* uses its barbed tongue-like appendature called the radula, which contains up to 27,000 teeth, to shred its food. This increases the surface area of its food for enzymatic degradation. The radula also allows the slug to eat even the toughest plant material in times where food is scarce. Due to the large portion of plant material in its diet, it is logical that the gut contains multiple enzymes that allow it to digest plant cell wall material into utilizable simple sugars. The *A. ater* gut is particularly interesting as a potential source of active enzymes given the variation in pH along its digestive tract and its ability to eat twice its body weight in vegetation per day. This efficiency in crop degradation has led to more than £30 million pounds a year being spent on slug pellets in the UK alone and a ~70-fold increase in utilization of molluscicides over three decades (Agular and Wink, 2005). Consequently, we have carried out in-depth analysis of the cellulolytic activity and associated microbial community of the terrestrial gastropod *A. ater*.

## 2. Materials and methods

### 2.1. Slug collection and dissection

Slugs were collected from a suburban area in North Cheshire (53.391463 N, 2.211214 W) two hours after last light. Individuals were allowed to feed on celery/lettuce cores for 12 h. Individuals were cooled to 4 °C prior to dissection to reduce metabolism and spontaneous mucus production during dissection. Whole gut tracts were removed, avoiding rupture that would result in loss or contamination of gut juices. Mucus that might interfere with the assays was removed by blotting. Total guts were further separated into 'crop' which denotes the region from the mouth up to and including the digestive gland and the 'gut' which corresponds to the gut after the stomach/digestive gland up to the anus (Fig. 1).
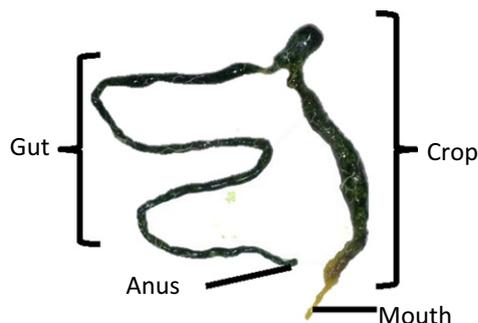


Fig. 1. Dissected whole gut tract of *Arion ater*.

### 2.2. Initial detection of total cellulolytic activity

Gut samples were cut up using a scalpel in a petri dish and then homogenized with a sterile glass rod in a 1.5 mL tube containing 200 μL of 0.2 M sodium acetate buffer (pH 5.2) followed by vigorous vortexing. To clear cell debris and food matter, samples were centrifuged at 13.3 krpm for 5 min. Supernatants were extracted, pooled (subsequently referred to as 'crude protein samples') and stored at −80 °C. Protein content of the crude samples was estimated using a standard Bradford assay (Bradford, 1976) using BSA to construct the standard curve. Total cellulase activity was measured using the dinitrosalicylic acid (DNSA) cellulase activity assay of Ghose (1987) with slight adjustments. This assay allows the detection of cellulolytic enzymes that hydrolyze cellulose internally or externally along with the breakdown of cellobiose, each of these actions produces reducing sugar free carbonyl groups which are measured in this assay. The cellulolytic activities of 50 μL of crop and gut samples were tested by mixing 1% carboxymethyl cellulose (CMC) (Sigma Aldrich) in a 100 mM sodium citrate buffer (pH 4.5). Samples were incubated at 50 °C for 30 min. Reactions were terminated by placing samples on ice, adding DNS reagent and heating to 95 °C for 10 min to allow color development. All samples were tested and boiled simultaneously. Samples were cooled to room temperature and absorbance read at 540 nm using a CMC control sample as a blank. Correction for background sugars in the sample was undertaken by subtracting a time 0 duplicate sample absorbance from the final result. All activities in this paper are expressed as specific enzyme activities, where 1 U/mg is equal to 1 μM glucose released per min per mg of protein.

### 2.3. pH and temperature profiling of crude protein cellulolytic activity

The cellulase detection assay previously described was modified to measure the pH profile of the crude protein cellulolytic activity against CMC, replacing the pH 4.5 buffer with 100 mM sodium citrate buffers ranging between pH 4 and 9 while all other conditions remained the same. To determine the temperature profile of the crude protein sample, the assay was modified by varying incubation temperature between 20 °C and 70 °C.

### 2.4. Identification of endoglucanases using CMC SDS PAGE zymography

CMC zymography was carried out following the procedure of Schwarz et al. (1987) and Willis et al. (2010). Samples were run using a 12% acrylamide SDS gel containing 0.2% CMC as a substrate for activity staining. Before polymerization was induced, solutions were heated to 30 °C and CMC was added slowly to the resolving gel mixture. Gels were allowed to polymerize for 2 h and used the same day. Crop and gut crude protein samples were thawed on ice followed by the addition of a modified Laemmli loading buffer (minus denaturants). Samples were then heated to 80 °C for 10 min followed by pulse centrifugation to denature proteins and prevent substrate digestion during electrophoresis. Size determination and separation was conducted by using 50 μg of each crude extract along with 15 μL of SeeBlue® Plus2 Pre-Stained Standard (Invitrogen). Gels were run at a constant 100 V for 4.5 h. For size estimation, the distances traveled by the pre-stained standard bands were measured prior to incubation/staining steps which cause the standards to become difficult to visualize, estimated Mw of bands is indicated on gels by an arrow at appropriate position. The CMC gel was washed in a 5% tritron X-100 solution for 30 min (repeated 5 times) to remove SDS. The gel was then rinsed with distilled water, placed in sodium phosphate buffer (50 mM, pH 6.5) and incubated for 2 h at 4 °C to exchange the buffer system and allow the renaturation of proteins in the gel. Phosphate buffer was refreshed and the gel was then incubated at 37 °C overnight. Following incubation, the gels were stained with 0.1% (w/v) Congo red for 1 h, and then destained with a 1 M NaCl solution for 3 h. To enhance the visualization of clear zones

acetic acid was added drop wise to the NaCl solution containing the gel, turning the Congo Red from red to a deep purple.

### 2.5. Identification of β-glucosidase enzymes using esculin hydrate — ferric ammonium citrate Native PAGE activity gel

Native PAGE was carried out as described by Kwon et al. (1994) using a 12% native tris–glycine gel. Native loading buffer was added to crop and gut crude protein extracts and 50 μg of each loaded on the gel. Gels were run at 100 V for 4 h. The gel was then placed in a 0.2 M sodium acetate buffer (pH 5.5) for 10 min to exchange the buffer system, then the gel was placed in a 0.2 M sodium acetate buffer (pH 5.5) containing 0.1% (w/v) esculin hydrate (Sigma) and 0.03% (w/v) ferric ammonium citrate (Sigma) and incubated for 3 h at 37 °C to allow in gel hydrolytic activity. Where β-glucosidase enzymes are present esculin is cleaved producing esculitin which goes onto react with ferric iron to produce a black precipitate. To stop the reaction, the gel was placed into a 10% glucose solution.

### 2.6. Identification of culturable cellulolytic microbes using esculin and CMC LB agar plate assays

Whole guts were extracted as previously described and homogenized in 500 μL of 1 quarter strength Ringer's solution. A range of dilutions was placed on LB agar plates containing 0.5% CMC and grown overnight at 25 °C. Replica plates were created and incubated for a further 24 h. This prevents false identification of cellulolytic bacteria through clearance zones caused by extracellular endoglucanase enzymes in the plated gut fluid. Replica plates were stained with a 0.1% Congo red solution for 1 h, followed by destaining with 1 M NaCl for a further hour. Colonies corresponding with zones of clearance were isolated from replica plates, grown overnight in LB broth. Isolates were then plated onto LB agar containing 0.1% esculin and 0.03% ferric ammonium citrate and incubated at 25 °C for 3 h to confirm β-glucosidase activity. Isolates were identified using 16S rDNA PCR using primers 8F (5′-AGAGTTTGATCCTGGCTC-3′) and 1512R (5′-ACGGCTACCTTGTTACGA-3′). Each amplified PCR product was sequenced using Sanger sequencing system big dye v3.1. Sequences were searched using BLASTn for matches in the 16S rDNA database in order to identify each microbe.

### 2.7. Culture independent microbe identification using DGGE analysis

Other members of the *A. ater* gut community were identified using denaturing gradient gel electrophoresis (DGGE). Metagenomic DNA was extracted from a whole gut using a modified version of the Meta-G-nome DNA isolation kit protocol (Epicentre) and extracted DNA was subjected to PCR targeting a 400 bp region of the 16S rDNA using primers F984GC and R1378 according to Heuer et al. (1997). PCR products were separated by sequence variation using a 30–60% gradient of urea and formamide in a polyacrylamide gel, using the protean 2 system run at a constant 100 V for 16 h at 60 °C. Gels were stained with Gel Red™ (Biotium, Inc.) and individual bands were excised and placed into wells of a 1% agarose gel and electrophoresed into agarose. Bands were then extracted using the Wizard gel extraction kit (Promega) and sequenced using big dye v3.1. Sequences were submitted to BLASTn for bacterial identification against the 16S rDNA database.

## 3. Results

### 3.1. Measurement of cellulolytic activity in A. ater gut samples

Total cellulase activity in the crop and gut regions (Fig. 1) of *A. ater* was assayed (Fig. 2A). Cellulase activity was observed in both the gut
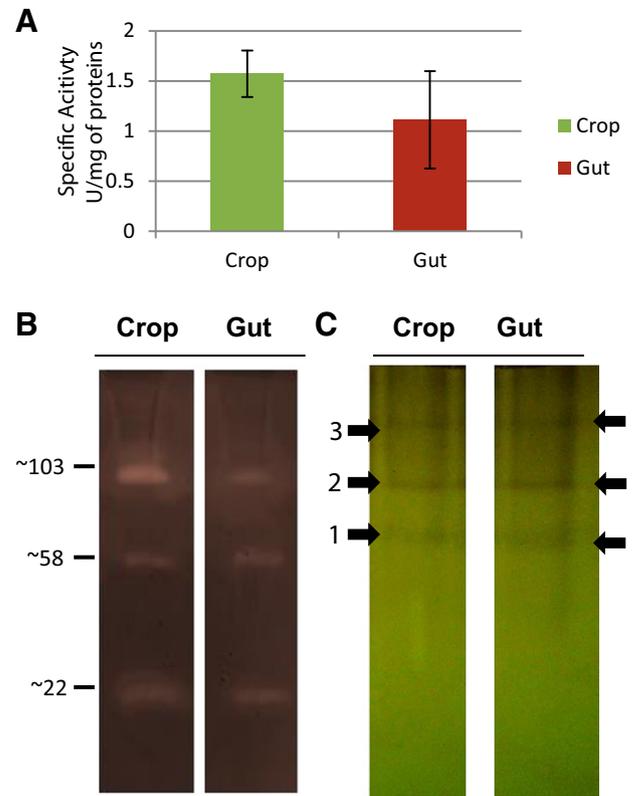


**Fig. 2.** (A) Total specific cellulolytic activity seen in the gut fluids from *Arion ater* against CMC at 50 °C and pH 5.0 using the cellulase activity assay described in Section 2.2. (B) A 12% SDS PAGE 0.2% CMC zymogram using 50 μg of crude gut and crop protein per lane. Gel stained with Congo red to allow activity visualization. Labels indicate approximate size of cellulolytic proteins (kDa). (C) A 12% native PAGE gel containing 100 μg of protein per lane, gels were incubated in a 0.2 M sodium acetate activity buffer containing 0.1% (w/v) esculin and 0.03% w/v ferric ammonium citrate for one hour. Black precipitate shows areas of activity, indicated by black arrows.

and crop with the crop portion showing the highest activity at 1.57 U/mg of protein and the gut showing 1.11 U/mg of protein.

### 3.2. Temperature and pH profiling of total gut cellulolytic activity

Both gut and crop samples showed resilience to heat up to around 50 °C at which point activity begins to decline, with both crude samples showing the greatest activity at 30–35 °C (Fig. 3A). The pH profiles for the two samples were however quite distinct, with the crop samples showing greatest activity at pH 5 and gut at pH 6 (Fig. 3B). At pH values higher than 6.5 the activity of both samples begins to decline up to pH 9 at which point activity is ~4 fold lower than at optimum pH for each sample.

### 3.3. CMC zymography and esculin hydrate activity gel assays

Due to the differences seen in the crop and gut cellulolytic activity profiles, CMC zymography (Fig. 2B) and esculin hydrate activity gel assays (Fig. 2C) were carried out in order to identify whether or not similar enzyme systems were being incorporated in the crop and gut digestive juices. In CMC zymograms we observed almost identical cellulose activity patterns. We observed 3 main bands in both crude samples, corresponding to proteins of approximately 103, 58 and 22 kDa in size. The β-glucosidase activity gels showed three bands at positions 1, 2 and 3 (indicated with black arrows) which appear to be at identical locations in the gel for both the gut and crop samples.
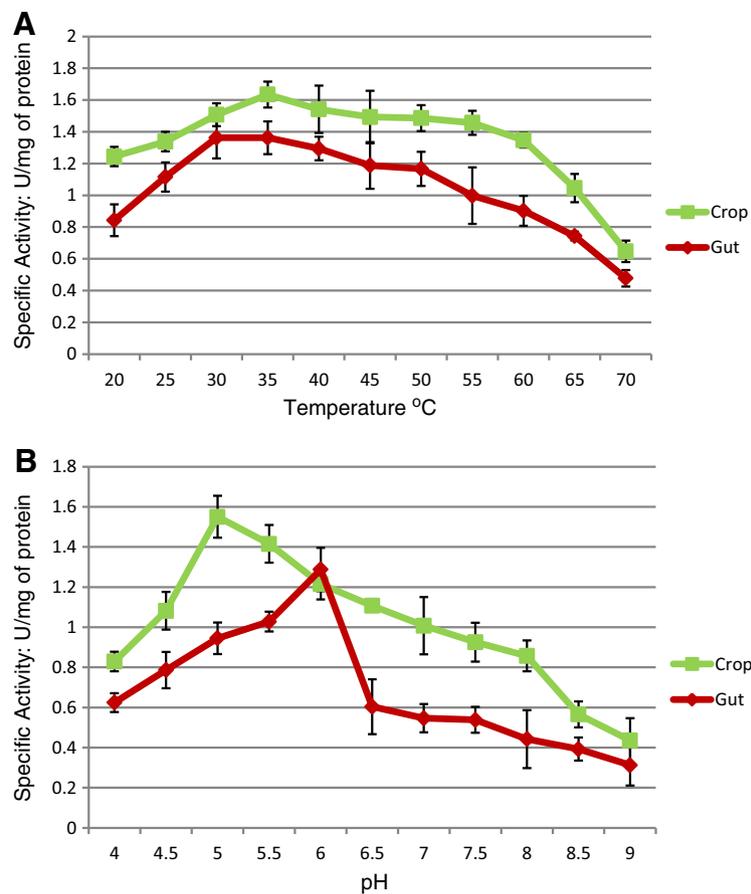
**Fig. 3.** Temperature profiles (A) and pH profiles (B) of the two crude gut protein isolations showing the total cellulolytic activity of each sample against a CMC substrate. Temperature and pH profiles were obtained using a modified cellulase assay with incubation steps at temperatures between 20 and 70 °C and at pH values between 4 and 9 respectively. Specific activity shown as enzyme units (U) where 1 U is equal to 1 μM glucose released per min per mg of protein.

### 3.4. Identification of cellulolytic microorganisms

To gain an insight into the origin of some of the cellulolytic activity seen in this study, gut microorganisms were isolated and tested for cellulolytic activity. Microbial isolates were grown on agar containing CMC and on agar containing ferric ammonium citrate and esculin hydrate to identify endoglucanase (Fig. 4B) and β-glucosidase (Fig. 4A) activity respectively. Twelve isolates showed both endoglucanase and β-glucosidase activity, including members of *Aeromonas*, *Acinetobacter*,

*Buttiauxella*, *Citrobacter*, *Enterobacter*, *Klebsiella*, *Kluyvera*, *Salmonella* and *Serratia* (Table 1). Only 4 of these microbes could be identified to within 97% similarity of bacterial 16S rDNA genes in the NBCI 16S rDNA and NR databases while the remaining 8 were seen to have

**Table 1**

NCBI BLASTn search results for each amplified 16S rDNA gene from cultured cellulolytic microbes (CA.a.*) and for uncultured microbes from the DGGE study (UA.a.*). Sequences were queried against the NCBI 16S rRNA database or the nr database if no match was found.

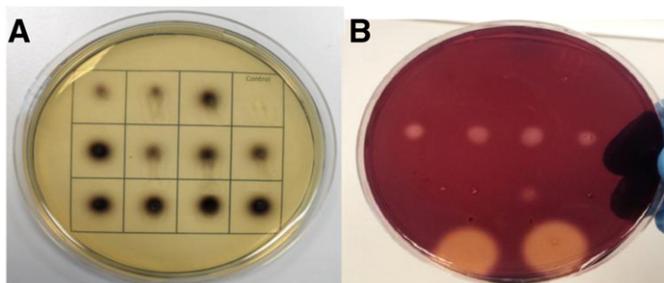| Name | Description | E-value | Identity | Accession |
|---|---|---|---|---|
| CA.a.1 | *Acinetobacter calcoaceticus* | 0 | 92% | NR_042387.1 |
| CA.a.2 | *Aeromonas hydrophila* | 0 | 99% | NR_104824.1 |
| CA.a.3 | *Buttiauxella agrestis* | 0 | 79% | DQ440549.1 |
| CA.a.4 | *Buttiauxella agrestis* | 0 | 99% | NR_041968.1 |
| CA.a.5 | *Citrobacter braakii* | 0 | 85% | NR_028687.1 |
| CA.a.6 | *Citrobacter freundii* | 0 | 99% | NR_028894.1 |
| CA.a.7 | *Enterobacter* sp. E6-PCAi | 0 | 94% | JN853247.1 |
| CA.a.8 | *Klebsiella pneumoniae* | 0 | 96% | NR_037084.1 |
| CA.a.9 | *Kluyvera intermedia* | 0 | 99% | KF724024.1 |
| CA.a.10 | *Salmonella enterica* | 0 | 91% | NR_044371.1 |
| CA.a.11 | *Serratia liquefaciens* | 0 | 86% | GU586145.1 |
| CA.a.12 | *Serratia marcescens* | 0 | 91% | NR_036886.1 |
| UA.a.1 | *Mycoplasma hyorhinis* | 1.00E-158 | 93% | NR_041845.1 |
| UA.a.2 | *Mycoplasma iners* | 4.00E-158 | 93% | NR_025064.1 |
| UA.a.3 | *Uncultured Citrobacter* | 0 | 99% | AY847172.1 |
| UA.a.4 | *Uncultured Serratia* | 0 | 100% | KC253894.1 |
| UA.a.5 | *Pectobacterium carotovorum* | 0 | 99% | NR_041971.1 |
| UA.a.6 | *Acinetobacter beijerinckii* | 0 | 98% | NR_042234.1 |
| UA.a.7 | *Pantoea* sp. 57917 | 0 | 99% | DQ094146.1 |
| UA.a.8 | *Erwinia amylovora* | 0 | 99% | NR_041970.1 |
| UA.a.9 | *Erwinia tasmaniensis* | 0 | 99% | NR_074869.1 |



**Fig. 4.** (A) An esculin hydrate plate assay demonstrating the β-glucosidase activity of microbial isolates. Isolates were grown on agar plates containing 0.1% (w/v) esculin and 0.03% (w/v) ferric ammonium citrate. A black precipitate indicates β-glucosidase activity. Untransformed top 10 *E. coli* (Invitrogen) was used as a negative control. (B) A CMC plate assay showing endoglucanase activity. Bacterial isolates were grown on agar plates containing 0.5% CMC after 16 h incubation plates were stained with congo red and destained with 1 M NaCl in order to visualize zones of clearing. 5 and 10 μL of 1 mg/mL *A. niger* cellulase were used as a positive control (indicated by black arrows).
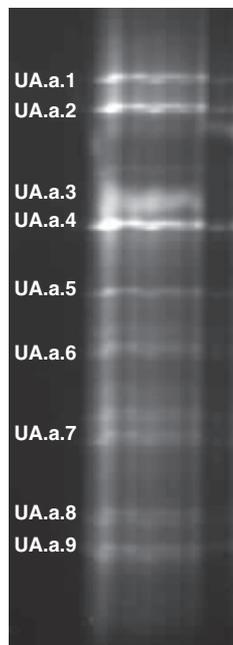
**Fig. 5.** Differential gradient gel electrophoresis gel, 30–60% gradient of formamide and urea. Labels show bands from which successful microbial identifications were deduced.

between 96 and 79% similarity to database entries. Subsequently, a DGGE study was carried out to identify microbes that might be present but which may be less easy to culture, using metagenomic DNA samples as templates for 16S rDNA targeted PCR (Fig. 5). This revealed multiple bands from which DNA was extracted and sequenced. Nine further microbes were identified, from the genera *Citrobacter*, *Serratia*, *Pectobacterium*, *Acinetobacter*, *Mycoplasma*, *Pantoea* and *Erwina* (Table 1). Sequences for cultured and uncultured 16S rDNA studies can be seen in Supplementary file 1.

## 4. Discussion

This study has further characterized the cellulolytic activity in the gut of *A. ater* through biochemical testing of different portions of the gut, along with the identification of multiple cellulolytic microorganisms and thus we begin to characterize the *A. ater* gut microbiome. Cellulase activity assays showed the overall cellulolytic activity in the gut of *A. ater* found in the North of England to be greater than that of many insects (Oppert et al., 2010), including members of the orders *Coleoptera*, *Isoptera*, *Orthoptera* and *Diptera*. We also demonstrate relative stability across a wide pH and temperature range, with optimal activity at pH values that would be feasible for use in modern industrial lignocellulose degradation methods. A separate investigation of the cellulolytic activity of *A. ater* of North American origin by James et al. (1997) showed higher overall cellulolytic activity than in this study, but with an optimal pH of 7 as opposed to the crop optimum of pH 5 observed here. A possible reason for this observed difference in optimal pH is the native environment from which individuals were taken, with the average soil pH for the area of North Cheshire being <5.0, whereas in Bellingham WA, the soil is at a pH of between 6 and 6.6, each correlating with the optimal pH values observed. Acidic environments have been observed in multiple land pulmonates such as *Helix aspersa*, (6.1–7.4) *Helix pomatia* (5.5–6.4), *Elona quimperiana* (5.3–6.6) (Charrier and Brune, 2003) and *Pomacea canaliculata* (6.0–7.4) (Godoy et al., 2013) which would suggest that members of this class harbor dietary enzymes that can function in acidic environments, including *A. ater*, as we have observed. Also, the cellulolytic systems appear to have varying temperature profiles, with our study showing crop

and gut samples retaining 90% and 85% activity respectively at 50 °C (where 100% is the highest activity observed in each assay for each crude sample) while the study of the North American species shows practically no activity against CMC in the same conditions. It is also important to note that the gut microbiome is a very dynamic environment which can be heavily altered by living in a different habitat, this has been demonstrated not only in humans (Huttenhower et al., 2012), but also in insects (Dillon and Dillon, 2004). The temperature profile we observed shows the crude enzyme extracts retain much of their activity even at 50 °C and demonstrates no clear optimum temperature. However this is not surprising when the complexity of the crude mixture is taken into account, as having multiple enzymes of different microbial origin would cause there to be variation in optimum temperatures for activity for cellulase enzymes of different glycoside hydrolase groups and even within groups.

Using modified cellulase zymograms and esculin hydrate activity gel assays we have also identified three highly abundant individual endoglucanase and β-glucosidase enzymes present in both the crop and gut juices, thereby demonstrating a similar cellulolytic system throughout the gut and therefore suggesting little activity compartmentalization throughout the gut regions. It is also important to take into consideration that the minimum detectable amount of active enzyme in the esculin hydrate activity gel assay is relatively low at >10 ng (Kwon et al., 1994). Our discovery of multiple endoglucanase and β-glucosidase producing bacteria suggests that there are much greater numbers of individual cellulolytic enzymes present than we observed in our gel methods. The individual microbes isolated may not make up a high enough proportion of the gut microbiome to produce their enzymes in sufficient abundances to be detectable using in gel separation methods.

Our study also confirmed that at least a portion of the cellulolytic activity seen in the gut of *A. ater* is due to symbiotic activity of gut microbes and, for the first time, isolated and identified individual cellulolytic microbes. Many studies have carried out growth plate assays successfully, quickly and accurately isolating gut cellulolytic microbes from gastropods (Antonio et al., 2010), insects (Huang et al., 2012) and mammals (Ruijssenaars and Hartmans, 2001). CMC and esculin hydrate activity growth plate assays allowed us to identify 12 cellulolytic gut microbes, only 4 of which could be identified with great confidence (>97% similarity). This strongly suggests that the *A. ater* gut microbiome contains uncharacterized microbes with uncharacterized cellulolytic systems that we have shown to have robust pH and temperature activity profiles. In the non-culture based DGGE study we identified 9 further microbes of which *Pectobacterium carotovorum*, *Erwinia amylovora* and *Erwinia tasmaniensis* species all have cellulolytic enzymes linked to their species in the NCBI database (http://www.ncbi.nlm.nih.gov/). In this study we have identified a high number of members of the gut that belong to the *Gammaproteobacteria* class, with only two *Mycoplasma* species being from outside that class. The microbes *Klebsiella pneumonia*, *Citrobacter freundii* and *Serratia liquefaciens* have also been identified in the gut of the *Bombyx mori* larvae (silk worm) and their cellulolytic activity was also observed (Anand et al., 2010). Multiple *Enterobacter* species, the species *Salmonella enterica* and *Serratia marcescens* have also been identified in the gut of herbivorous beetle larvae during their development (Azambuja et al., 2004; Butera et al., 2012). Further to this, a metagenomic study into the gut microbiome of the giant African Snail interestingly shares all but one of the microbial species identified here (Cardoso et al., 2012), this suggests that there may be a set of gut microbes on which multiple land gastropods rely to aid their digestion of lignocellulose. This also indicates that the gut microbe host interaction could have played an important role in the evolutionary dietary transitions of land gastropods as it is thought to have in insects (Hansen and Moran, 2014).

Gastropods have not been the main focus of recent cellulase prospecting studies due to the initial successes with the insect families,

most notably studies looking at termites (Tokuda and Watanabe, 2007) but also beetles (Wei et al., 2006a,b) and grasshoppers (Oppert et al., 2010; Willis et al., 2010). However the recent study into the microbiome of the giant African snail has identified thousands of glycoside hydrolase enzymes and carbohydrate binding modules of microbial origin (Cardoso et al., 2012). Our findings and these promising results from related species give a strong indication that the gut of *A. ater* is a viable target for more intense study to identify novel, plant cell wall degrading enzymes which may be key to improving contemporary biochemical methods in the biofuel industry. In addition, further understanding of the essential biochemical pathways involved in slug feeding could be used to develop more target-specific pest control measures for slugs. Here for example, the identification of these different classes of enzymes demonstrates that the slug gut has the capability to digest the cellulose portion of its diet from long polymer cellulose to individual, utilizable, glucose monosaccharides. This therefore confirms that the slug has the ability to efficiently utilize the cellulose portion of plant matter it consumes as a source of carbon and we have also identified that gut microbes play a significant role in making this glucose accessible. Increases in physiological understanding are especially important given the detection of high levels of the generic slug pellet poison metaldehyde in water in the UK (Kay and Grayson, 2013) and the recent European Union regulation, which imposes a complete ban on sales of traditional slug pellets by 19th September 2014 (Commission Implementing Regulation 187/2014).

## Acknowledgments

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cbpb.2014.08.003.

## References

Agular, R., Wink, M., 2005. How do slugs cope with toxic alkaloids? Chemoecology 15, 167–177.
Anand, A.A.P., Vennison, S.J., Sankar, S.G., Prabhu, D.I.G., Vasan, P.T., Raghuraman, T., Geoffrey, C.J., Vendan, S.E., 2010. Isolation and characterization of bacteria from the gut of Bombyx mori that degrade cellulose, xylan, pectin and starch and their impact on digestion. J. Insect Sci. 10.
Antonio, E.S., Kasai, A., Ueno, M., Kurikawa, Y., Tsuchiya, K., Toyohara, H., Ishihi, Y., Yokoyama, H., Yamashita, Y., 2010. Consumption of terrestrial organic matter by estuarine molluscs determined by analysis of their stable isotopes and cellulase activity. Estuar. Coast. Shelf Sci. 86, 401–407.
Azambuja, P., Feder, D., Garcia, E.S., 2004. Isolation of Serratia marcescens in the midgut of Rhodnius prolixus: impact on the establishment of the parasite Trypanosoma cruzi in the vector. Exp. Parasitol. 107, 89–96.
Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. 72, 248–254.
Butera, G., Ferraro, C., Colazza, S., Alonzo, G., Quatrini, P., 2012. The culturable bacterial community of frass produced by larvae of Rhynchophorus ferrugineus Olivier (Coleoptera: Curculionidae) in the Canary island date palm. Lett. Appl. Microbiol. 54, 530–536.
Cao, W.X., Sun, C., Liu, R.H., Yin, R.Z., Wu, X.W., 2012. Comparison of the effects of five pretreatment methods on enhancing the enzymatic digestibility and ethanol production from sweet sorghum bagasse. Bioresour. Technol. 111, 215–221.
Cardoso, A.M., Cavalcante, J.J., Cantao, M.E., Thompson, C.E., Flatschart, R.B., Glogauer, A., Scapin, S.M., Sade, Y.B., Beltrao, P.J., Gerber, A.L., Martins, O.B., Garcia, E.S., de Souza, W., Vasconcelos, A.T., 2012. Metagenomic analysis of the microbiota from the crop of an invasive snail reveals a rich reservoir of novel genes. PLoS ONE 7, e48505.
Charrier, M., Brune, A., 2003. The gut microenvironment of helicid snails (Gastropoda: Pulmonata): in-situ profiles of pH, oxygen, and hydrogen determined by microsensors. Can. J. Zool. 81, 928–935.
Cotta, M.A., 2012. Ethanol production from lignocellulosic biomass by recombinant Escherichia coli strain FBR5. Bioengineered 3, 197–202.
Dillon, R.J., Dillon, V.M., 2004. The gut bacteria of insects: nonpathogenic interactions. Annu. Rev. Entomol. 49, 71–92.
Dojnov, B., Pavlovic, R., Bozic, N., Margetic, A., Nenadovic, V., Ivanovic, J., Vujcic, Z., 2013. Expression and distribution of cellulase, amylase and peptidase isoforms along the

midgut of Morimus funereus L. (Coleoptera: Cerambycidae) larvae is dependent on nutrient substrate composition. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 164, 259–267.
Edwards, M.C., Henriksen, E.D., Yomano, L.P., Gardner, B.C., Sharma, L.N., Ingram, L.O., Peterson, J.D., 2011. Addition of genes for cellobiase and pectinolytic activity in Escherichia coli for fuel ethanol production from pectin-rich lignocellulosic biomass. Appl. Environ. Microbiol. 77, 5184–5191.
Ghose, T.K., 1987. Measurement of cellulase activities. Pure Appl. Chem. 59, 257–268.
Godoy, M.S., Castro-Vasquez, A., Vega, I.A., 2013. Endosymbiotic and host proteases in the digestive tract of the invasive snail Pomacea canaliculata: diversity, origin and characterization. PLoS ONE 8, e66689.
Hansen, A.K., Moran, N.A., 2014. The impact of microbial symbionts on host plant utilization by herbivorous insects. Mol. Ecol. 23, 1473–1496.
Heuer, H., Krsek, M., Baker, P., Smalla, K., Wellington, E.M., 1997. Analysis of actinomycete communities by specific amplification of genes encoding 16S rRNA and gel-electrophoretic separation in denaturing gradients. Appl. Environ. Microbiol. 63, 3233–3241.
Huang, S.W., Sheng, P., Zhang, H.Y., 2012. Isolation and identification of cellulolytic bacteria from the gut of Holotrichia parallela larvae (Coleoptera: Scarabaeidae). Int. J. Mol. Sci. 13, 2563–2577.
Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q.D., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Brooks, J.P., Buck, G.A., Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Dunne, W.M., Durkin, A.S., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H.Y., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haake, S.K., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H.Y., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.D.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.C., Lozupone, C.A., Lunsford, R.D., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J.G., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Sathirapongsasuti, J.F., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z.Y., Ward, D.V., Warren, W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu, Y.Q., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y.Z., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y.J., Zhu, Y.M., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methe, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K., White, O., Consortiu, H.M.P., 2012. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214.
Ibrahim, M.M., El-Zawawy, W.K., Abdel-Fattah, Y.R., Soliman, N.A., Agblevor, F.A., 2011. Comparison of alkaline pulping with steam explosion for glucose production from rice straw. Carbohydr. Polym. 83, 720–726.
James, R., Nguyen, T., Arthur, W., Levine, K., Williams, D.C., 1997. Hydrolase (beta-glucanase, alpha-glucanase, and protease) activity in Ariolimax columbianus (banana slug) and Arion ater (garden slug). Comp. Biochem. Physiol. B Biochem. Mol. Biol. 118, 275–283.
Kay, P., Grayson, R., 2013. Using water industry data to assess the metaldehyde pollution problem. Water Environ. J. 28, 410–417.
Konig, H., Li, L., Frohlich, J., 2013. The cellulolytic system of the termite gut. Appl. Microbiol. Biotechnol. 97, 7943–7962.
Kwon, K.S., Lee, J., Kang, H.G., Hah, Y.C., 1994. Detection of beta-glucosidase activity in polyacrylamide gels with esculin as substrate. Appl. Environ. Microbiol. 60, 4584–4586.
Oppert, C., Klingeman, W.E., Willis, J.D., Oppert, B., Jurat-Fuentes, J.L., 2010. Prospecting for cellulolytic activity in insect digestive fluids. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 155, 145–154.
Ruijssenaars, H.J., Hartmans, S., 2001. Plate screening methods for the detection of polysaccharase-producing microorganisms. Appl. Microbiol. Biotechnol. 55, 143–149.
Schwarz, W.H., Bronnenmeier, K., Grabnitz, F., Staudenbauer, W.L., 1987. Activity Staining of Cellulases in Polyacrylamide Gels Containing Mixed Linkage Beta-Glucans. Anal. Biochem. 164, 72–77.
Shi, W.B., Ding, S.Y., Yuan, J.S., 2011. Comparison of insect gut cellulase and xylanase activity across different insect species with distinct food sources. Bioenergy Res. 4, 1–10.
Tokuda, G., Watanabe, H., 2007. Hidden cellulases in termites: revision of an old hypothesis. Biol. Lett. 3, 336–339.

Watanabe, H., Tokuda, G., 2010. Cellulolytic systems in insects. Annu. Rev. Entomol. 55, 609–632.

Watanabe, H., Noda, H., Tokuda, G., Lo, N., 1998. A cellulase gene of termite origin. Nature 394, 330–331.

Wei, Y.D., Lee, K.S., Gui, Z.Z., Yoon, H.J., Kim, I., Je, Y.H., Lee, S.M., Zhang, G.Z., Guo, X., Sohn, H.D., Jin, B.R., 2006a. N-linked glycosylation of a beetle (*Apriona germari*) cellulase Ag-EGase II is necessary for enzymatic activity. Insect Biochem. Mol. Biol. 36, 435–441.

Wei, Y.D., Lee, K.S., Gui, Z.Z., Yoon, H.J., Kim, I., Zhang, G.Z., Guo, X., Sohn, H.D., Jin, B.R., 2006b. Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, *Apriona germari*. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 145, 220–229.

Willis, J.D., Klingeman, W.E., Oppert, C., Oppert, B., Jurat-Fuentes, J.L., 2010. Characterization of cellulolytic activity from digestive fluids of *Dissosteira carolina* (Orthoptera: Acrididae). Comp. Biochem. Physiol. B Biochem. Mol. Biol. 157, 267–272.