Mining saltmarsh sediment microbes for enzymes to degrade recalcitrant biomass

Juliana Sanchez Alponti

PhD

University of York

Biology

September 2019

Abstract

The recalcitrance of biomass represents a major bottleneck for the efficient production of fermentable sugars from biomass. Cellulase cocktails are often only able to release 75-80% of the potential sugars from biomass and this adds to the overall costs of lignocellulosic processing. The high amounts of fresh water used in biomass processing also adds to the overall costs and environmental footprint of this process. A more sustainable approach could be the use of seawater during the process, saving the valuable fresh water for human consumption and agriculture. For such replacement to be viable, there is a need to identify salt tolerant lignocellulose-degrading enzymes. We have been prospecting for enzymes from the marine environment that attack the more recalcitrant components of lignocellulosic biomass. To achieve these ends, we have carried out selective culture enrichments using highly degraded biomass and inoculum taken from a saltmarsh. Saltmarshes are highly productive ecosystems, where most of the biomass is provided by land plants and is therefore rich in lignocellulose. Lignocellulose forms the major source of biomass to feed the large communities of heterotrophic organisms living in saltmarshes, which are likely to contain a range of microbial species specialised for the degradation of lignocellulosic biomass. We took biomass from the saltmarsh grass *Spartina anglica* that had been previously degraded by microbes over a 10-week period, losing 70% of its content in the process. This recalcitrant biomass was then used as the sole carbon source in a shake-flask culture inoculated with saltmarsh sediment. Cultures were grown for 8 weeks and then analysed using meta-omic approaches. Meta-genomics were used to investigate the microbial community present in the final recalcitrant biomass, while combined meta-proteomics and meta-transcriptomics were used to identify putative CAZymes (Carbohydrate active enzymes). Candidate enzymes have been cloned, heterologous expressed in *E. coli* and characterized according to their salt tolerance.

List of contents

List of tables

List of figures

List of figures

Acknowledgments

I would like to thank God and Our Lady for being always present in my life, guiding me through difficult times and helping me towards the realization of this 'dream' of doing a PhD abroad.

I would like to thank both my supervisors, Prof. Simon McQueen Manson and Prof Neil Bruce, who have trusted me in the development of this project and who have provided me with immense support and guidance during these four years. I am sincerely grateful to you and I could not have asked for better mentors. Thank you so much!

Besides my supervisors, I would also like to thank both my Thesis Advisory Panel (TAP) members Prof. Peter Young and Prof. Gavin Thomas for their invaluable advice and for having always encouraged me in the TAP meetings. Thank you, this work would not have been the same without your feedback.

I would like to thank the National Council for Scientific and Technological Development (CNPq; process number: 232506/2014-0) for providing me with the financial support to develop this project during these four years. Also to The Radhika V. Sreedhar Scholarship Fund from the Department of Biology for the financial aid provided me during the final months of my research.

I also would like to express my sincere gratitude to Daniel Leadbeater, who has shared much of his knowledge about saltmarshes, bioinformatics and methodologies with me; to Luisa Elias for sharing all her knowledge regarding cloning and protein expression; to Katrin Besser for the immense help with methodologies, data analysis and discussions; Giovanna Pesante for the help with protein purification; Carla Bothelho Machado and Mariana Silva for their friendship. I am also very happy to have been part of the Centre for Novel Agricultural Project (CNAP), where I have learned immensely and want to thank all of its other members in no particular order: Alexander Setchfield; Alexandra Lanot; Amira Abood (thank you for your help with ion-exchange purification); Aritha Dornau, Caragh Whitehead, Claire-Steele-King, Daniel Upton, David Neale, Federico Sabbadin, Heather Eastmond, Janina Hossbach, Jessica Dobson, Julia Crawford, Laura Faas, Leonardo Gomez, Linda Sainty, Liz Rylott, Margaret Cafferky, Nicola Oates, Paulina Dani, Rachael Hallam, Susannah Bird (thank you for some clarifications about the DNA sequencing analysis), Thierry Tonon and Veronica Ongaro.

Also a special thanks to other members of the University and staff members: Dr. Yi Li from CNAP for the contig assembly of my RNA sequencing; Dr. Adam Dowle and Dr. Swen Langer from the Technology Facility for the proteomic studies and ferulic acid analysis; Dr. Deborah Rathbone and Susan Heywood from the Biorenewables Development Centre (BDC), for all the help regarding DNA

sequencing and for giving me permission to use BDC's facility; Monica Bandeira and Amanda Barnes from the Graduate office for their friendly and helpful assistance anytime needed; and David Nelmes for his assistance anytime computers decided to give me troubles.

Last but not least, I would like to thank my family: To my mum and Godmother Cidinha; even though I was not born from her, she is the person I love unconditionally since the first second of my life and who I am completely sure will be by my side forever. My dad Domingos, who even if on another plane, I know is always by my side. My aunt who is also my second mother Regina, who has always supported me at all times. My cousins 'sisters and brothers' Regiane and Joao Vitor, who have always been by my side supporting me in any decisions. Finally, my partner Adam for his support, dedication and sympathy even during the hardest moments of this journey; without him I am not sure I would have mentally survived this PhD… I love you all!

Declaration

I declare that this thesis is a presentation of original work and I am the sole author, except for where due reference has been given to colleagues and collaborators. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1 General introduction

1.1 Lignocellulose biomass

Global commitments to reduce greenhouse gas emissions mean that we can no longer rely on fossil fuels to produce the commodities that drive our industrial economy. Academia, industrial and governmental bodies have been mobilizing sectors in order to develop technologies to replace the high usage of fossil fuels up to date. Although renewable sources of electricity (for example solar, wind and hydropower) are already being widely deployed around the world, there is still a need to efficiently replace the petroleum used in the production of bio-based chemicals and biofuels. In this context, the production of biofuels from plants emerges as an attractive alternative and even though countries such as UK, USA and Brazil for decades have been successfully producing first generation biofuels (from wheat, corn and sugar cane, respectively) there is a need for an alternative feedstock in order to avoid competition with food. The most promising alternative is the production of second generation biofuels from lignocellulose, a renewable, cheap and abundant feedstock available worldwide [1, 2].

Lignocellulose biomass is the most abundant raw material on the planet, it is usually low-priced [3, 4] and comprises a range of potential feedstock including forest products (wood and softwood) to general wastes such as municipal wastes, industrial waste (paper, textile and clothing) and agricultural wastes (wheat straw, corn straw, sugar cane straw, bagasse and oil palm residues) [2, 5]. Although promising, the conversion of lignocellulose into biofuels is challenging due to the recalcitrant nature of lignocellulose (resistance of plant cell walls to deconstruction). While fermentation of corn starch or sucrose from sugarcane juice is a well-established and relatively easy technology, the conversion of lignocellulosic biomass involves the hydrolysis of polysaccharides into monosaccharides (a process called saccharification) prior to microbial fermentation. This process is difficult and complex because lignocellulose has evolved in nature to resist degradation, conferring protection to the plant against chemical and biological attacks, which hinders the access to its monosaccharides.

1.2 What makes lignocellulose biomass difficult to digest?

Lignocellulose biomass is mainly present in the secondary cell wall of plants and its formation happens after the primary cell wall is completed and cell expansion is finished [6]. Secondary cell walls are the thick layer present in plants and confer strength and resistance against degradation, stabilizing

the structure of the plant as a whole [7]. The main constituents of secondary plant cell wall are cellulose, hemicellulose and lignin (figure 1.1) with some minor amounts of pectin and structural proteins present in grasses [8], but the specific composition and three dimensional structures of these polymers varies according to the feedstock [4, 6].

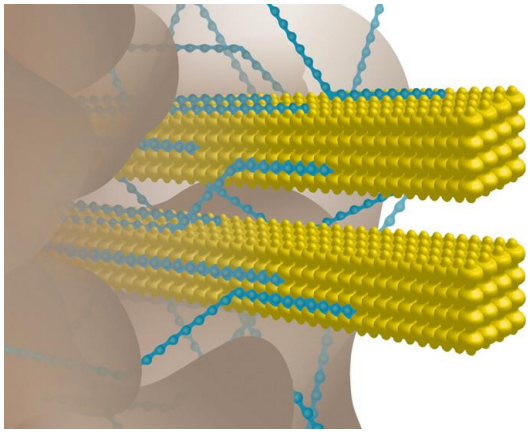


Figure 1.1 General representation of the plant secondary cell wall. In yellow is shown the cellulose microfibrils, mainly responsible for the plant cell wall structure; embedding cellulose is hemicellulose (in blue), which is a more complex polymer formed by different sugars and side chains that can interact with cellulose by hydrogen bonding and connect these polysaccharides to a more complex polymer, called lignin; Lignin (in brown) is an amorphous and heterogeneous phenolic polymer that surrounds both polysaccharides and offers protection to the secondary cell wall as a whole. Reproduced from Marriot *et al.*, 2016 [6].

1.2.1 Cellulose

Cellulose is the major component of lignocellulose and is the most abundant biomass on the planet [9]. Because it is a polymer of glucose, it represents a valuable renewable source of carbon to be used for the production of valuable chemicals and biofuels [10, 11]. Cellulose is composed of linear β-1,4 glucans with sequential glucose residues being rotated 180° to one another (figure 1.2a), and unlike other polymers of glucan the repeating unit in cellulose is the disaccharide cellobiose instead of the glucose [4, 6]. This configuration results in an extended and stable conformation for the molecule, giving rise to long and straight chains. Multiples of these glucan chains aligned side by side, form the cellulose microfibrils and because this structure lacks side chains the microfibrils interact with one another through several intra and intermolecular bonds, resulting in a crystalline structure that is highly insoluble and resistant to microorganisms and enzymatic attack [6, 12, 13]. Sometimes, glucan chains form a less organized region along the cellulose microfibrils (called amorphous region)

that tend to be more easily digested by enzymes [13] (figure 1.2b). These amorphous regions of cellulose are believed to be areas of links between hemicellulose and cellulose [14].

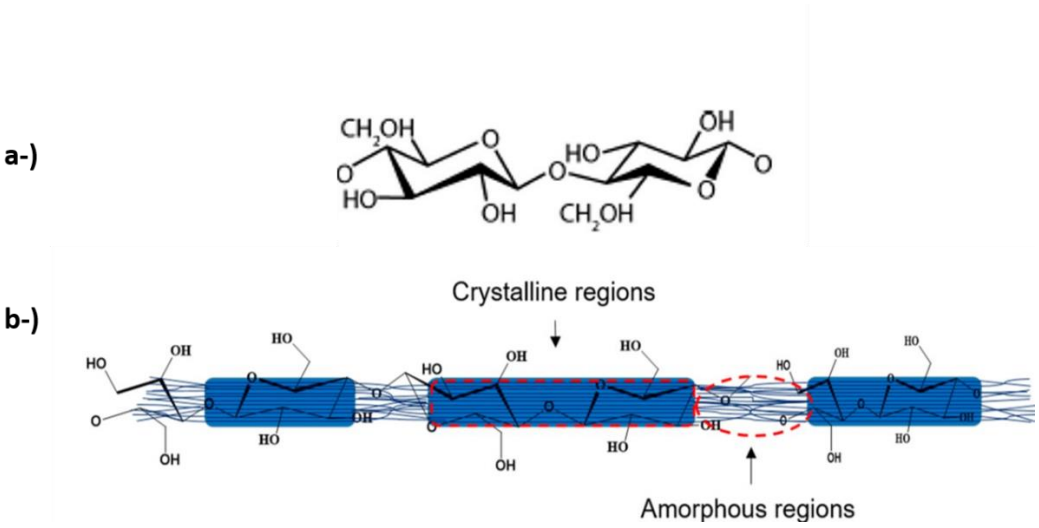


Figure 1.2 a) Cellobiose, the repeating structure of cellulose. **b)** General schematic representation of cellulose microfibril showing the amorphous and crystalline structure. Reproduced from Tayeb *et al*., 2018 [15].

1.2.2 Hemicellulose

Unlike cellulose, hemicellulose is a more complex heteropolysaccharide and its composition, types of glycosidic bonds, degree of polymerisation and side chains varies greatly according to the plant species [16]. Even though hemicellulose constitutes 15-35% of plant biomass and could be a great source of sugar for industrial purposes, because of its heterogeneity and high amounts of pentose sugars (not easily fermented by yeast), hemicellulose currently has few applications in industry [7]. Hemicellulose is usually formed by a β-1,4 linked backbone with an equatorial configuration and because it is highly substituted it does not form crystalline structures, but interacts with cellulose and lignin instead [6, 17]. The polysaccharides in the hemicellulose are typically named according to their backbone sugar and it includes mannans and glucomannans, xyloglucans, mixed linkage glucans (MLG) and xylans [16, 17]. In the next few paragraphs, each of these polysaccharides will be briefly discussed, with emphasis in their occurrence into grass cell walls.

Mannans and glucomannans are formed by a backbone of β-1,4 mannosyl residues or β-1,4 glucosyl-mannosyl residues. If the mannosyl residue is branched with a galactosyl residue, they are called galactomannans or galactoglucomannans (figure 1.3a and 1.3b). These types of polysaccharides only appear as minor amounts in the hemicellulose of grasses [17]. Xyloglucans (figure 1.3c) are formed by a backbone of β-1,4 glucosyl residues highly substituted with xylosyl residues groups. These xylosyl residues can be decorated with galactosyl and/or arabinosyl residues and the galactosyl residues can still be decorated with a fucosyl residue [6, 17]. Xyloglucans only represents minor amounts of hemicellulose of grasses, where the structure is usually less branched than in dicot plants [18]. MLG (figure 1.3d) are an unbranched polymer formed by a backbone of β-1,3-1,4-glucosyl residues and it is usually composed of 70% β-1,4 linked and 30% β-1,3 linked [6]. MLG are exclusive to grasses and is mainly present in primary cell walls of grasses, with minor amounts also present in the secondary cell wall [8]. MLG (as well as mannans and glucomannans) are particularly interesting from a fermentation point of view because they are formed by hexoses sugars, which are more easily fermented by yeast than the pentose sugars present in xylans and xyloglucans. Xylans are a diverse group of polysaccharides sharing the common feature of being formed by a backbone of β-1,4-xylosyl residues. This backbone can be substituted to several levels by arabinosyl residues, glucuronic and methylated glucuronic acid, or acetyl side chains. Arabinosyl residues can also be substituted with a xylosyl residue and/or ferulic acid. Xylans decorated with arabinosyl residues and glucuronic acids are called arabinoxylans (AX) and glucuronoarabinoxylans (GAX) (figure 1.3.e) and are the main constituent of grass cell walls [6, 13].

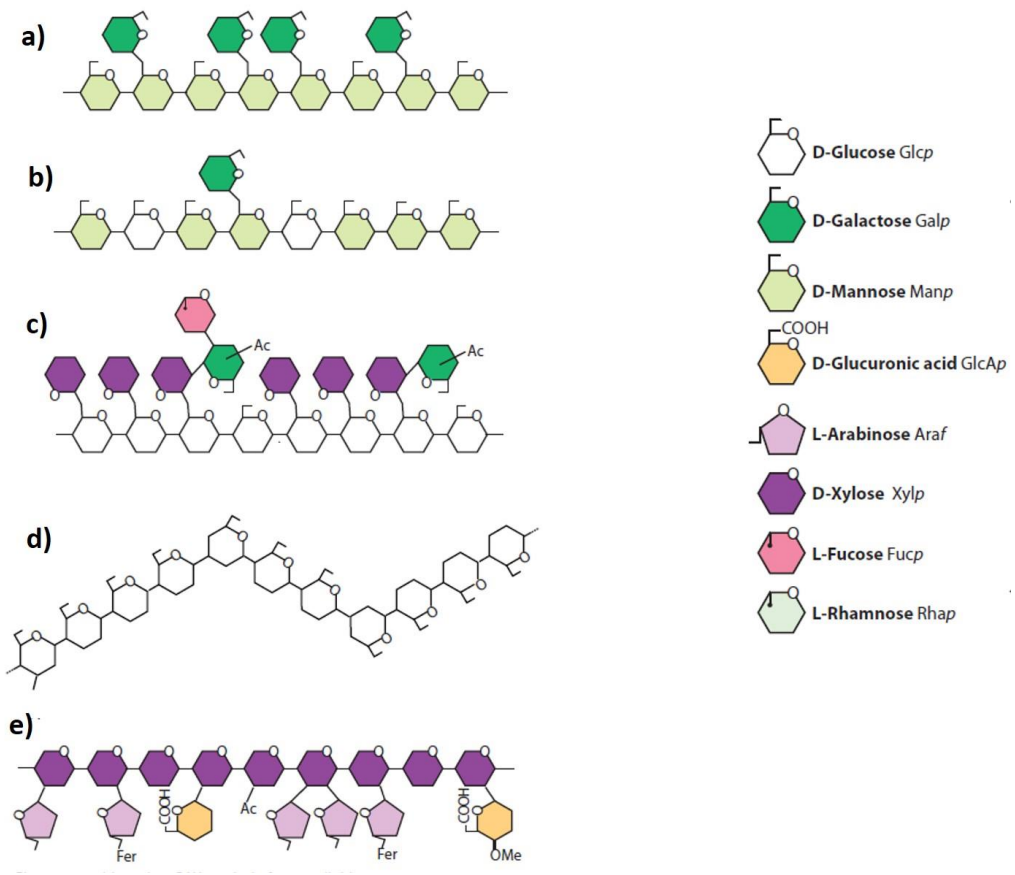


Figure 1.3 Schematic of different types of hemicellulose. a) galactomannan; b) galactoglucomannan; c) xyloglucan: in this representation, galactose residues are acetylated (Ac); d) mixed linkage glucan; e) glucuronoarabinoxylan: in this representation, the glucuronic residue can be methylated or not and the arabinose can be linked to a ferulic acid or not. Reproduced from Henrik *et al*., 2010 [17].

In xylans of grasses the amount of arabinosyl substitution can largely vary from 1:2 Ara:Xyl to 1:30. Moreover, these arabinosyl residues can be attached by ester linkages to ferulic acid (FA) and to a lesser extent, coumaric acid (pCA) [19]. The levels and pattern of these substitutions vary from species to species and directs how strongly they can interact with other polysaccharides, thus affecting the properties of the wall as a whole [20]. In fact, it is well accepted that grass cell walls are uniquely cross linked by FA [19, 21, 22]. Importantly, FA is not only ester linked to hemicellulose, it is also capable of oxidatively coupling to lignin or to another FA [22, 23] in the hemicellulose. Thus, through the formation of diferulates and through the esterification of arabinosyl residues, FA promotes the linking of one chain of hemicellulose to another and because FA and/or diferulates can covalently link to lignin by ether bonds, FA also connects hemicellulose to lignin (figure 1.4). This connection point between lignin and hemicellulose is also known as lignin-carbohydrate complex (LCC) and because

lignin is the most recalcitrant composite polymer in the cell wall, the degree of these cross linking is directly related to the digestibility of AX (or GAX) [6, 19, 24].

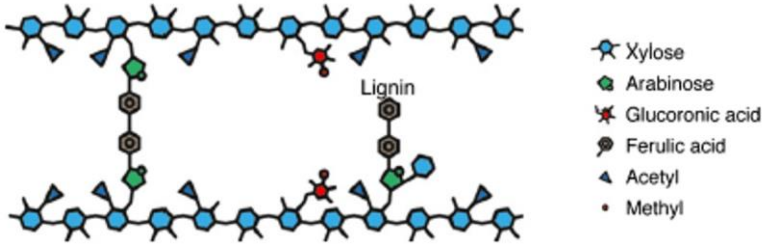


Figure 1.4 Schematic of feruloylation with the formation of diferulates. FA can link hemicellulose chains by the esterification of arabinose residue or can link hemicellulose to lignin by ether bonds. Reproduced from Marriott *et al*., 2016 [6].

GAX of grass cell walls can also be highly decorated with acetyl and glucuronic acid which can be methylated or non-methylated. The roles of these decorations is not completely known but according to a model proposed by Busse-Wicher *et al*., 2014 [25] the arrangement of acetyl and glucuronic acid (GlcA) turn the GAX structure into a helical conformation interacting with both hydrophilic and hydrophobic cellulose faces, suggesting the importance of acetylation and GlcA in the interaction with cellulose microfibrils (figure 1.5).

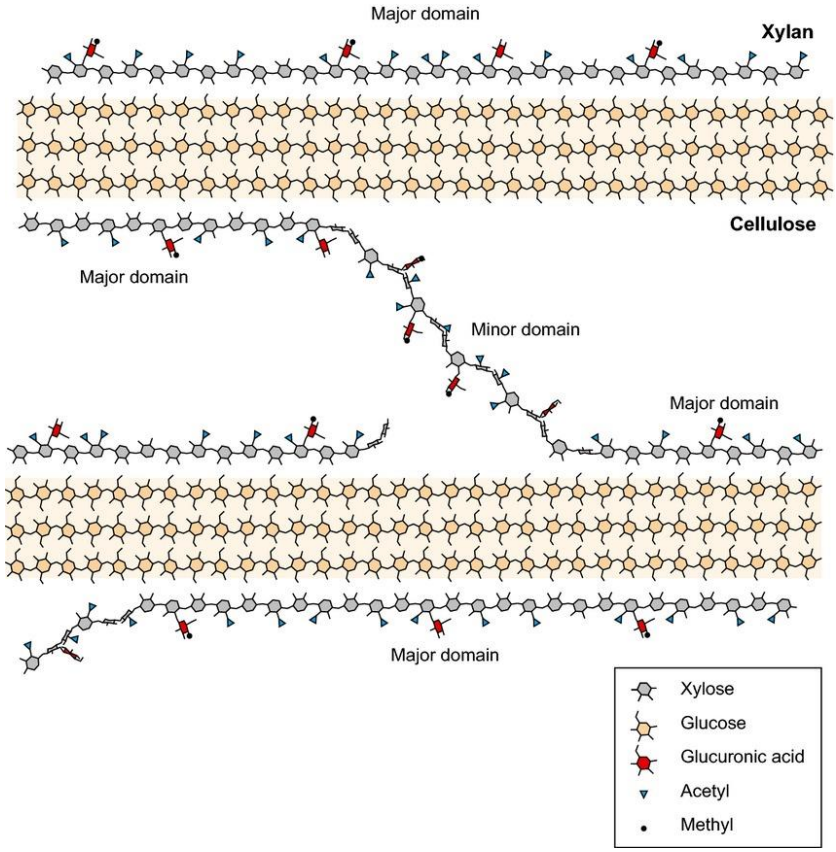


Figure 1.5 Schematic of the interactions of acetyl and glucuronic acid with cellulose microfibrils. Depending on the pattern of acylation and glucuronic acid, chains of hemicellulose will interact with hydrophilic (major domain) or hydrophobic faces (minor domain) of cellulose microfibrils. Reproduced from Busse-Wicher *et al.*, 2014 [25].

1.2.3 Lignin

In the secondary plant cell wall, cellulose and hemicellulose are embedded in a complex hydrophobic polymer called lignin. Lignin is an amorphous and heterogeneous phenolic polymer formed mainly from three basic units, p-hydroxyphenyls (H), guaiacyl (G) and syringyls (S) (originated from the p-coumaryl, coniferyl and sinapyl hydroxycinnamyl alcohols, respectively - figure 1.6) through a variety of ether and carbon-carbon linkages [26, 27]. Lignin structure is believed to have a random formation free of biological control [28] resulting in a polymer highly branched and amorphous. Moreover, because of its aromatic nature, lignin forms a hydrophobic coat surrounding the polysaccharides, which protects and confers high resistance to the plant towards degradation [29]. The lack of a repetitive pattern in lignin's structure is the reason why it is so difficult to find microorganisms and enzymes able to directly degrade lignin [6].

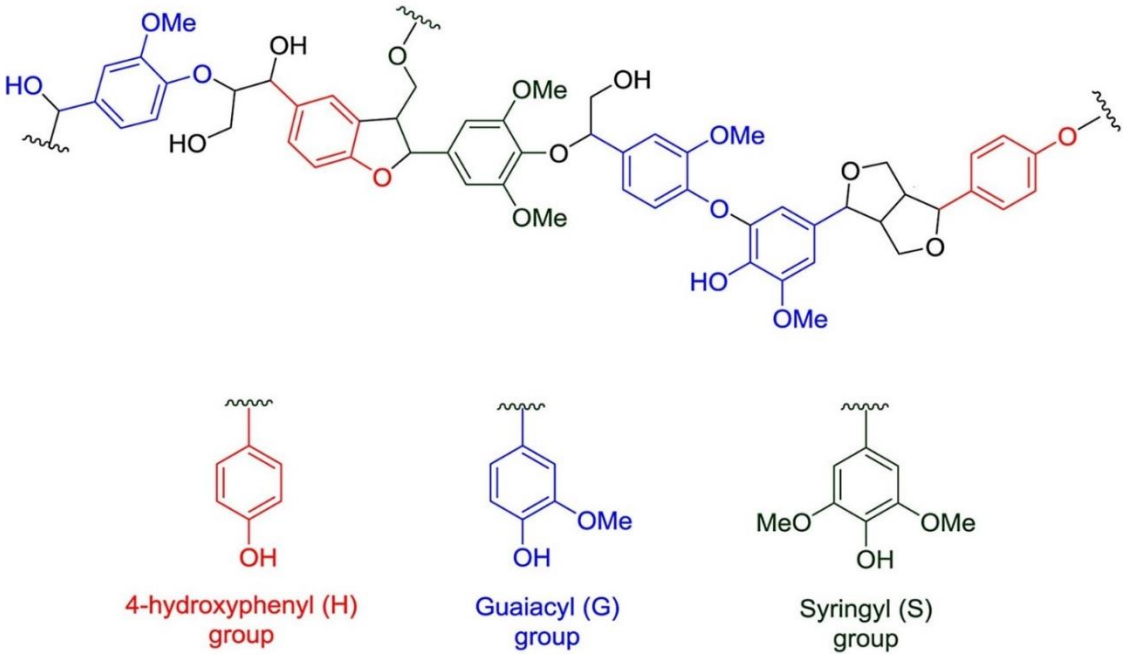


Figure 1.6 Representation of lignin structure. In red, blue and green are the three basic units of H, G and S units, respectively. Reproduced from de Gonzalo *et al*., 2016 [26].

1.2.4 Pectins

Pectins are very complex polysaccharides that typically contains high amounts of galacturonic acid (GalA) in their structure. The mains types of pectin (figure 1.7) in plants cell wall are homogalacturonan (HG), rhamnogalacturonan I (RG-I), rhamnogalacturonan II (RG-II) and xylogalacturonan (XGA) [30]. These types of pectins have a backbone of α-1,4 linked GalA residue that can be methylated or acetylated, and in RG-I the backbone is formed by alternation in GalA and rhamnose residues. Many pectins are highly decorated with different amounts of different sugars, but RG-I is mainly decorated with arabinosyl and galactosyl residues and XGA is mainly decorated with xylosyl residues [30-32].

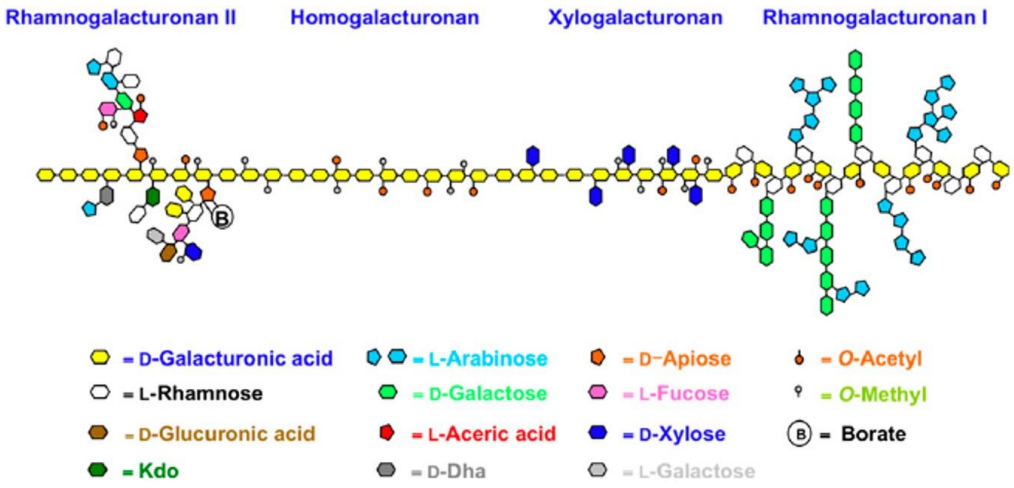


Figure 1.7 Schematic representation of 4 different types of pectin. The backbone of GalA with different patterns of methylation and acetylation (or alternated GalA and rhamnose for RG-I) and the different possibilities of decorations for each type of pectin is shown. Reproduced from Harholt *et al*., 2010 [30].

Despite pectins being present in grass secondary cell walls only in minor amounts [6, 8], there is evidence that it influences the efficiency of biomass saccharification [33, 34]. The roles of pectins in biomass digestibility is not yet completely elucidated; however, some studies suggest that pectin embeds cellulose and hemicellulose in a pectin matrix, which might block the access of lignocellulose-degrading enzymes to cellulose and hemicellulose [31, 35, 36]. Thus, even in small amounts, pectin could contribute to lignocellulose recalcitrance.

As described above, biomass recalcitrance is directly related to its chemical composition and physical spatial structure. The presence of lignin, pectin, hemicellulose and its decorations, and the crystallinity of cellulose, as well as the cross linking and interactions between each of these composites, act as a barrier and prevent the access of degradative enzymes to the polysaccharides. Because of this, even though lignocellulose is typically 75% composed of polysaccharides with potential to be converted into biofuels and bio-based products, the saccharification step remains the bottle-neck of the process [13]. Typically, harsh conditions and chemicals are usually employed to efficiently promote hydrolysis of lignocellulose, which might cause negative impacts in the environment [12]. In contrast, the use of lignocellulose-degrading enzymes during saccharification are typically associated with smaller environmental impacts once it can be conducted under mild

conditions of pH and temperature. Therefore, there is a growing interest in the discovery of novel and more efficient lignocellulose-degrading enzymes to be employed in the saccharification process [37].

1.3 CAZymes

Due the complexity of lignocellulose, several enzymes acting synergistically are needed to convert the polysaccharides of biomass into its monosaccharides. These enzymes are generally referred to as carbohydrate active enzymes (CAZymes). The CAZy database (http://www.cazy.org/) is a collection containing all the known enzymes up to date related to either carbohydrates assembly or carbohydrates breakdown [38] and these enzymes are classified in different families according to similarity of their amino acid sequences [39]. Enzymes related to carbohydrate assembly belong to the glycosyltransferases (GT) family and the ones related to the carbohydrates deconstruction are classified in four different groups: glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs) and auxiliary activities (AA) families. In addition, there are also the carbohydrate binding modules (CBMs), which do not exhibit catalytic activity and are grouped together [38].

Glycoside hydrolases (GHs) are the biggest group of CAZymes to date, they have different substrate specificity and according to the CAZy database are classified in 165 different families based on structure and activity. Because these enzymes are directly related to the hydrolysis and/or rearrangement of glycosidic bonds, most (if not all) cellulases and hemicellulases known to date, belong to this group of enzymes. Unlike GHs, Polysaccharide lyases (PL) are enzymes that cleave polysaccharides containing uronic acid through an elimination mechanism instead of hydrolytic cleave [40]. Because these enzymes are active against uronic acid-containing polysaccharides, they are typically associated with degradation of pectins and are currently divided into 37 different families. The third group, carbohydrate esterases (CEs) are a smaller group of enzymes currently divided into 16 different families and are characterized for hydrolysing ester linked substitutions from polysaccharides. These enzymes have different substrate specificity, but because they act on ester groups, they are typically responsible for removing acetyl and GalA groups from pectins and/or for removing acetyl, GalA and ferulic acid groups from side chains of hemicelluloses [41-43]. The fourth group, auxiliary activity (AA), are a group recently created in the CAZy database to accommodate enzymes involved in lignocellulose degradation through redox mechanisms [44]. AA are currently divided into 16 groups and accommodates families of enzymes related to lignin modification, such as lignin peroxidases and catalases to the lytic polysaccharide mono-oxygenases (LPMOs). Finally, the

carbohydrate binding modules (CBMs) are a group also covered by the CAZy database that includes associated domains without catalytic activity but with carbohydrate-binding activity. CBMs are defined as a contiguous amino acid sequence within the CAZymes that promote the association of the enzyme with its substrate [45]. They are also classified according to their sequence of amino acid and are currently divided into 85 different families.

1.3.1 Enzymatic cellulose degradation

Cellulases are the common name given to enzymes directly related to the degradation of cellulose. In nature, some microorganisms are able to produce a set of enzymes capable of promoting cellulose degradation, called multi enzymatic complex [46], which originated a classical model for enzymatic hydrolysis of cellulose. In this model, the main enzymes involved in the degradation of cellulose are endoglucanases, cellobiohydrolases (also known as exo-glucanases) and β-glucosidases. Endoglucanases act randomly on the cellulose microfibrils, especially on low crystallinity regions, mainly releasing oligosaccharides with free ends. Cellobiohydrolases cleaves the bonds on the free ends of cellulose microfibril and oligosaccharides releasing mainly cellobiose units. Lastly, β-glucosidases cleave short cello-oligosaccharides and cellobiose into its final monosaccharide, glucose [47, 48]. Recently, this classic model has been reviewed (figure 1.8) and polysaccharide lytic monooxygenases (LPMOs) has been included. These enzymes oxidatively cleave internal glycosidic bonds from the crystalline cellulose microfibrils, enhancing the action of cellobiohydrolases [49, 50].

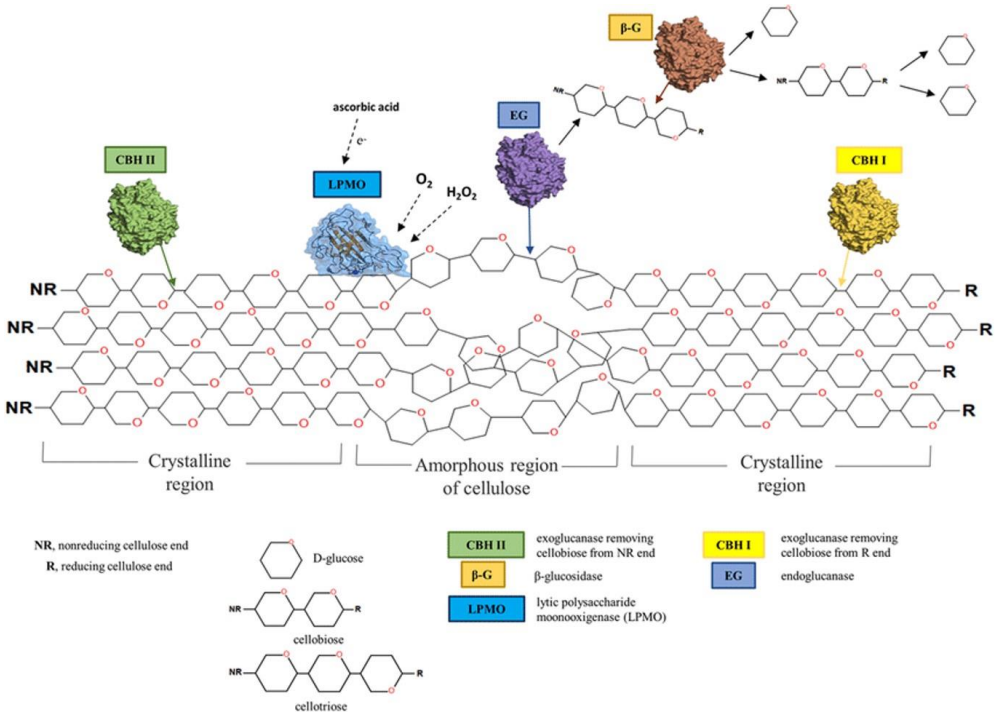


Figure 1.8 Schematic representation of degradation of cellulose by the synergistically action of endoglucanases, cellobiohydrolases, LPMOs and β-glucosidases. Reproduced from Andlar *et al.*, 2018 [51].

1.3.1.1 β-glucosidases

Beta-glucosidases are a highly heterogeneous group of enzymes that can be found in many organisms such as bacteria [52], fungi [53], plants [54] and animals [55], and among others activities, they are responsible for the hydrolysis of cellulose oligosaccharides, such as cellobiose and cellotriose, into glucose [56]. All β-glucosidases belong to the GH super family and they are mainly grouped into families 1 and 3 [57], but some representatives have also been found in families 5, 9, 6 and 30, for example [48]. The β-glucosidases belonging to GH1 family are generally from archaebacteria, mammals and plants, whereas β-glucosidases belonging to GH3 family are mainly β-glucosidases from bacteria, yeast and fungi [48].

In the final step of saccharification, cellobiose and other short cello-oligosaccharides (cellotriose, for example) are hydrolysed by β-glucosidases to yield glucose. This is an important step of the entire cellulolytic process, as endoglucanases and cellobiohydrolases are inhibited by cellobiose and short cello-oligosaccharides [48, 58]. By preventing the accumulation of inhibitory levels of cellobiose and short cello-oligosaccharides, β-glucosidases play a crucial role in the whole process of saccharification. However, since β-glucosidases are often sensitive to the presence of glucose [57], which is the main product of their catalysis, their application in commercial scale are restricted.

Although some examples of glucose-tolerant or glucose-stimulant β-glucosidases (mostly from fungi and members of family GH1 or GH3) have already been reported [52, 59, 60] their mechanisms and reasons for such a feature is yet not completely known and are currently focus of investigation [61-63]. For this reason, it is important to conduct searches for novel β-glucosidases in both ambits: to provide new information that could help to understand and elucidate their diversity and properties, as well as the identification of novel glucose-tolerant/stimulant β-glucosidases to be included into enzyme cocktails for biomass hydrolysis.

1.3.2 Enzymatic hemicellulose degradation

For decades only cellulose hydrolysis has been the focus of researchers' attention. However, recent studies have shown that enzymatic cocktails containing hemicellulases and other accessory enzymes like carbohydrate esterases, in addition to cellulases, are more efficient for the saccharification of lignocellulosic biomass that has been mildly pre-treated, resulting in higher yields of fermentable sugars with lower amounts of enzyme [64, 65]. As discussed in previous sections (1.2.1 to 1.2.4), the access of cellulases to the cellulose microfibrils is restricted due the barrier provided by the hemicellulose, lignin and, to a lesser extent, pectin. In this section the focus is on the degradation of hemicellulose to gain an understanding of how some enzymes involved in hemicellulose degradation can help in the overall saccharification process.

Xylans are an abundant class of hemicellulose in grasses and as it was mentioned before, in grasses it presents different levels of decorations. Consequently, due to its chemical and structural heterogeneity, several types of bonds, and the presence of different monomeric units, the efficient hydrolysis of xylan requires a complex enzymatic system. This system of enzymes acting on xylan is known as the xylanolytic complex and it involves synergistic action of enzymes on the main backbone of xylan as well as on xylan side chains [66]. Figure 1.9 shows a schematic for the arabinoxylan structure and the sites of action for its xylanolytic enzymes.

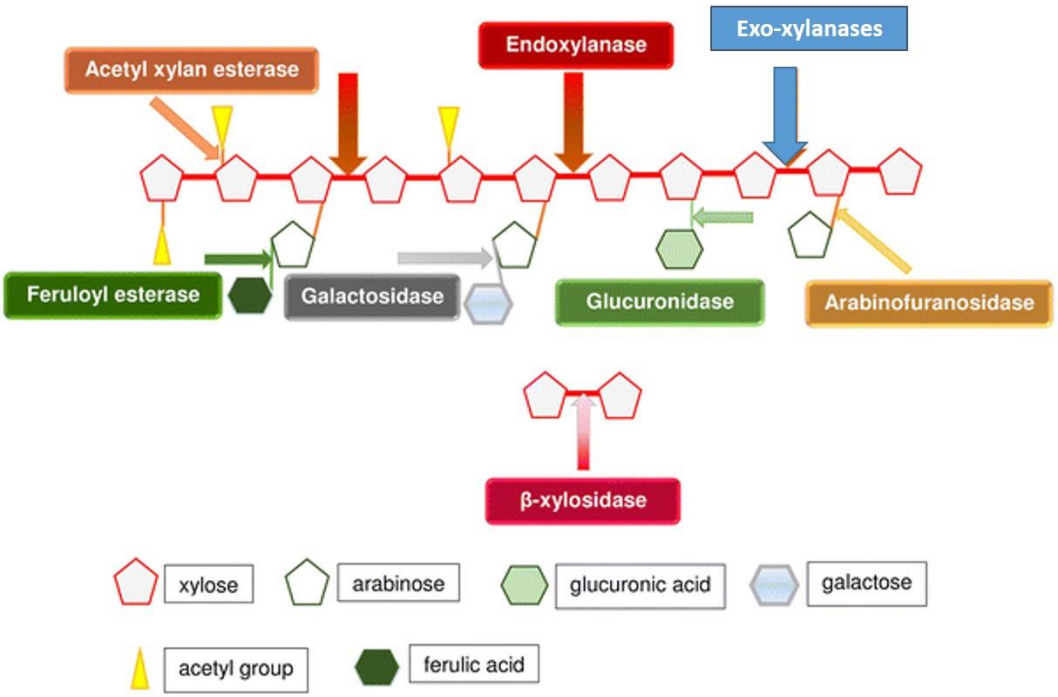


Figure 1.9 Schematic representation of arabinoxylan (mainly constituent of hemicellulose of grasses) and the xylanolytic enzymes involved in its degradation. Reproduced from Gündüz *et al*., 2016 [67].

The main enzymes of the xylanolytic complex are: endoxylanase, which hydrolyse internal glycosidic bonds of the main xylan chain releasing xylo-oligosaccharides and xylobiose [66, 68]; and β-xylosidases, which acts on xylo-oligosaccharides and xylobiose, releasing xylose [69]. The debranching is catalysed by a range of different accessory enzymes, as arabinofuranosidases, which remove arabinosyl residues; α-glucuronidases, which release glucuronic or methyl-glucuronic acid; α-galactosidases, which remove galactosyl residues; acetyl xylan esterases, which remove acetyl groups; and feruloyl esterases, which remove ferulic (and to a lesser extent, p-coumaric) acids [17, 23, 37, 70]. Among these enzymes acetyl xylan esterases and feruloyl esterases belong to the CE super family because they act on the ester bonds that connects their specific residues to the arabinoxylan, while all the remaining enzymes belongs to the GH super family, because they act on glycosidic bonds.

Endoxylanases and β-xylosidases act synergistically on xylan hydrolysis and because a large proportion of known endoxylanases are inhibited by its product (xylobiose and xylo-oligosaccharides) [71], in addition to catalysing the final hydrolysis step, β-xylosidases play an important role by relieving the inhibition of endoxylanases, enabling greater efficiency of the process as a whole. Moreover, due to the branched nature of xylans, a strong synergism is also observed and needed between endoxylanases and some accessory enzymes. Decorations of arabinosyl residues, for example, might

hinder the action of endoxylanases on the main backbone chain and the removal of these residues by arabinofuranosidases can enhance the action of endoxylanases [24].

1.3.2.1 Arabinofuranosidases

Arabinofuranosidases (AFases) are enzymes that act releasing arabinosyl residues from the non-reducing end of polysaccharides (such as arabinoxylan) and others arabino-oligossacharides [72]. They are accessory enzymes of the xylanolytic complex that by removing arabinosyl residues from the side chains of xylans, aids the action of endoxylanases. AFases are mainly found in bacteria [73] and fungi [71], but some members from plants have also been reported [74]. They belong to the GH super family and are mainly present in families 43, 51 and 62, but also have representatives in families 2, 3 and 54 [24, 75]. Although always responsible for the removal of arabinosyl residues, AFases are complex and varied with respect to substrate preferences. In the family 43, has been reported AFases specialized for removing arabinosyl residues from mono substituted xylan [76] but also AFases specialized in the removal of residues from di substituted xylan [77, 78]. It is also in the family 43 that AFases with bifunctional arabinofuranosidase/β-xylosidase activities have been reported [24] (although some bifunctionality has also been reported for members of family 51 [75]). Family 62 is exclusively formed by AFases and members of this family are typically specialized in the release of arabinosyl residues from mono-substituted xylans with some activity in arabinans (but not in de-branched arabinan) [24, 79]. Family 51 contain the largest number of studied AFases and most of them are from bacterial origin [75]. Members of this family are reported as having a wide substrate specificity, being able to remove arabinosyl residues from mono and/or di substituted xylan, from arabinans and from arabino-oligosaccharides [24, 72, 80].

AFases have gained some attention in past years as they were reported to have positive effects on the hydrolysis of pre-treated biomass [81]. However, later studies [82] have shown that wheat arabinoxylans treated with AFases have enhanced inhibition of cellobiohydrolases, suggesting that addition of AFases to enzymatic cocktails could lead to a decrease in saccharification. In fact, most of the AFase studies to date have focused in their ability to debranch arabinoxylans (and/or arabinans) and on their synergistic action with other xylanolytic (or pectinolytic) enzymes but there is a need for more studies and accurate information in the real roles of AFases in the digestibility of biomass as a whole.

1.3.2.2 Feruloyl esterases

As previously mentioned, in secondary cell wall of grasses, ferulic acid (FA) plays an important role promoting the cross linking between arabinosyl residues from different xylan chains, as well as between arabinosyl residues and lignin [22, 23]. Feruloyl esterase (FAE) is the name given to the class of enzymes that are able to remove FA and cross linking polysaccharides by the cleavage of the ester bonds connecting arabinosyl residues and FA [83]. These enzymes belong to the CE1 family and can be found mostly in fungi but also in bacteria [84]. Due to their action breaking the bonds between polysaccharides and phenolic compounds, it is believed that FAEs reduce biomass recalcitrance by facilitating the access of GHs to the polysaccharides [42]. Synergistic action between FAEs and cellulases, xylanase and pectinases has been reported [83, 85] and its effect in improving biomass saccharification has also been described [86, 87]. Besides their application in the process of saccharification, FAEs (and the products that it releases, as FA and other phenolic compounds) are of great interest for diverse biotechnological applications, such as food, cosmetic, pulp and paper, and pharmaceutical industries [83, 88]. Therefore, discovery of novel FAEs with different applications and features are not only important to understand different patterns of cross linking in secondary plant cell wall, but it is also of considerable interest for biotechnological applications [23, 83, 88].

1.4 Challenges for the use of seawater in biorefineries

The use of lignocellulosic biomass in biorefineries, as well as its conversion into bio-based chemicals are very attractive and promising from an environmentally friendly and sustainable point of view. Lignocellulose is abundant in nature, does not affect food security nor occupy land destined for food production, has less emissions of greenhouse gases to the atmosphere when compared to the combustion of fossil fuels and is renewable. However, due to the structural complexity of lignocellulose, biomass conversion on an industrial scale is not yet economically viable. While the sugars in corn starch and/or sugar cane juice are easily accessible and fermented by bacteria or yeast for the production of first generation biofuels, the digestion of cellulose to produce glucose for fermentation is still challenging. Due to the complexity of lignocellulose, to overcome the barriers offered by lignin, hemicellulose and pectins, a step of pretreatment (chemical or biological) is needed to expose cellulose, which only then can be saccharified to render the final fermentable sugars. These two extra steps demand more financial investments and time, which makes the final product also more expensive.

Another important aspect to be considered during the degradation of biomass and its conversion into biofuels is the large amounts of fresh water (1.9-5.9 m^3 water per m^3 of biofuel [89]) used in this process. This is a major concern as fresh water is a valuable and scarce resource. Large parts of the world are currently experiencing water stresses and this is expected to be worse with climate change and increasing in population [90]. Thus, considering the existing shortage of fresh water in some places in the world, the heavy use of fresh water by those industries could become unsustainable in the future [91]. As a result, the possibility of using non-potable water resources, especially seawater, in steps of pretreatment and saccharification has been gaining interest, which could save the fresh water otherwise used in these processes, for agriculture and public consumption.

So far, the use of seawater in the pretreatment of biomass has been poorly explored, but studies investigating the effects of the addition of salts in the pretreatment of different biomass have been reported [92, 93]. Also, more recent studies reported by Fang *et al*., 2015 show that leaflets of date palm pretreated with seawater resulted in lower cellulose crystallinity than leaflets pretreated with fresh water and that no significant differences were observed for the ethanol yield of liquids obtained from both conditions of pretreatment [89]. These results are encouraging and show that the replacement of fresh water by seawater in pretreatment steps could be feasible.

Regarding saccharification, it is well known that different ions, even at low concentrations can affect the activity (inhibit or stimulate) of many CAZymes [53, 94, 95]. The biggest issue though, is regarding salt concentrations. In the presence of high salt concentrations, most enzymes have low or no catalytic activity, which has been attributed to the effect of ions on the structure and dynamics of the water [96]. It is well known that the biological function and structure of a protein is critically affected by its surface interactions with molecules of water, which tend to interact with polar groups in the surface of the protein. At the same time, water molecules tend to form organised cages of water molecules joined by hydrogen bonds surrounding hydrophobic regions of the proteins [97]. At high saline concentrations, ions sequester molecules of water, which limits the availability of free molecules for protein hydration. In addition, these ions also disturb the organized local structures of water molecules by disrupting intermolecular hydrogen bonds, and disrupting electrostatic interactions between side chains of charged amino acid residues [96]. Taken together, these effects interfere with the structure and function of proteins, their solubility, stability, and ability to interact with other molecules, including other proteins or interaction between subunits of the same protein [89, 96]. In addition to concentration, the nature of the ions in solution is also important with regard to the destabilizing effects on protein structure and function. In general, the destabilizing effect of an ion can be predicted by its position in the Hofmeister series [98], which describes the ability of ions to

salt in (when protein-ions interactions prevent protein-protein interactions, increasing solubility) or to salt out (when high concentration of ions lead to osmotic dehydration, facilitating protein-protein interaction, causing precipitation). Even though salt concentrations of seawater vary widely according to geographic location, its composition is well known (Table 1.1) and according to the Hofmeister series, most of the ions present in the seawater (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Br^{-}, Cl^{-}) may have destabilizing effects on different enzymes and may thus negatively influence enzymatic hydrolysis of lignocellulosic biomass [89, 96].

Table 1.1 Average composition of seawater. Reproduced from de Maria *et al*., 2013 [91].

Component	**Composition in seawater (g/L)**
NaCl	27,133
$MgCl_2$	2,504
$MgSO_4$	3,382
$CaCl_2$	1,17
KCl	0,74
$NaHCO_3$	0,21
NaBr	0,08
Total salts	35,22
Remnant water	964,78

Due to all these effects, the replacement of fresh water with seawater during the saccharification steps would not be possible using the current enzymes employed in the process. Although techniques of molecular biology could be used aiming to improve the salt tolerance of these enzymes, another approach would be to identify and to use salt tolerant lignocellulose-degrading enzymes already presents in the world, as they have evolved in nature to survive under these conditions.

1.5 Saltmarshes are source of salt tolerant enzymes

Saltmarshes are unique ecosystems that are located between land and the ocean and are characterized by being repeatedly flooded by seawater. Saltmarshes are highly productive ecosystems recognized for their importance regarding to nature conservation, sea and coastal protection, nursery areas for marine species, nesting for wild birds, among others. Saltmarsh formation happens on the coast, where the deposition of sediment brought by the seawater is stabilized by salt-tolerant terrestrial vegetation. As soon as vegetation becomes established, the growth of the saltmarsh is made possible by the accumulation of sediment and organic matter, that now are trapped by a bigger layer of material and by roots present underneath the surface [99, 100]. Because of its localization, saltmarshes are inhabited by a range of organisms and microorganisms with both, marine and terrestrial origins, and as in any other intertidal habitats they are exposed to physical stress, as such as flooding, salinity and climate changes, for example [99]. Therefore, Ecologists and Biologists have studied saltmarshes for a while in order to understand how these stressful conditions affect the interactions of these organisms, and more recently work has been done to explore and investigate the microbial diversity in this environment [101-104]. However, from a biotechnological point of view, these environments have not been well explored. As a result of their location, saltmarshes are dominated by salt-tolerant land plants and the lignocellulosic material from these plants forms the major source of biomass to feed the large communities of heterotrophic organisms living in these environments. Thus, there are likely to be a range of salt tolerant marine microbial species specialised for the degradation of the lignocellulosic biomass found there, and potential for novel species and enzymes. In this work, sediment from saltmarsh will be used as a source of microbial diversity in the attempt to find salt-tolerant lignocellulose-degrading enzymes.

1.6 Aims of this project

As described above, although promising, the use of lignocellulose to produce bio-based chemicals and biofuels is not yet feasible due the recalcitrant nature of plant biomass. In nature, plants have evolved to resist microbial degradation and enzymatic attack, resulting in a complex structure of the plant cell wall. Thus, in order to access the sugars present in lignocellulose, microorganisms and the enzymes produced by them need to degrade and/or modify lignin, pectin, hemicellulose and overcome the crystallinity of the cellulose. In this project, the objective was to find enzymes that degrade the most recalcitrant portions of lignocellulose. To do this, highly recalcitrant biomass (that has been previously degraded for 10 weeks) was used as the only source of carbon to enrich a community of microbes that originate from a lignocellulose-rich intertidal saltmarsh (Welwick,

Humber, UK). To interrogate the degradation process, a combination of meta-genomics, meta-transcriptomics and meta-proteomics were employed to identify potentially interesting enzymes for further study. Selected target enzymes were cloned, expressed and characterised for their enzymatic activity.

Chapter 2 Materials and Methods

2.1 Chemicals and reagents

The reagents and kits used in this work, if not stated otherwise, were obtained from Agilent technologies, Cambio, Cambridge Biosciences, GE Healthcare, Illumina, Merck, New England BioLabs, Promega, Qiagen, Sigma-Aldrich and Thermo Fisher Scientific.

Deionised water dH_2O was the main solvent used in this project and unless stated, it was obtained using an Elga PureLab Ultra water polisher under resistivity of 18 MΩ/cm.

Artificial seawater was prepared by dissolving 34 g of sea salt mixture (SeaChem) per 1 L of H_2O. The solution was heated to aid in the complete dissolution of the salts and allowed to cool at room temperature before its use.

2.2 Production of recalcitrant biomass

2.2.1 Initial recalcitrant biomass production

The initial recalcitrant biomass resulted from incubation of 35 g of *Spartina anglica* biomass (28 g > 1.12 mm and 7 g < 1.12 mm > 500 µm) in 700 mL of seawater (10 mM NH_4Cl) with 7 g of saltmarsh sediment collected in the Humber estuary as inoculum. It was retrieved after 10 weeks of incubation at 30 °C and 180 rpm in shake flasks by 5 consecutive washes with water through a 200 µm nylon mesh, followed by one wash with 1% SDS (at 60 °C for 15 minutes with agitation) and 5 more washes with dH_2O to remove the SDS. This biomass was then freeze-dried and used as the initial recalcitrant biomass in this project.

2.2.2 Final recalcitrant biomass and weight loss

A new experiment was set up following the same methodology as mentioned in section 2.2.1 using the initial recalcitrant biomass as the only source of carbon instead of the *Spartina* grass and fresh saltmarsh sediment as inoculum. In total, six shake flasks were set up containing the saltmarsh inoculum and one was used as a blank control without inoculum added. These flasks were incubated for another 8 weeks at 30 °C and 180 rpm, after which three of these shake flasks were used to estimate biomass degradation expressed as the weight loss by comparing the remaining recalcitrant

biomass with the initial recalcitrant input material (at time zero). The other three flasks were used to perform the meta-omics analyses (section 2.4).

2.3 Biomass composition analysis

To investigate how the composition of lignocellulose varied from the initial recalcitrant biomass to the final recalcitrant biomass, assays to analyse lignin, hemicellulose and crystalline cellulose content were performed. The details for each of these analyses are shown in the next subsections (2.3.1, 2.3.2 and 2.3.3) and these experiments were performed using 5 replicas.

2.3.1 Lignin content

Lignin content was measured using the acetyl bromide soluble lignin (ABSL) method [105]. For this, 250 µl of freshly prepared 25% acetyl bromide solution (25% v/v acetyl bromide in glacial acetic acid) was added to flasks containing 5 mg of finely ground biomass in order to break phenolic bonds and solubilise lignin. The flasks were heated at 50 °C for 2 h followed for an additional 1 h of incubation at same temperature (with agitation every 15 minutes). After cooling, the remaining liquid was transferred to volumetric flasks and mixed with 1 mL of 2 M NaOH and 175 µl of 0.5 M hydroxylamine HCl. The sample was then diluted 1:10 with glacial acetic acid and through the absorbance measured at 280 nm, the amount of lignin was determined as percentage of ABSL using the following equation:

% ABSL = [absorbance/(coefficient x path length)] x [(total volume x 100 %)/biomass weight)] x dilution, where the coefficient used was 17.75 (for grasses).

2.3.2 Hemicellulose content

Hemicellulose content was analysed using the trifluoroacetic acid (TFA) method [106]. Five mg of finely ground biomass was hydrolysed with 500 µl of 2 M TFA. The mixture was heated at 100 °C for 4 hours, mixing a few times during this process, separated into TFA-insoluble pellet and TFA-hydrolysate. The TFA was evaporated from the hydrolysate in a speed vacuum concentrator (SPD131DDA, Thermo Scientific) at 55 °C for two hours. The dried TFA-hydrolysate was washed twice with 500 µL of isopropanol, dried, resuspended in 200 µL of dH_2O. The supernatant containing the TFA-soluble sugars was filtered through 0.45 µm filters and submitted to analyses of the monosaccharides by High-Performance Anion-Exchange Chromatography (HPAEC). A mixture of nine monosaccharides (arabinose, fucose, galactose, galacturonic acid, glucose, glucuronic acid, mannose,

rhamnose and xylose, each at 100 µM) prepared in 3 different concentrations served as standards, which were treated according to the same procedure described above. Quantification was performed using the Chromeleon software package (version 6.80 SR16 Build 5387, Thermo). The TFA-insoluble pellet was stored and later used for analysis of the crystalline cellulose content (section 2.3.3).

2.3.3 Crystalline cellulose content

Crystalline cellulose content was determined using the anthrone-sulfuric acid method [106]. The TFA-insoluble pellet from the previous step (section 2.3.2) was washed once with dH_2O followed by 3 additional washes with acetone and left to dry on bench overnight. The next day, 70 µL of 72% (w/w) sulfuric acid was added to the sample and it was incubate for 4 hours at 25 °C in the heating block. After this time, 1890 µL of dH_2O was added to dilute the sulfuric acid to 3.2% and samples were again incubated for 4 hours at 120 °C in the heating block. Samples were allowed to cool and centrifuged for 5 minutes at 10000 rpm. The glucose content of the supernatant was determined using the colorimetric anthrone assay [107] against a glucose standard curve. For this purpose, 40 µl of samples were mixed with 360 µl of dH_20 and 800 µl of anthrone reagent. Samples and glucose standards were incubated at 80 °C for 30 minutes, transferred to optical plate and the amount of glucose present in the samples is determined by comparing the absorbance at 620 nm with the standard curve.

2.4 Meta-"omics" approaches

2.4.1 Combined genomic DNA (gDNA) and total RNA extraction

Reagents and materials preparations

- All water used in this extraction was diethyl pyrocarbonate (DEPC)-treated for 2 hours at 37 °C and autoclaved at 121 °C for 15 min.
- The beads used in this methodology were previous prepared incubating 0.5 g of 0.5 mm glass beads (Sigma G9268) and 0.5 g of 0.1 mm glass beads (Sigma G8893) in 2 mL cap tubes, with 1 mL of concentrated HCl for 1 hour with agitation. The beads were then washed with enough DEPC-treated water for complete removal of HCl (pH near to neutral) and finally autoclaved in 1 mL of DEPC-treated water.
- Phosphate buffered saline (1X PBS) was prepared by mixing 137 mM of NaCl, 2.7 mM of KCl, 8 mM of Na_2HPO_4, and 2 mM of KH_2PO_4 and adjusting the final solution to pH 7.4, unless otherwise stated.

Genomic DNA (gDNA) and total RNA were extracted simultaneously from microbial communities by the bead beating method. For the extraction, around 35 mL of the mix containing biomass and supernatant from the flasks used for the production of the final recalcitrant biomass (section 2.2.2) were transferred to a 50 mL falcon tube, filled to 50 mL with 1X PBS pH 8.15 and centrifuged for 30 min at 4500 rpm. The supernatant was discarded and the pellet was washed twice with 1X PBS as mentioned before. The residual pellet was mixed and 0.5 g was transferred to the pre-prepared 2 mL RNAse treated glass bead tube (the residual water was previously removed), followed by the addition of 0.4 mL cetyl trimethylammonium bromide (CTAB) extraction buffer, containing 1 µL/mL of β-mercaptoethanol (freshly added). After mixing in vortex, 0.3 mL of phenol/chloroform/isoamyl alcohol (25:24:1) pH 8, was added to the tubes and the biomass was homogenised using the Qiagen Tissue Lyzer for 2 cycles of 1.5 min at 30/sec frequency. The tubes were centrifuged at 13000 rpm, 4 °C for 15 min and the supernatant containing the DNA/RNA mixture was transferred to a 2 mL Eppendorf tube, extracted with equal volume of Chloroform/isoamyl alcohol (24:1) and again centrifuged in the same conditions mentioned before. The aqueous phase (free of phenol) was transferred to new 1.5 mL Eppendorf tubes and two volumes of PEG precipitation solution (PEG8000 Sigma) were added in order to precipitate the mix DNA/RNA content. The tubes were mixed by gentle inversion and left for total precipitation on ice, at 4 °C for 6 hours. The mixture of DNA/RNA was collected by centrifugation at 13000 rpm, 4 °C for 30 minutes. The pellet was washed twice with 1 mL of 75% ice-cold DEPC ethanol and allowed to dry for 10-15 min for complete removal of ethanol, before resuspension in 30 µL of DEPC-treated water. An aliquot of the extracted gDNA and total RNA mix was applied to agarose gel electrophoresis, and upon confirmation of successful extraction, the mixture was stored at -80 °C for further experiments (Chapters 4 and 5).

2.4.2 DNA preparation for meta-genomics and DNA sequencing

All the experiments described under this topic (section 2.4.2 and subsections) were performed under supervision of Susan Heywood at the Biorenewables Development Centre (BDC) at the University of York. Experiments were performed for only one time point (RNA/DNA extracted in section 2.4.1) and were performed in triplicate.

2.4.2.1 RNase treatment, DNA cleaning and concentration

The samples containing RNA/DNA previously extracted from microbial communities associated with biomass degradation (section 2.4.1) were incubated in a heating block at 37 °C with

10 mg/µL of RNaseA for 30 min in order to eliminate the RNA present. The remaining gDNA was purified and concentrated using Genomic DNA Clean and Concentrator 25 from Zymo Research, following the manufacturer's instructions. In details, 22 µL of gDNA and 44 µL of DNA binding solution were mixed, the mixture was transferred to the column supplied and centrifuged for 30 s at 1200 rpm. The flow through was discarded and samples were continuously added to the same column in order to concentrate the DNA. Next, the mixture was washed twice by the addition of 400 µL of DNA wash buffer in the column, which was centrifuged for 1 min at 12000 rpm. After the second wash, the column was transferred to new tubes, 25 µL of nuclease-free water was added and the column left for 3-5 min at room temperature. The column was centrifuged for 1 min and the eluate collected. One extra addition of 25 µL of nuclease-free water was performed and the second eluate was again collected (in the same tube). The DNA was quantified measuring absorbance at 280 nm using NanoDrop and stored at -20 °C.

2.4.2.2 PCR amplification of 16S rRNA.

PCR reactions of the 16S rRNA were performed on the gDNA from the biomass degradation cultures in order to analyse the bacterial community present in these samples. The primers used for these reactions were kindly provided by Dr. Daniel Leadbeater (a colleague in the lab who has previously performed these analyses using saltmarsh environment samples) and they target the very established V4 region of bacterial genomes. Primer 515f-Y was chosen because it supports detection of *Crenrchaea* and *Thaumarchaeota* [108], which are both considered abundant constituents of the archaeal saltmarsh profile [109, 110] and primer 806R was selected because it allows detection of SAR-11 clade in marine samples [111]. Additionally, a random dodecamer sequence NNNHNNNWNNN (5'-3') was added to the forward primer aiming to increase Illumina cluster ID accuracy [112]. In Illumina sequencing, a cluster is a clonal group of library fragments on a flow cell.. Thus the strategy of using a dodecamer was adopted because amplicon libraries typically have low diversity (since the same region is amplified), which can be problematic for the cluster identification by Illumina as it usually uses the first 12 base pairs to determine the cluster. The primers used for PCR amplification are listed below (table 2.1), where the dodecamer sequence is shown in red and the Illumina Nextera adaptor (added to both, forward and reverse primers) is shown in blue.

Table 2.1 Primers used for the PCR amplification of 16S rRNA. These primers were chosen according to the literature and target the V4 region of the ribosome.

Primer	Sequence
515f-Y	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGANNNHNNNWNNNHGTGYCAGCMGCCGCGGTAA
806R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGACTACNVGGGTWTCTAAT

PCR reactions were performed using 2 ng/µL of gDNA, 0.3 µL of enzyme (Phusion High-Fidelity DNA Polymerase from ThermoFisher Scientific), 5 µL of HF buffer, 1.25 µL of each primer (100 µM), 0.5 µL of dNTP mix (10 mM) and dH_2O enough to complete 25 µL total. The conditions for the PCR reaction are shown in table 2.2.

Table 2.2 PCR conditions for 16S rRNA amplification using 515f-Y and 806R primers.

Step	**Temperature (°C)**	**Time**	**Repeat**
Initial denaturation	98	30s	
Denaturation	98	10s	
Annealing	53	30s	x28
Extension	72	15s	
Final extension	72	10 min	
Hold	4	Forever	

PCR products were separated in agarose gels and after confirmation of expected amplicons, they were cleaned using AMPure XP beads purification (Agilent Genomics) and analysed using a TapeStation (Agilent Genomics) (section 2.4.2.3.), which provided their specific sizes.

2.4.2.3 Amplicon cleaning and purification by AMPure XP beads and Tapestation analysis

Amplicon cleaning and purification was performed by AMPure XP beads (Agilent Genomics) following the manufacturer's instructions. Magnetic beads were defrosted, allowed to reach room temperature and homogenised by vortexing before 20 µL were added to the tubes containing the PCR amplicons, mixed by pipetting and left at room temperature for 2 min to homogenise. Samples were transferred to a 96-well plate and placed in a magnet stand until the supernatant was clear (around 2 min), which was carefully removed and discarded. The mixture (beads + samples) was then washed

twice with ethanol and after its removal left to air-dry for no longer than 10 min. Next, tubes containing the mix were removed from the magnetic stand, 52.5 µL of Tris (10 mM, pH 8.5) was added, gently mixed by pipetting and incubated again at room temperature for 2 min. Finally, the samples were placed back into the magnetic stand until the supernatant was clear, and 50 µL was transferred to a new tube. The amplicons, now cleaned, were analysed using a TapeStation (Agilent Genomics) and the screentape High Sensitivity D1000 quick assay. In detail, samples and ladder were prepared by mixing 2 µL of High Sensitivity D1000 buffer with 2 µL of amplicon (or High Sensitivity D1000 Ladder) and were placed in the machine for the exact determination of amplicon size.

2.4.2.4 Index PCR reaction

After purification and confirmation of the right sizes of each 16S rRNA amplicon, the next step was to attach the Illumina sequencing adapters to the amplicons by PCR, using the Nextera XT index kit. The PCR reaction was carried out using 5 µL of amplicon, 25 µL of 2X KAPA HiFi HotStart Ready Mix, 5 µL of Nextera XT Index 1 Primer (N7XX), 5µL of Nextera XT Index 2 Primer (S5XX) and 10 µL of nuclease-free water. The specific primers used for each sample is shown in the table 2.3 and the PCR conditions used are listed in table 2.4.

Table 2.3 Illumina Index primer adapters used for each one of the 16S rRNA amplicons.

Index	**N705**
S506	16S1
S507	16S2
S508	16S3

Table 2.4 PCR conditions for the inclusion of Illumina adapters to 16S rRNA amplicons

Step	**Temperature (°C)**	**Time**	**Repeat**
Initial denaturation	95	3 min	
Denaturation	95	30s	x8
Annealing	55	30s	
Extension	72	30s	
Final extension	72	5 min	
Hold	4	Forever	

2.4.2.5 Library quantification, normalization, and pooling

Products obtained by the index PCR were once again cleaned/purified by AMPure XP beads and the new sizes were confirmed by Tapestation analysis. According to Illumina's recommendation, the quantification of library DNA was performed by a fluorometric method that uses dsDNA binding dyes and samples were normalized to 4 nM each in 5 µL total volume. Thus, DNA quantifications were performed using Qubit Fluorometric Quantification from ThermoFisher. Firstly, Qubit working solution was freshly prepared by diluting Qubit dsDNA HS Reagent in Qubit dsDNA HS buffer. Then, the two Qubit standards and each sample were diluted in the working solution, they were gently mixed, placed at room temperature for 2 min and applied on Qubit Fluorometric Quantification from ThermoFisher. DNA was quantified by comparison with the standard curve and all samples were normalized to 4 nM in 5 µL by dilution in Tris (10 mM pH 8.5) and re-read on Qubit. After each sample was normalized, they were all pooled together and the final concentration of the pool was verified by a new reading on Qubit.

2.4.2.6 Library denaturation and MiSeq sample loading

Following Illumina's recommendation, the pooled 16S rRNA amplicon library was denatured and hybridized with hybridization buffer before the final steps of cluster generation and sequencing. For this, 5 µL of freshly prepared NaOH (0.2M) was combined with 5 µL of the normalized pool, homogenised by vortexing and placed for 5 min at room temperature. Then hybridization buffer was added twice in order to serial dilute the pool to 20 pM and 4 pM, respectively. The 20 pM pool was kept and stored at -20 °C and the experiment continued with the pool at 4 pM, which was kept in ice while the same procedure of denaturation and dilution was repeated for PhiX (which is used as an Illumina's internal control, increasing the diversity of libraries). Pool and PhiX were combined (to 25% of PhiX) in one tube, heat denatured for 2 min at 96 °C, kept in ice for 5 min and finally loaded onto the MiSeq platform.

2.4.3 Bioinformatic analysis and microbial community profile pipeline

The results obtained by MiSeq sequencing were carefully analysed and a bioinformatic pipeline was developed in order to create a microbial community profile for the bacteria based on the 16S rRNA sequence information. The bioinformatics steps were carried out under supervision of Dr. Daniel Leadbeater. First, files were individually unzipped to generate fastq format files. Next, forward

and reverse primers of the 16S rRNA sequences sharing the same index were merged using Vsearch version 1.11.1 [113]. Then, both reads were trimmed from the nextera linker and forward reads were trimmed from the random dodecamer sequence using Cutadapt version 1.11 [114]. Fastaq files were split and the fasta file obtained had the headers formatted to Usearch format ("barcodelabel=sample_id;sequence_number_integer"). Replicates were concatenated into a single file, trimmed from the primers used followed by global trim to 250 bp lengths using Usearch version 9 (fastx_truncate). Files were assigned by abundance and sorted by size using Usearch version 7 (derep_fulllength) [115]. Then, files were clustered into Operational taxonomic units (OTUs) using the UPARSE algorithm [116] with simultaneous *de novo* chimera detection using Usearch version 9 (cluster_otus) with a 97% identity threshold. OTUs were relabelled from sample IDs to OTU numbers using Usearch (fasta_number.py). Representative sequences for each OTU were then mapped to the original sequences using Usearch version 7 (usearch_global). Taxonomy was assigned using QIIME version 1.9 (assign_taxonomy.py) against the Greengenes 13.8 database. Finally, the output file was converted to text file format by Usearch (uc2otutab.py), and further converted to .biom and .tsv (Biom: convert) for custom analysis in Python (version 3.6). Table 2.5 details all the commands used for this pipeline:

Table 2.5 Commands used for the analysis of 16S rRNA amplicon database. Steps 1 to 6 were performed individually for each of the files (16s1, 16s2 and 16s3), which were put together in step 7. Ref_set refers to the database file used for the taxonomy and R1 to R15 refers to each of the input and/or output files.

Step	Function	Command
1	Unzip files to a .fastq format	gzip -d *.gz
2	Merge pair ends	vsearch --fastq_mergepairs 16s1_S1_L001_**R1**_001.fastq --reverse 16s1_S1_L001_**R2**_001.fastq --fastqout 16s1**R3**.fastq
3	Remove of Nextera linker	cutadapt -g GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGT -a CTGTCTCTTATACACATCTGACGCTGCCGACGA --overlap 10 -o 16s1**R4**.fastq --discard-untrimmed 16s1**R3**.fastq
4	Remove the dodecamer	cutadapt --cut 13 -o 16s1**R5**.fastq 16s1**R4**.fastq
5	Fasta split	convert_fastaqual_fastq.py -c fastq_to_fastaqual -f 16s1**R6**.fastq
6	Format header	python qiime_to_usearch_compatible_lib_format_conversion2dan.py

7	Concatenate	cat Usearch_formatted_sequence_file_16s1**R6**.fna Usearch_formatted_sequence_file_16s2**R6**.fna Usearch_formatted_sequence_file_16s3**R6**.fna > 16sR7.fna
8	Remove primers	cutadapt --cut 19 --minimum-length 200 -o 16s**R8**.fna 16s**R7**.fna
9	Global trim	usearch_v9 -fastx_truncate 16s**R8**.fna -trunclen 250 -fastaout 16s**R9**.fasta
10	Dereplicate	usearch_v7 --derep_fulllength 16s**R9**.fasta --output 16s**R10**.fasta --log --sizeout -- minuniquesize 2
11	Sort by size	usearch_v7 -sortbysize 16s**R10**.fasta -output 16s**R11**.fasta -minsize 2
12	Cluster	usearch_v9 -cluster_otus 16s**R11**.fasta -otus 16s**R12**.fasta -minsize 2
13	Relabel samples ID to OUT numbers	python drive5/fasta_number.py 16s**R12**.fasta OTU_> 16s**R13**.fasta
14	Map OTUs	usearch_v7 -usearch_global 16s**R8**.fna -db 16s**R13**.fasta -strand plus -id 0.95 -uc 16s**R14**.fasta
15	Assign taxonomy	assign_taxonomy.py -i 16s**R13**.fasta -o 16s**R14**.txt --similarity 0.9 -r ref_set.fasta -t ref_set.txt
16	OUT table	python drive5/uc2otutab.py 16sR14.fasta > 16s**R15**.txt

2.4.4 RNA preparation for meta-transcriptomics and RNA sequencing

2.4.4.1 DNase treatment

The samples containing RNA/DNA (section 2.4.1) were first treated with DNase Max Kit MoBio following the manufacture's guidance. In details, 40 µL of sample (DNA/RNA) were mixed with 10 µL of 10X DNase Max buffer, 1 µL of DNase Max enzyme and nuclease-free water to 100 µL, and the mixture was incubated in the heating block at 37 °C for 20 min. Next, 10 µL of DNase Max Removal Resin (prior mixed and homogenised) were added and left at room temperature for 10 min (with gentle inversions every 2 minutes). The samples were centrifuged at 13000 g for 1 min and the supernatant (free of resin) was transferred to a new tube.

2.4.4.2 Total RNA cleaning and concentration

The samples previously treated with DNase were cleaned and concentrated using the RNA clean and concentrator kit from Zymo Research following the manufacture's guidance (for RNA > 17 nt). For this purpose, 100 µL of the total RNA were mixed with 200 µL of RNA binding buffer and 300 µL of ethanol, 100%. The mixture was transferred to the column supplied and centrifuged for 30 s at 12000 rpm. The flow-through was discarded and samples were concentrated by continuously adding more RNA solution to the same column. Next, 400 µL of RNA prep buffer was added to the column, which was centrifuged and washed twice with 700 µL and 400 µL of RNA wash buffer, respectively. After the second wash, the column was transferred to new tubes, 15 µL of nuclease-free water was added and the column was left for 1-2 min at room temperature, centrifuged for 1 min and the eluate was collected. Another 15 µL of nuclease-water was added and the second eluate collected again (in the same tube). An aliquot of the final concentrated total RNA was quantified by 280 nm absorbance using NanoDrop and the quality and integrity of the total RNA was analysed using a Bioanalyzer. After confirmation of its integrity, total RNA was immediately stored at -80 °C and kept for further investigations.

2.4.4.3 Total RNA depletion, cleaning and concentration

Samples containing total RNA obtained in the step above were depleted of ribosomal RNA (rRNA) using Ribo-Zero Magnetic Kit (Epidemiology) from Illumina according to the manufacture's instruction. Firstly, 90 µL of the magnetic beads solution were transferred to 1.5 mL RNase-free tubes and placed in a magnetic stand, for 1 min with caps open. The supernatant was discarded and 225 µL of nuclease-free water was added and vortexed in order to resuspend the beads. The tubes were once again placed in magnetic stand and the supernatant removed. Finally, 35 µL of magnetic bead resuspension solution and 0.5 µL of RiboGuard RNase Inhibitor were added to the tubes, which were gently mixed for resuspension and kept at room temperature. Next, probes present in the removal solution hybridize to rRNA present in the samples. For this, 4 µL of Ribo zero reaction buffer, 14 µL of total RNA and 2 µL of Ribo zero removal solution were added to 1.5 mL tubes and well mixed by pipetting (10-15 times) before incubation at 68 °C for 10 min in a water bath. The tubes were gently centrifuged to collect the condensation and left at room temperature for 5 min. Removal of the ribosomal RNA was carried out by the addition of the hybridized samples to the tubes containing the magnetic beads solution, followed by immediate mixing by pipetting (10-15 times), incubation at room temperature for 5 min and incubation in a heating block at 50 °C for 5 min. The samples were removed from the heating block and immediately transferred to the magnetic stand, with caps open. After 1

min, the supernatant was transferred to a new tube and the RNA samples (now rRNA-depleted) were cleaned and concentrated. Steps of cleaning and concentration were performed in a similar way as described above (section 2.4.4.2), however at this time, the protocol followed was for RNA > 200 nts. For this, equal volumes of RNA binding buffer and ethanol were mixed together and 2 volumes of this mix were added to the depleted RNA samples prior to their application onto the column. All the other steps were as detailed above (section 2.4.4.2) and the remaining messenger RNA (mRNA) was analysed and quantified by the Bioanalyzer before the best samples were selected for sequencing.

2.4.4.4 RNA sequencing

RNA sequencing was performed at the Next Generation Sequencing Facility at the University of Leeds using HiSeq3000 from Illumina Technology to generate the required 150 bp paired end data. The library construction was completed using Illumina's TruSeq stranded mRNA library protocol, starting at the RNA fragmentation step as suggested by Illumina.

2.4.4.5 Contig assembly of the transcriptome

Contig assembly for this work was performed by Dr Yi Li at the University of York. In short, files containing the paired-end raw reads were downloaded in .fastq format from the Illumina website. The raw reads were mapped to a ribosomal database rRNA_115_tax_silva_v1.0 (downloaded from the SILVA database https://www.arb-silva.de/) using the Bowtie2 software [117] and ribosomal RNA contaminations were removed. Sequences were pooled and assembled into a reference file with the Trinity assembly software package version 2.2.0 [118]. Finally, individual reads were mapped against the reference file created with BWA software [119] and the read count was performed with the Samtools software packet [120]. The final file was used as a database for searches with the proteome library.

2.4.5 Protein extraction and extracellular protein purification

Extracellular proteins were recovered from both the supernatant and bound fractions (proteins that are bound to the biomass) following the methodology previously described by Alessi *et al*., [121]. For this purpose, around 50 mL of a mix containing biomass and supernatant from the final recalcitrant biomass flasks (section 2.2.2) were transferred to a 50 mL (falcon) tube and centrifuged

at 4500 rpm for 20-30 min. The supernatant was transferred to new 50 mL tubes (supernatant fraction) and the pellet was kept (bound fraction).

2.4.5.1 Protein precipitation from the supernatant fraction

Around 40 mL of the supernatant fraction were transferred to Sorvall tubes and ultra centrifuged for 30 min, at 4 °C and 12000 rpm. Supernatants were filtered through 0.22 µM PES filters and 5 mL of the filtered supernatant were transferred to a new 50 mL tube. Five volumes of 100% ice-cold acetone were slowly added and after gentle mixing by inversion, samples were left overnight at -20 °C for total precipitation. In the next day, the mixture was centrifuged for 20 min, at 4 °C and 4500 rpm, the acetone discarded and the pellet washed twice with ice-cold 80% acetone. For complete removal of acetone, the samples were left for 30-45 min under a snorkel-extractor before the pellets were resuspended in 1 mL of 0.5X PBS and stored at -80 °C for further experiments.

2.4.5.2 Biotinylation and precipitation of bound fraction proteins

Enough ice-cold 0.5X PBS was added to the pellet from section 2.4.5 until the volume reached the 50 mL mark of a falcon tube. The pellet was resuspended by vortexing and centrifuged at 4500 rpm, 4 °C for 20 minutes. The Supernatant was discarded and the pellet was washed two more times as described above. Aliquots of 2.5 g biomass were transferred to new falcon tubes containing 19 mL of 0.5X PBS and 10 mM of freshly prepared biotin solution (Biotin EZ-link-Sulfo-NHS-SS-biotin from Thermo Scientific) in 0.5X PBS. Tubes were placed in a rotator with slow agitation (at 10-12 rpm), at 4 °C for 1 h to allow biotinylation (tagging of extracellular proteins associated with the biomass). Next, samples were centrifuged at 4500 rpm, 4 °C for 10 min, supernatant was discarded and the reaction was quenched by the addition of 25 mL of 50 mM Tris-HCl pH 8.0. Samples were centrifuged for another 30 min (same conditions as mentioned before), the supernatant was discarded and the pellet washed twice with 20 mL of 0.5X PBS. Supernatant was removed and the extraction of proteins from the remaining pellet was made by the addition of 10 mL of pre-heated (60 °C) SDS 2%. Samples were incubated in a rotator with slow agitation for 1 h at room temperature, centrifuged and the supernatants transferred to a new 50 mL tube. From this stage, protein precipitation was performed as described above (section 2.4.5.1), the pellet was dried under a snorkel-extractor and resuspended in 1 mL of 0.1% SDS/PBS, filtered through 0.22 µm PES filters and reserved for future application to Streptavidin columns (see section 2.4.5.3).

2.4.5.3 Purification of biotinylated proteins

For the purification specifically of the biotinylated proteins (extracellular proteins associated with the biomass) from the extract prepared in section 2.4.5.2, Streptavidin 1 mL columns from GE Healthcare were used. Firstly, the columns were washed with 10 mL of 0.1% SDS/PBS using a peristaltic pump at a flow rate of 1 mL/min. Next, 1 mL of the biotinylated sample was loaded onto the column using 1 mL syringes at a maximum flow rate of 0.5 mL/min. To aid binding, the columns were incubated with the protein extract at 4 °C for 1 hour and subsequently washed with 10 mL of 0.1% SDS/PBS, again using the peristaltic pump at the same conditions as stated before. The elution was performed by the addition of 1 mL freshly prepared 50 mM dithiothreitol (DTT) to the column and incubation overnight at 4 °C. The next day, another 1 mL of DTT was loaded onto the column and the first elution was collected. The columns were incubated for 1 hour at 4 °C, and through an extra addition of 1 mL DTT, the second elution was collected. The eluates were kept on ice and before proceeding to the buffer exchange (section 2.4.5.4).

2.4.5.4 Buffer exchange

Both protein fractions (from supernatant and from biomass bound fractions derived from the degradation reactions of recalcitrant saltmarsh biomass) were passed through 5 mL Zeba Spin columns (7k MWCO - ThermoFisher) in order to be desalted. The Zeba columns were firstly washed with ultra-pure water, then placed into 15 mL tubes before the samples from supernatant and bound fractions were individually and slowly applied to the Zeba columns. After centrifugation for 2 min at 1000 g, samples containing the proteins were freeze-dried, resuspended in 300 µL dH_2O and kept at -80 °C for future use.

2.4.6 Proteomic analysis

Both desalted protein fractions (from supernatant and bound fractions, section 2.4.5.4) were submitted to proteomic analysis. For this, 26 µL of each sample were mixed with 4 µL of NuPAGE reducing agent and 10 µL of NuPAGE loading buffer (both Invitrogen), and heated for 10 min at 70 °C. The samples were then, individually applied to Invitrogen NuPAGE 10% Bis-Tris precast gels (maximum load of 40 µL sample) and submitted for a short electrophoresis run of 5-6 min at 200 V, only long enough for the proteins to enter the gel. The samples were stained with Coomassie Blue solution for 1 hour and destained under water for 30 min. The visualised band containing all proteins of each

sample were excised from gels and sent to the Bioscience Technology Facility at the University of York https://www.york.ac.uk/biology/technology-facility/proteomics/) for further analysis.

2.4.6.1 Protein Identification by Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) analysis

The proteomic analysis described in this section (2.4.6.1) was performed by Dr. Adam Dowle. To identify the proteins contained in the excised gel samples, in-gel tryptic digestion was performed after reduction with dithiothreitol (DTE) and S-carbamidomethylation with iodoacetamide. Resulting peptides were analysed by label free LC-MS/MS over a 125 min gradient using a Waters nanoAcquity UPLC interfaced to a Bruker maXis HD mass spectrometer as detailed in [122]. Protein identification was performed by searching tandem mass spectra against the transcriptomics database previously obtained (section 2.4.4.5) using the Mascot search program (http://www.matrixscience.com/) and filtered to accept only peptides with expect scores of 0.05 or better. Molar percentages were calculated from Mascot emPAI values by expressing individual values as a percentage of the sum of all emPAI values in the sample [123].

2.4.6.2 Protein annotation

Identified proteins were annotated using dbCAN [124] (a specialised web server and database for automated carbohydrate active enzymes -CAZymes - annotation) and BlastP searches against the non-redundant NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using a Linux platform, followed by manual inspection for similarity to known CAZymes. These steps will be detailed in depth in the results and discussion section of chapter 4, addressing some complications of the process.

2.5 Molecular Biology techniques

2.5.1 Host organisms for cloning and protein expression

Bacteria were chosen as the host organism for gene cloning and recombinant protein expression in this project. *Escheria coli* (*E. coli*) strains Stellar™ ultra-competent cells (ClonTech) were used for the cloning steps. *E. coli* Rosetta-gami™ 2 (DE3) and BL21 competent cells from Novagen and *E. coli* ArcticExpress (DE3) competent cells from Agilent Technologies were utilised for heterologous expression.

2.5.2 Media

The media used for the bacterial growth cultures were Lysogeny Broth (LB), Super Optimal broth with Catabolite repression (SOC), auto-induction medium (AI) and M9 minimal medium.

2.5.2.1 LB medium

LB medium is one of the most routine media used in the laboratory to grow cultures. It is a nutritional rich medium and was prepared by dissolving 25 g of LB Broth Miller (Fisher BioReagents) in 1 L of dH_20, adjusting the pH to 7.0 and autoclaving at 121 °C for 15 min. LB-agar, used for growing individual bacterial colonies on a semi solid surface in petri dishes (agar plates), was prepared by dissolving 40 g of LB-agar Miller (Formedium) in 1 L of dH20 and autoclaving as described before.

2.5.2.2 SOC medium

SOC medium is also commonly used to grow cultures. Because this medium is richer in nutrients, it is typically used to improve efficiency of transformations by providing better growth conditions. The SOC medium used in this work was obtained as ready to use liquid from Sigma.

2.5.2.3 Auto-induction medium

Auto-induction medium was used as an alternative to LB in the attempt to optimise levels of expression for some proteins. Basically, this medium contains different carbon sources that are metabolized and consumed differentially by the bacteria, promoting the growth of the culture followed by induction of protein expression from lac-based promoters when the media is depleted of glucose. The auto-induction medium used in this work was obtained from Formedium and was prepared by dissolving 55.85 g into 1 L of dH_2O, adjusting the pH to 7.0 and autoclaving at 121 °C for 15 min

2.5.2.4 M9 minimal medium

M9 minimal medium was also used as an alternative to LB in the attempt to optimise levels of expression for some proteins. This medium contains only the minimal essential nutrients for the growth of cells and is usually supplemented with various amino acids and carbon sources. Because it is not a rich medium, the growth of cells is slower, which is desirable in cases where cells can produce

toxic substances. Also, when levels of protein expression are lower and slower, there is more time and better chances for proteins to be folded in the correct form.

To prepare this medium the salts were made up first by adding 64 g Na_2HPO_4 heptahydrate, 15 g KH_2PO_4, 2.5 g NaCl and 5.0 g NH_4Cl to 800 mL of dH_20 and autoclaving at 121 °C for 15 minutes. The M9 minimal medium was completed by mixing 200 mL of the salt solution with 2 mL sterile 1M $MgSO_4$, 20 mL sterile 20% glucose, 100 µL sterile 1 M $CaCl_2$ and dH_20 to 1L.

2.5.3 Polymerase Chain reaction (PCR)

Since the annotation of the target proteins always referred to a prokaryote organism, all the steps of cloning were performed using genomic DNA (section 2.4.2.1) as template for PCR reactions. Pairs of primers were designed external to the target sequences and whenever possible, a new pair of nested primers were designed internal to the first primer sequence, but still external to the target sequence (figure 2.1). Nested PCR was used in cases where PCR reactions using the first pair of primers had apparently failed or produced several visible DNA bands in agarose gels due to unspecific amplification.

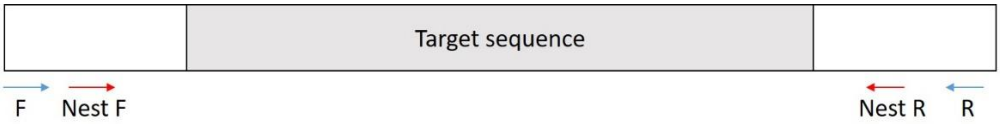


Figure 2.1 Strategy of the primer design for gene cloning. Target sequence is exemplified in grey. Arrows and colours represent each of the primers designed: F and R in blue for external forward and reverse primers; Nest F and Nest R in red for the forward and reverse nested primers, which are external to the target sequence but internal to the external primers.

Primers were designed so that each primer consisted (when possible) of 15-22 bp with melting temperatures (Tm) of primer pairs as similar as possible to each other avoiding differences in Tm higher than 5 °C. For calculations of the Tm the online calculator from ThermoFisher was used (https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/tm-calculator.html). All primers used for cloning are listed in table 2.6. The PCR reactions were performed using 2 ng/µL gDNA, 10 µM of each forward and reverse primer, 4 mM dNTP mix, 4

µL 5X Phusion HB buffer and 0.2 µL Phusion polymerase (Thermo Fisher) adjusted to a final reaction volume of 20 µL with nuclease-free water. PCR conditions are listed in the table 2.7. After PCR amplification, 3 µL of the reactions were analysed by agarose gel electrophoresis to confirm the presence and length of the PCR products.

Table 2.6 List of primers used for the amplification of the genes targets. Targets 2, 16, 17 and 18 were excluded because they were truncated forms of other selected targets. F and R are forward and reverse primers, respectively. Nest F and Nest R are forward and reverse nested primers, respectively. Sequences are shown in 5' to 3' orientation

Target		Primer	Sequence
1	GH5	F	CATAATTTCGGAGAAGTGATTATGC
		R	TTTTTATTACGCTCAGGCTTTATTA
		Nest F	TTATGCTTTGTAGCGATACATTGCG
		Nest R	TTTATTAATTACCGGCCCACAAG
3	GH5	F	AGATCGTGATGTGACTGGAGTTC
		R	GGCCAAGAAAAGGCTGAAAA
		Nest F	TTCCGATCTCTCGAGGTGATAATAA
		Nest R	ATTGTCCGGGCTTTTTGG
4	CE6	F	ATTTACCAAAAGTGAAATATATGAA
		R	CAGAGGGGGATTAAGAATGTAA
		Nest F	GTGAAATATATGAACAATAAATTATTAACG
		Nest R	GTGGTGCGGGGTAATT
5	CE10	F	ACAACAGACAATTCCAAACATACGG
		R	CCTACACGACGCTCTTCCGATCT
		Nest F	TAGGTCGAATGGCTTATATATCTG
		Nest R	CGCCGTTTTCAATTGAAATATT
6	GH10	F	GCTGTTTATATTTCATACAGTTTTTAGG
		R	TGTCAAACCTTCATCAAACCCTA
		Nest F	TTGTAAGCTAAAGATAATAAATAATGG
		Nest R	AACCTTCATCAAACCCTAGTTT
7	PL9	F	CTCGGATTGCCCCAATTCTTAC
		R	ACAGGAACTACCAGCCGAAAC

		Nest F	ACCAAGCAACGTTATCCAGCTGAA
		Nest R	ACGCGGTTATTGCGGTG
8	GH51	F	CGGCCAATTTGGAATTTGAATAGA
		R	AAAATGCGCGGCACCAAG
		Nest F	GAGATATATAACTAGGAGATTTTGAGAA
		Nest R	GCCGCGCAGTAATCTAAAT
9	GH3	F	AATCCGGGAGAATCAGCGTC
		R	GAACCTACAAGGGCGGCTTTG
		Nest F	TTTGCCCCGGAAGACCTTTA
		Nest R	GCAAAATATGTCATATCGATACTGACCTA
10	AA2	F	GATCTCTGAAGACCATATGCATG
		R	TTTTAATGATACGGCGACCAC
		Nest F	TCAACATAAGGCGGAGTGAAGTAA
		Nest R	ATACGGCGACCACCGAGATCTA
11	GH3	F	CCGATCTAAACTGTCAAAAATCAAA
		R	GATACGGCGACCACCGAGATCTACA
		Nest F	CTGTCAAAAATCAAATCAAATTATTATG
		Nest R	CACCGAGATCTACACTCTTTCCCTA
12	CE1	F	TTCCGATCTATTTTCAATGCGC
		R	CATTGCGAGTTCGCTCTTTAAAG
		Nest F	TTTATAATAAGTGAGTGAGATAATTATG
		Nest R	GACACACCCAAATAATGAATTT
13	GH11	F	TCAGGTGCTCTCCTGCGACGTTTAA
		R	ACGGCGACCACCGAGATCTACACTC
		Nest F	ACTTTTTACCAATCATGTTTAACGC
		Nest R	ATCTACACTCTTTCCCTACACGACG
14	GH3	F	TAAAGCCCGGCGAGACACATA
		R	TTAAAGCGAGGGTTGCGG
		Nest F	TTTGACTGTGGAAGAATTTTGAGGAG
		Nest R	TGTAGTCGGGTCATGGACCAG
15	GH5	F	GCATATCGCCGAAGATGACAA
		R	TCATATCGGGAATGCCCG
		Nest F	AAGCTGCACATTCTAAAAAGGAGC

		Nest R	AAAGCCGTAAAAACTTTGCGTTTTT
19	Hydrolase	F	AAAATGCTCAACCGCCGC
		R	ACCCGGCCTGCGTCA
		Nest F	ATGCTCAACCGCCGC
		Nest R	TGCAGGCCGGAGATGAGT
20	CE10	F	GTGCTCTTCCGATCTTTCTAAACAG
		R	CGAAGACATGCCCGACATC
		Nest F	CGATCTTTCTAAACAGGCGATTTAT
		Nest R	CCCGGATGGAGTAGGAAGGA
21	GH6	F	TTAAGTGCCAACATTAACTGCTC
		R	CCGGGTTCTTGATACATCTAAAAA
		Nest F	GTTACAGCAAAGTTTAGGGAGA
		Nest R	TACATCTAAAAAGGATCAGTTTTTA
22	CE1	F	AGTATCGACTTGAAACCGACG
		R	CAGTTAAGAAAATTGAAATTAAAGC
		Nest F	ACCATTAGCGGTGGTTATGC
		Nest R	TGGTTGCTTTGATCGATTAAT
23	GH10	F	TCTCTCCGTTCTCATCCTCAAT
		R	AATCGAATCGAAAAGCATCAGC
		Nest F	TGGAACACCAATGAATTTATTGATA
		Nest R	TTTCGGCGAATCTCACAATCA
24	GH109	F	AAAAAGTAAAACTTGTTTTGCTTTT
		R	ACCACCGAGATCTACACTCTTT
		Nest F	TGCTTTTAATAATTCAATACAAATG
		Nest R	TACACTCTTTCCCTACACGAC
25	CE15	F	AGACGTGTGCTCTTCCGATCTAGC
		R	TGATACGGCGACCACCGAGAT
		Nest F	AACTCTAAGTTGCCTGATCCGTTCA
		Nest R	CACCGAGATCTACACTCTTTCCCTA
26	CE15	F	AACACCTGTCGATACCAAAGAAAAA
		R	ATCTCATGCGGATCCGGC
		Nest F	TATGATGCGCTACGTTTATGGTATG
		Nest R	AGTGTCGTCGTGAAGAAATTCTGG

27	GH3	F	TACTCCAGGGAGCGACCTTC
		R	CTACACGACGCTCTTCCGATCTA
		Nest F	ACTCGTGCGCCGCGT
		Nest R	---
28	GH3	F	CAAGCTTTCTTTGCTCCCA
		R	TTCAAATGTTCTTTTGAGATTTCA
		Nest F	ATCGTCCTCAAAGGAGATCCCA
		Nest R	TTTCCGTTGTCGCGCTG
29	CE10	F	GCATACGCAACAATTCTTTATGATT
		R	AAGGCGAATCTTGAAGGATCAA
		Nest F	ATTTAACAAAAGAAATAACACATTA
		Nest R	GCCAACAAAAATTTATTATCAGGTA
30	AA2	F	AGACGTGTGCTCTTCCGATCTCA
		R	GTTGGAGCAACGCATCCTT
		Nest F	CGATCTCAGGAGATAACACAATG
		Nest R	GCTGAAAAATATCGACTAAGGATAA
31	GH67	F	TGATGCCCAAGCTGCCCTATTAC
		R	CTACACGACGCTCTTCCGATCTC
		Nest F	TCAAGGTGAGGCTTTGGCA
		Nest R	---
32	Peroxidase	F	TTGGCATAGACCCCATATATCGAC
		R	AGACAAAAGAGACAACTCGCCA
		Nest F	AACATCAATAAACCATTAAAGAGGA
		Nest R	TTTTGGAAGCTTTAAGATTAAC
33	Peroxidase	F	TTCGCGAATAGAAACCCACTAAA
		R	GCGACCACCGAGATCTACACTC
		Nest F	TTTAAAGAGGAATTTAAGATGGCCG
		Nest R	---
34	CE1	F	ATCTTTAACTCTTGGCCGCG
		R	AGTCTTTTCATTCTCAAAATCTCC
		Nest F	AGCCATGTCTCGCGA
		Nest R	ATATCTCTAATTAATTAGGTCTATTCAAA
35	AA3	F	ATTATAGTGCAATAACAAGAACTGA

		R	ACCTAAAATAATTGAGGATGTTTTT
		Nest F	AAAGCTATCAAATCTTACGGGTTT
		Nest R	TACGTCTCTTCTTTTCATTTTATAA
36	AA6	F	CAGGCACTCCTCGAACTGAAC
		R	GGTTTGCGCCGCGAT
		Nest F	CATAAAGCACGCGTGAGGG
		Nest R	GATCTGCCGTCCCGATCA
37	CE8	F	ATCCGTTGTGCGTGCG
		R	AAAAAGGTGGCGGCCATAT
		Nest F	AAAAAGTAATTGGGAGAATTTAACC
		Nest R	CACCTTTTTTTTAGATATCCGTGTG

Table 2.7 PCR conditions for the amplification of the genes targets.

Step	Temperature (°C)	Time	Repeat
Initial denaturation	98	2 min	
Denaturation	98	10s	
Annealing	60	30s	x35
Extension	72	30s	
Final extension	72	10 min	
Hold	4	Forever	

2.5.4 Agarose gel electrophoresis

PCR products were applied to agarose gel electrophoresis to separate and visualise DNA fragments of different sizes. The gel was prepared at a concentration of 1% agarose solubilised in 0.5x Tris-borate-EDTA buffer (TBE, 45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8.0) by heating in a microwave at full power for 1-2 min. The agarose-TBE solution was cooled at room temperature and 0.5 µg/mL ethidium bromide were added before pouring the gel solution into a casting tray with a comb inserted to mould the wells for sample loading once the gel has set. After solidified, the comb was removed and the gel was placed in an electrophoresis tank filled with 0.5x TBE buffer. The samples were mixed with 5x loading buffer (50 mM Tris, 5 mM EDTA, 50% (v/v) glycerol, 0.3% (w/v) orange G, pH 8.0) and loaded into the wells alongside 5 µL of the molecular weight marker 1 kb Hyperladder (Bioline) to estimate the size of separated DNA fragments. The gels were usually electrophoresed at

120-140 V for 30-45 minutes (depending on the gel size) and the DNA was visualised under UV light using a UVITEC transilluminator (Cambridge).

2.5.5 DNA purification

PCR reactions consisting of a single amplification product identified as a single gel band upon electrophoresis were purified straight from the reaction solution (liquid phase) using the Wizard SV gel and PCR clean-up System (Promega) following the manufacturer's recommendations. In detail, an equal volume of membrane binding solution was added to the PCR amplification products, and the mixture was transferred to a column provided, incubated for one minute at room temperature and centrifuged at 16,000 g for one minute. After discarding the flow-through, 700 µL of membrane wash solution were applied to the column before centrifugation at 16,000 g for one minute. The step was repeated with 500 µL of membrane wash solution and 5 min centrifugation. After allowing evaporation, the column was transferred into a new tube, and bound DNA eluted with 50 µL of nuclease-free water. An aliquot was analysed by gel electrophoresis and upon confirmation of the purification process, the PCR product was stored at -20 °C until further use.

PCR reactions containing unspecific bands were completely applied to agarose gel electrophoresis, the band with the correct size was excised and the DNA was extracted from the gel. For this, an equal amount (w/v) from the membrane binding solution was added to the gel slice containing the DNA and the mixture was incubated at 60 °C for 15 min with gently inversion every 2-3 min. Once the gel was liquefied, the sample was transferred to the column provided and the procedure followed as above.

2.5.6 Gene cloning using StrataClone technology

In this project, it was decided to first clone the genes of interest into a cloning vector using StrataClone Blunt PCR cloning kit (Agilent Technologies). This strategy was adopted for several reasons: to increase the efficiency of cloning; to facilitate further steps of subcloning into expression vectors; and to be able to verify cloning of the correct sequence by DNA sequencing. The kit contains a mix of two blunt-ended linearized parts of a vector, each arm charged with the Topoisomerase I at one and the loxP recognition site at the other end. Blunt-ended PCR products are efficiently ligated to these vector arms using Topoisomerase I as mediator. The reaction is very quick (5 min of incubation at room temperature) and once the ligation is completed, a transformation (section 2.6.2) is performed using competent cells provided with the kit without any cleaning steps necessary. These

specialized cells express the Cre recombinase, which is an enzyme that recognizes the loxP sites and circularizes the vector with the insert. Figure 2.2 (taken from the manufacturer's manual) exemplifies the steps mentioned above. The vector provided in the kit has ampicillin and kanamycin resistance as its selection markers, is 4269 base pairs long and includes a lacZ α-complementation cassette for blue/white colony screening.

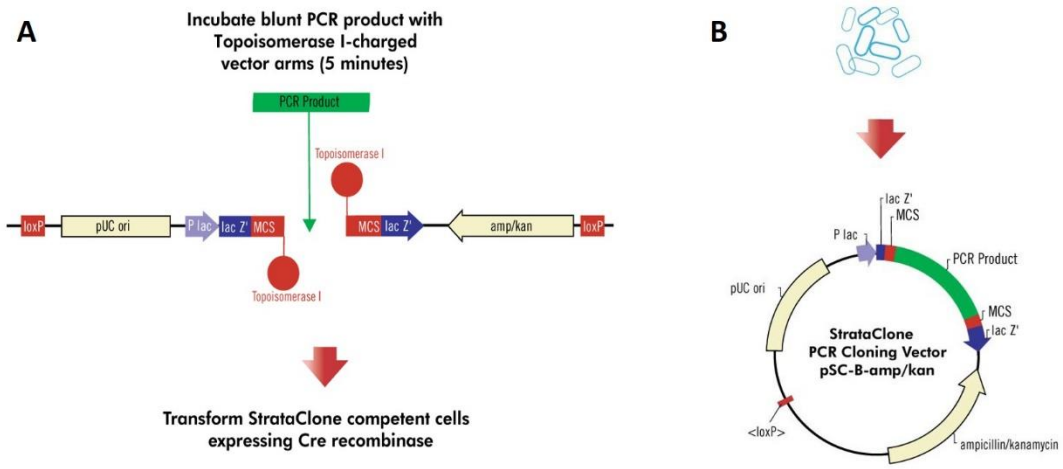


Figure 2.2 Operating scheme of the StrataClone technology (taken from the manufacturer's manual). **A** shows the two arms of the linearized vector charged with Topoisomerase I and loxP on each arm and where the PCR product will be ligated into the vector. **B** show the circularized vector after transformation using StrataClone solo competent cells, which express Cre recombinase.

Cloning was performed following the manufacturer's instructions. For this, 3 µL of StrataClone Blunt Cloning Buffer, 2 µL of the PCR product (diluted if necessary) and 1 µL of StrataClone Blunt Vector Mix were gently mixed by pipetting and the reaction was left at room temperature for five minutes and then placed on ice. 1 µL of the reaction was used to transform an aliquot of the StrataClone SoloPack competent cells following the standard procedure described in section 2.6.1. An aliquot of the transformation mixture was plated on LB-ampicillin plates containing 40 µL of 1% w/v X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) in DMF (dimethylformamide) to allow blue/white colony screening (section 2.5.7) and incubated at 37 °C overnight.

2.5.7 Colony screening by colony PCR

The identification of positive transformant bacteria was performed by colony PCR. This procedure involves picking a colony from the plate (only white colonies picked in the case of blue/white screening) using a sterile pipette and dipping it into a PCR mixture containing 5 µL of DreamTaq Green PCR Master Mix 2x (Fermentas), 10 µM of each forward and reverse primers and nuclease-free water to a total volume of 10 µL. Colony PCR conditions are listed in table 2.8. For colony PCR of transformation reactions using the StrataClone vector, M13 forward and reverse primers were used, while for reactions using the pet52b+ vector (used for protein expression) T7 promoter and terminator primers were used (table 2.9). Amplification products of the colony PCR was applied to agarose gel electrophoresis and colonies with DNA inserts of the expected size, had their plasmid DNA extracted and sent for sequencing (section 2.5.8).

Table 2.8 PCR conditions for the colony PCR.

Step	Temperature (°C)	Time	Repeat
Initial denaturation	95	2 min	
Denaturation	95	30s	
Annealing	55	30s	x35
Extension	72	30s	
Final extension	72	5 min	
Hold	4	Forever	

Table 2.9 Primers used for colony PCR reactions. Pair M13 Forward and M13 Reverse were used for the StrataClone vector and pair T7 promoter and T7 terminator were used for the pet52b+ vector.

Primer		Sequence
M13	Forward	TGTAAAACGACGGCCAGT
	Reverse	AGGAAACAGCTATGACCAT
T7	Promoter	TAATACGACTCACTATAGGG
	Terminator	GCTAGTTATTGCTCAGCGG

2.5.8 Plasmid DNA extraction and Sanger sequencing

Colonies identified as positive transformants were transferred to 5 mL of LB media containing the appropriate antibiotic/s for selection and were incubated in a shaker overnight at 37 °C and 180 rpm. On the next day, cells were pelleted by centrifugation and plasmid DNA was extracted using the Wizard Plus SV Miniprep DNA Purification kit (Promega), according to the manufacturer's instructions. In detail, the pellet was resuspended into 250 µL of cell resuspension solution and inverted to mix. 10 µL of alkaline protease solution was added, gently mixed and incubated for 5 min at room temperature. Next, 350 µL of neutralization solution was added, once again the mixture was gently inverted and centrifuged at top speed for 10 min. The supernatant was transferred to a provided column and centrifuged at top speed for 1 min. The Flow-through was discarded and the column washed by the addition of 750 µL of wash solution. The tube was centrifuged and a new wash was performed using 250 µL of wash solution. Samples were centrifuged at top speed for 2 min, columns transferred to a new collection tube and plasmid DNA eluted by addition of 100 µL of nuclease-free water followed by centrifugation. The plasmid DNA was stored at –20 °C for further experiments.

2.5.9 Nucleic acid quantification

When required, the quantification of DNA and/or RNA was performed using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). 1 µl of sample was used and the concentration was measured by absorbance at 280 nm against a blank. A 260/280 ratio was considered pure for DNA at 1.8 and at 2.0 for RNA, with lower values indicating possible contamination with protein, phenol or other substances.

2.5.10 DNA sequencing (Sanger sequencing)

An aliquot of extracted plasmid DNA (section 2.5.8) was sent for sequencing at GATC Biotech (https://www.gatc-biotech.com/en/index.html), following the instructions provided by the company. The M13 forward and reverse primer pair was used for sequencing the StrataClone vector and the T7 promoter and terminator primer pair for sequencing the pet52b+ vector. The chromatograms obtained were analysed using BioEdit software.

2.5.11 Subcloning into the expression vector pet52b+ using In-Fusion HD cloning

The gene targets that were cloned successfully into the StrataClone vector were subcloned into the pet52b+ expression vector (Novagen) using the In-Fusion HD cloning system (Clontech laboratories). In this technology, the chosen vector (pet52b+ in this case) has to be linearized by conventional PCR reaction first. For the success of this technology, the primers used for the amplification of genes of interest are designed in a way that each forward and reverse primer have 15 nucleotides complementary to the ends of the linearized vector. These overlapping ends will be recognized by the In-Fusion enzyme, which ligates the DNA fragments in order to circularize the final product.

The primers used for linearization of the pet52b+ plasmid were designed to eliminate the multiple cloning site, thrombin recognition site and the histidine tag, but to keep the start codon, the N-terminal streptavidin II tag, the HRV 3C recognition sequence (for the cleavage of strep tag, if needed) and a native stop codon. PCR reactions were performed using 2 ng/µL of plasmid template, 10 µM of each forward and reverse primers, 4 mM dNTPmix, 4 µL 5X Phusion HB buffer and 0.2 µL of Phusion enzyme to a final volume of 20 µL reaction and with the PCR conditions shown in table 2.10. PCR products were incubated at 37 °C for 2 hr with the DpnI restriction enzyme (New England Biolabs) to remove the remaining methylated DNA of the original template.

Simultaneously, all successfully cloned gene targets were also prepared for insertion into the vector by PCR. The reaction was performed following the protocol for the plasmid linearization, except that in this case, each one of the cloned targets was used as template for amplification with their specific primers. In-Fusion primers for the target genes were designed to eliminate the native signal peptide (as a cytoplasmic expression system was chosen) and the overhangs were designed to insert the gene of interest between the HRV 3C site and the stop codon. PCR conditions used for the plasmid linearization and insert preparation are shown in the table 2.10 and all the primers used in this section are listed in the table 2.11. PCR products were applied to agarose gel electrophoresis (1%), bands of correct size were excised from the gel, cleaned and purified as described above (section 2.5.5) and cloned into the expression vector using the In-Fusion technology.

Table 2.10 PCR conditions for the plasmid linearization.

Step	Temperature (°C)	Time	Repeat
Initial denaturation	98	2 min	
Denaturation	98	10s	x35
Annealing	60 (plasmid) 54 (insert)	30s	
Extension	72	2.5 min	
Final extension	72	10 min	
Hold	4	Forever	

Table 2.11 Primers used for PCR reactions of plasmid linearization and insert preparation. For the insert preparation, primers were designed containing complementary tags to the linearized vector (in red).

	Target		Sequence
	Pet52b+ linearization	Forward Reverse	GGGTCCCTGAAAGAGGACTTCAAG TAATTAACCTAGGCTGCTGCCACC
1	**GH5**	Forward Reverse	CTCTTTCAGGGACCCTTAATGTCTGCCTGTG AGCCTAGGTTAATTATTAATTACCGGCCC
3	**GH5**	Forward Reverse	CTCTTTCAGGGACCCGCAGGTTTGAGTGCC AGCCTAGGTTAATTACTACTGCGGTTGGACTG
5	**CE10**	Forward Reverse	CTCTTTCAGGGACCCCAAGTTAGATACGTCGATG AGCCTAGGTTAATTATTATCGGAAAAGTACCTTTT
6	**GH10**	Forward Reverse	CTCTTTCAGGGACCCGCCTGTGGCAACGAG AGCCTAGGTTAATTACTAGTTTCGACCCAAGTATTCC
7	**PL9**	Forward Reverse	CTCTTTCAGGGACCCAATGAGCCTTCGCTTGAA AGCCTAGGTTAATTACTATTGGACGCTCGTATTGG
8	**GH51**	Forward Reverse	CTCTTTCAGGGACCCCAGAACGCCGTCACTC AGCCTAGGTTAATTACTATTCAATTACCCAAACG
9	**GH3**	Forward Reverse	CTCTTTCAGGGACCCGCCAGCACAGGATTAG AGCCTAGGTTAATTACTAGAAAGAGCAGCTCGA
12	**CE1**	Forward Reverse	CTCTTTCAGGGACCCTCTACATCAAGTAGTGGTTC AGCCTAGGTTAATTATTATGGAAGGGTAAACC

14	**GH3**	Forward	CTCTTTCAGGGACCCGCGCCTGCCCCAAAA
		Reverse	AGCCTAGGTTAATTATTAGAATGTGCAGCTTGCCT
15	**GH5**	Forward	CTCTTTCAGGGACCCTATCCCGAACCTAAGGCT
		Reverse	AGCCTAGGTTAATTACTATAGGCCGAGCGCTT
20	**CE10**	Forward	CTCTTTCAGGGACCCGCTTCAGCGACAACC
		Reverse	AGCCTAGGTTAATTACTATTTTGCAAATCCGC
21	**GH6**	Forward	CTCTTTCAGGGACCCGGCACAAACCCAACTCCA
		Reverse	AGCCTAGGTTAATTATTAGTATTCACTTTGTCCCAATGGT
22	**CE1**	Forward	CTCTTTCAGGGACCCATGGTCAGTGGTGAATAC
		Reverse	AGCCTAGGTTAATTATTAAAACTGCGTAATAAAA
26	**CE15**	Forward	CTCTTTCAGGGACCCATGGCTGCCAAGTAT
		Reverse	AGCCTAGGTTAATTATTAGTTCAATACTGTATCCA
28	**GH3**	Forward	CTCTTTCAGGGACCCATGAGAAGGTCATTTCTTATCAC
		Reverse	AGCCTAGGTTAATTATCAGTCACCGCTGACTG
29	**CE10**	Forward	CTCTTTCAGGGACCCCAAGAACGATTTCTCGA
		Reverse	AGCCTAGGTTAATTATTATTCAAATACTACCTTTTC
32	**Peroxidase**	Forward	CTCTTTCAGGGACCCATGGGCGTATTAGTTGG
		Reverse	AGCCTAGGTTAATTATTACAGCTTGTCGTGGT
34	**CE1**	Forward	CTCTTTCAGGGACCCATGGCGACCACTCTGT
		Reverse	AGCCTAGGTTAATTACTATTCAAATTCCAAATTG
35	**AA3**	Forward	CTCTTTCAGGGACCCATGGATATGTTGCGCA
		Reverse	AGCCTAGGTTAATTACTATATTTCTTTGTTATTTAATGCT
36	**AA6**	Forward	CTCTTTCAGGGACCCATGCCGCCCGATACGA
		Reverse	AGCCTAGGTTAATTATCAGCTGGTGATCTTTCCGG

In-Fusion cloning reactions were performed using a 1:2 ratio of linearized insert to vector in the following reaction mix: 4 µL of linearized vector (25 – 100 ng/µL), 2 µL of insert (5 – 100 ng/µL), 2 µL of 5X In-Fusion HD enzyme and nuclease-free water to a total volume of 10 µL. The reaction was incubated at 50 °C for 15 min and after allowing to cool on ice, the ligation was used for transformation (section 2.6.1) into Stellar competent cells from Clontech.

2.6 Recombinant protein expression

2.6.1 Competent cells transformation

After each step of cloning, it is necessary to insert the plasmid containing the gene of interest into a chosen organism in order to perform further downstream experiments. This process is called transformation. Thus, plasmids obtained from the cloning and subcloning steps were chemically transformed into a chosen competent cell. For this, 1-2 ng of plasmid DNA was added to 25-50 µL of competent cell suspension previously thawed on ice. The mixture was incubated on ice for 20 min followed by a heat-shock at 42 °C for 45 s and incubation on ice again for another 2 min. SOC medium (75 – 300 µL) was added to the mix and the cells were incubated in a shaker at 180 rpm and 37 °C for 1-2 hours. The cells were plated on LB agar plates containing the appropriate antibiotic/s for selection of successfully transformed cells and incubated at 37 °C overnight. At the next day, colony PCR (section 2.5.7) was performed to select the positive transformants, which were sent for sequencing (section 2.5.10) to confirm the cloning/subcloning.

2.6.2 Bacterial protein expression

Genes that have been successfully subcloned into the expression vector pet52b+ were submitted to expression trials in different *E. coli* cell strains, media, temperatures and inducer concentrations. BL21 cells were tested as it is a routine procedure in most laboratories. Rosetta-gami™ 2 (DE3) strains were tested as they are improved to express proteins predicted to contain disulphide bridges as well as codons rarely used in *E.coli*. ArcticExpress cells were tested as they grow and express at lower temperatures, as well as contain chaperones that can help the correct folding of proteins. Trials were performed for LB, AI and M9 minimal media, with different concentrations of isopropyl β-D-1-thiogalactopyranoside (IPTG) from 0.5 mM to 2.5 mM (except for AI medium, which does not need the addition of any protein inducer) and temperatures tested were 37 °C for 5 hours; 30 °C for 5 hours and overnight; 20 °C overnight; and 16 °C overnight. Purified plasmids containing the protein targets were transformed into the cells and grown at 37 °C overnight on LB agar plates containing the appropriate antibiotic/s. At the end of the next day, one colony of each plate was used as pre-inoculum for overnight growth of cells at 37 °C in 5 mL of LB liquid medium (containing the appropriate antibiotic/s), which were used in the next day as inoculum for the expression of protein by adding 50 µL of each inoculum individually to 5 mL of the medium containing the appropriated antibiotic/s. Cultures were incubated at 37 °C (Rosetta-gami 2 and BL21) or 30 °C (ArcticExpress) until the optical density recorded at 600 nm (OD600nm) reached 0.6 to 0.8. At this point, IPTG (0.5 – 2.5 mM) was added to each flasks for induction of protein expression (except for AI medium where addition of IPTG

is not required) and each flask was placed at the desired temperature and incubated for a certain amount of time. Each of these conditions was also applied to cells containing the empty vector, pet52b+, which was used as control. After the desired expression period, cells were centrifuged, the supernatant discarded and the pellet stored at -20 °C.

For the analysis of protein expression, cell pellets were lysed using BugBuster from Novagen. For this, each pellet was resuspended in 250 µL of BugBuster mixture (Novagen) containing 1x BugBuster in 1X PBS, 10 µg/mL DNase I (NEB), 0.1 mM of the protease inhibitor 4-benzenesulfonyl fluoride hydrochloride (AEBSF) and 1 mM $MgCl_2$. The mix was incubated at room temperature on a shaking platform at 1000 rpm for 30 min. Next, the samples were centrifuged at 14,000 rpm for 30 min and aliquots of the pellet and the supernatant (insoluble and soluble fractions, respectively) were analysed by SDS-PAGE (section 2.6.3) and Western Blot (section 2.6.4). Trials that showed soluble protein expression were selected and a new expression reaction was set up in bigger culture volumes. Pellets from these expression experiments were resuspended in lysis buffer containing 1X PBS buffer, (0.1 mM AEBSF, 10 µg/mL DNase and 1 mM $MgCl_2$). The cells were lysed by sonication (3 minutes: 3 seconds on, 7 seconds off) on ice using an S-4000 ultrasonic liquid processor (Misonix, Inc). Soluble expressed proteins were separated from insoluble cell debris by high speed centrifugation in a Sorvall Evolution RC (Thermo) equipped with an SS-34 angled rotor at 16,000 rpm for 30 min at 4 °C. Supernatant (soluble fraction) was collected and subsequently purified (section 2.7).

2.6.3 Sodium Dodecyl Sulphate Poly Acrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE gels used in this project were Mini-Protean TGX Precast gels (Bio-Rad) with a polyacrylamide gradient from 4 to 20% and the electrophoresis was carried out using a Mini Protean II apparatus (Bio-Rad). Samples to be analysed were mixed with 5x SDS loading buffer (1% SDS, 10% glycerol, 0.1% bromophenol blue and 100 mM 2-mercaptoethanol) and heated at 100 °C for 5 min to allow denaturation of proteins. The volume of samples loaded onto the gel varied according to the well size. To estimate protein size, 5 µl of pre-stained 1 kb HyperLadder (Bioline) were separated alongside the samples. Electrophoresis was performed in 25 mM Tris, 192 mM glycine, 0.1% SDS, pH 8.3, at 200 V until the disappearance of the marker dye-front from the gel (typically 45 min). Proteins were stained with InstantBlue Coomassie stain (Expedeon) and de-stained by washing several times with dH_20. Gel images were taken using Syngene PXi gel documentation imaging system.

2.6.4 Western Blot (WB) analysis

To help identify target protein expression, WB analyses were performed. In this technique, protein extracts are incubated with a specific antibody against the protein of interest after protein separation by SDS-PAGE. Thereby, recombinantly expressed proteins can be identified and distinguished from other proteins of similar size also present in the cell extract. For WB analysis, proteins were loaded and separated in SDS-PAGE as described above (section 2.6.3) but instead of being stained with Coomassie Blue the gel was incubated with ethanol 20% (for 5 min) and then proteins were transferred onto a nitrocellulose membrane using iBlot 2 dry blotting system (Thermo Fisher Scientific) and iBlot transfer stacks (Thermo Fisher Scientific) according to the manufacturer's instructions. After blotting, the membrane was washed with dH_2O and incubated with blocking buffer (1X BPS with 5% (w/v) skimmed milk powder) for one hour at room temperature on a shaking platform to decrease non-specific background signals. After blocking, the membrane was washed with 1X PBS containing 0.05% Tween 20 three times for five minutes and gentle rocking, before incubation with the anti-Strep II antibody conjugated to horseradish peroxidase (HRP) (Novagen) at 1:5000 dilution in 1X PBS containing 0.05% (v/v) Tween 20 and 1% (w/v) BSA for two hours at room temperature with shaking. Afterwards, a wash step was performed (three times for five minutes and gentle shaking with 1X PBS containing 0.05% Tween 20) and the recombinant proteins were detected by the addition of 2 ml of a mix containing 50% stable peroxidise solution and 50% luminol/enhancer solution (SuperSignal West Pico PLUS Chemiluminescent Substrate kit, Thermo Scientific). HRP activity resulting in chemiluminescence was visualised and recorded with the Syngene PXi gel documentation imaging system.

2.6.5 Protein quantification by Bradford

Proteins were quantified using the Bradford method [125]. In detail, 5 µL of each sample and standard (bovine serum albumin (BSA) from 0.03125 to 1.5 mg/mL in dH_2O) were added to 250 µL of Coomassie Plus Protein Assay Reagent (Thermo Fisher Scientific) in a 96 well optical plate. Blanks were prepared with buffer only (without protein extract). After incubation at room temperature for 5 min, the absorbance at 595 nm (OD595) of the samples was recorded using a Sunrise plate reader (Tecan). A linear standard curve was produced with the absorbance values of the BSA standards and was used to calculate the protein concentration of the samples.

2.7 Protein purification

2.7.1 Affinity chromatography

Successfully expressed protein targets 8GH51, 14GH3 and 34CE1 were purified by affinity chromatography against a StrepTrap HP column (GE Healthcare Life Science) using ÄKTA start (GE Healthcare Life Science) following the manufacturer's instructions. For this, soluble protein fractions obtained in section 2.6.2 were filtering through a 0.45 µm syringe filter and loaded onto a 5 mL StrepTrap HP column at a rate of 1 mL/min. Washes were performed with 1X PBS and the elution was carried out with the same buffer with the addition of 2.5 mM desthiobiotin (Sigma) and a rate of 1 mL/min. Eluted fractions showing absorbance peaks at 280 nm were analysed by SDS-PAGE to confirm the presence of the recombinant protein. These fractions were combined, concentrated (if necessary – section 2.7.2) and quantified by Bradford assay (section 2.6.5).

2.7.2 Protein concentration

When required, purified proteins were concentrated by centrifugation using either Vivaspin 2 (Sartorius) or Microsep Advance (Pall Corporation) centrifugal devices. The appropriate molecular weight cut-off size was chosen depending on the size of the protein to be concentrated and samples were centrifuged until reduced to the desired volume according to the manufacturer's recommendations.

2.7.3 Gel filtration chromatography

When required, gel filtration was performed to clean proteins of interest from other contaminants. The purification was performed with the ÄKTA start (GE Healthcare Life Science) using a HiLoad 16/600 Superdex 75 pg column (Ge Healthcare Life Science) equilibrated with 1X PBS. Elution fractions with an absorbance peak were collected and verified by SDS-PAGE. Fractions containing the protein of interested were combined, concentrated and quantified by Bradford assay. Samples were stored at 4 °C until further use in characterization assays.

2.8 Characterization of soluble targets

Soluble purified proteins were submitted to activity tests in order to investigate their function in lignocellulose degradation. A range of different substrates were tested according to the predicted function of each target. Once the substrate against which the highest enzyme activity was detected

was determined, an in-depth characterisation (pH and temperature optimum, seawater influence and salt tolerance against NaCl) was performed with the selected substrate. All enzymatic activities were measured in its linear phase to guarantee that maximal activity was obtained and that results from different enzymes could be compared to each other.

2.8.1 Reagents and substrates

A range of different model substrates based on the release of 4-nitrophenol (pNP) were used to investigate the activity of the target enzymes. The nitrophenyl substrates tested in this work were as follow: ortho-Nitrophenyl β-D xylopyranoside (oNP βxyl), 4-Nitrophenyl α-D manopyranoside, 4-Nitrophenyl β-D manopyranoside, 4-Nitrophenyl-α-L rhamnopyranoside, 4-Nitrophenyl β-D-fucopyranoside, 4-Nitrophenyl β-D galactopyranoside, 4-Nitrophenyl α-D galactopyranoside, 4-Nitrophenyl-α-L-arabinofuranoside (pNP-Ara), 4-Nitrophenyl α-D-xylopyranoside, 4-Nitrophenyl β-D-glucopyranoside (pNP-Glc), 4-Nitrophenyl β-D-xylopyranoside (pNP-Xyl) and 4-Nitrophenyl Acetate (pNP-Ace). They were obtained from Sigma or from Santa Cruz Biotechnology. For the activity tests, purified protein and each of the substrate tested were incubated overnight at 30 °C. All o- and pNPs substrates were used at a final concentration of 0.5 mM and reactions were terminated by the addition of 1 M sodium carbonate (Na_2CO_3) to a final concentration of 0.5 M. Activity was assessed against a 4-Nitrophenol dilution series as standards and the absorbance at 405 nm was measured using a Sunrise plate reader (Tecan).

The polysaccharides arabinoxylan (0.1%), unwashed arabinoxylan (0.1%), debranched arabinan (0.1%) and gum arabic (0.1%) from Megazymes were used to investigate the activity of the AFase GH51. Tests were performed by incubation of the enzyme with each of these substrates overnight at 30 °C. On the next day samples were precipitated with ethanol, resuspended in dH_2O and the supernatant was analysed in HPAEC, compared to standard solutions of arabino-oligosaccharides and xylo-oligosaccharides.

The model substrate methyl ferulate (MFA) from Santa Cruz Biotechnology was used to assess the activity of the FAE CE1. Tests were performed by incubation of the purified CE1 with MFA at 0.5 mM in citrate-phosphate buffer (pH 6) at 1 mL final volume. Reactions were incubated at 30 °C for zero, 10 and 20 min followed by 5 min heating at 100 °C to inactivate the enzyme. For the time zero, the mix containing enzyme and buffer was first boiled at 100 °C for 5 min, before the substrate MFA was added and the reaction was incubated at 30 °C for 20 min. Standards containing only MFA and buffer were also prepared and incubated at 30 °C for 20 min to investigate if spontaneous hydrolysis

of the substrate could occur. At the end of the reaction, samples were centrifuged at top speed for 10 min, supernatant was transferred to a new tube and stored at -20 °C. These samples were sent for analysis by HPLC to investigate the release of ferulic acid (FA). HPLC analysis was performed by Swen Langer at the technology facility at the University of York.

2.8.2 Determination of optimum pH

Buffers used for the enzymatic characterisation were as follow: McIlvaine (citrate-phosphate) [126] buffer for the range of pH 3-7; Tris-HCl for the range of pH 7-9; and Glycine-NaOH for pHs 9 and 10. McIlvaine buffer was prepared by mixing different volumes of 0.2 M disodium phosphate (Na_2HPO_4) and 0.1 M citric acid stock solutions according to the desired pH. Tris-HCl buffer was prepared by mixing different volumes of 1 M Tris stock solution with appropriated volume of HCl according to the desired pH. Glycine-NaOH buffer was prepared by mixing different volumes of 0.5 M Glycine and 0.032 M NaOH stock solutions according to the desired pH. For the meeting pH point of two different buffers, activity tests were performed in both buffers. All reactions were performed in 5 replicates.

Optimum pH was determined by the incubation of each purified protein with the appropriate pNP substrate (0.5 mM final concentration) and buffer in 50 µL final volume reaction. Mixtures were incubated at 30 °C for 30 min and the reaction was terminated by addition of 50 µL 1 M Na_2CO_3. 90 µL of the final reaction were transferred to a plate reader and the activity was measured by absorbance at 405 nm compared to a standard curve of 4-Nitrophenol.

2.8.3 Determination of optimum temperature

Optimum temperature was determined by the incubation of each purified protein with the appropriate pNP substrate (0.5 mM final concentration) and citrate-phosphate buffer pH 7 (optimum pH obtained for the targets tested) at 50 µL final volume reaction. Mixtures were incubated for 30 min at the following temperatures: 0 °C, 10 °C, 20 °C, 30 °C, 40 °C, 50 °C, 60 °C, 70 °C and 80 °C. Reactions were performed in a PCR cycler set to constant temperature (or on ice for 0 °C) and terminated by the addition of 50 µL 1M Na_2CO_3. Again, 90 µL of the final reaction was transferred to a plate reader and the activity was measured by the absorbance at 405 nm compared to a standard curve of 4-Nitrophenol.

2.8.4 Influence of seawater in the optimum temperature

To investigate the influence of seawater to the optimum activity of each enzyme, the experiment performed to determine the optimum temperature (section 2.8.3) was repeated, using artificial seawater instead of buffer.

2.8.5 Salt tolerance against NaCl

To investigate the salt tolerance of each enzyme against NaCl, the experiment performed to determine the optimum temperature (section 2.8.3) was once again repeated, but using different concentrations of NaCl solutions instead of the buffer. The final concentrations of NaCl in the reactions were 0.5 M, 1 M, 2 M and 3 M. Each solution of NaCl was prepared by dissolving the desired amount of NaCl in the citrate-phosphate buffer at pH 7. If necessary, the pH of the final solution was adjusted using either 0.1 M citric acid or 0.2 M Na_2HPO_4 stock solutions.

Chapter 3 Production of the recalcitrant biomass and its compositional analysis

3.1 Introduction

Lignocellulosic biomass (also known as woody plant biomass) is the largest underexploited renewable carbohydrate source on the planet [2] and because it is not used for food, it provides a potential feedstock for the production of second generation biofuels and chemicals, in an environmentally beneficial way without negative impacts on food security. Lignocellulosic biomass is mainly composed of plant cell walls, which are mostly composed of polysaccharides. Its specific composition varies according to the plant source [6], but generally it is made up of three major constituents (cellulose, hemicellulose and lignin) and other minor components such as water, small amounts of pectin, proteins, and minerals. Secondary plant cell walls generally dominate the composition of plant biomass and are responsible for stabilizing the structure of the plants as a whole [7].

Cellulose is the principal component of plant cell walls and is the most abundant renewable organic polymer in nature. It is a polymer of β-1,4 linked D-glucose and occurs in the plant as crystalline and non-crystalline phases. The crystalline structure makes the cellulose highly insoluble and recalcitrant to degradation by microorganisms and enzymatic attack while the non-crystalline forms (also known as amorphous regions) are more easily digested by enzymes [127]. Cellulose microfibrils are embedded in matrix polysaccharides (also known as hemicellulose) which are composed of different amounts of pentose and/or hexose sugars formed by β-1,4 linked backbones and various side chains. Because they are highly branched, hemicelluloses do not form crystalline structures, instead they form long chains that interact with the cellulose [6] and lignin. The abundance, types of glycoside linkages and side chain compositions, broadly varies in hemicellulose according to the plant species and the tissues where they occur [127]. In secondary cell walls, the cellulose and hemicellulose network is interpenetrated by and cross-linked to a hydrophobic polymer called lignin. Lignin is an amorphous and heterogeneous phenolic polymer formed mainly from three monolignols, p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) alcohols [128], through a variety of ether and carbon-carbon linkages [129]. Due to its aromatic nature and extensive cross-linking, lignin is recalcitrant towards degradation and because lignin embeds both cellulose and hemicellulose, it offers protection against microbial and enzymatic degradation. Therefore, even though lignocellulose is typically composed of 75% polysaccharides that can potentially be transformed into biofuels and other

products of industrial interest, the conversion of polysaccharides into monosaccharides presents the major bottleneck for the industrial process, due to the recalcitrance of biomass [127]. In order to overcome this issue, it is desirable to identify efficient enzymes to convert cellulose and hemicellulose into monosaccharides [130], as well as more effective lignin-modifying enzymes. In this chapter I will explain the approach that has been used in aiming to find enzymes that could help in the degradation of recalcitrant biomass. In order to find enzymes robust enough to tolerate salt conditions, this work was performed using sediment from a saltmarsh environment.

3.2 Aims of the chapter

This chapter describes the production of the biomass that is recalcitrant to digestion (referred to as recalcitrant biomass from hereon) used in this project and its characterisation. In order to find enzymes targeting the recalcitrant components of lignocellulose, this recalcitrant biomass residue was used as the only carbon source for the selective enrichment of microorganisms and its composition compared before and after microbial digestion was performed.

3.3 Results and discussion

In order to produce recalcitrant biomass to be used in the search for enzymes targeting the most recalcitrant biomass components, it was decided to use biomass that had already been extensively degraded by microorganisms. As a further aim of the project was to identify halotolerant enzymes, biomass from *Spartina anglica* (a saltmarsh grass) was used and inoculum taken from the sediments of the same saltmarsh. Data from previous experiments performed in our laboratory by Dr. Daniel Leadbeater was used to establish conditions and time for incubation. Dr Leadbeater performed an experiment to analyse the degradation of biomass in saltmarsh sediments and recorded its weight loss over time. He observed that most of the weight loss occurred in the first 3 weeks of incubation (figure 3.1). Interestingly, we can see that from week 4 to week 8, mass loss is very slow and no significant degradation of biomass was observed even though 40% of biomass remained. These findings encouraged us to believe that the remaining 40% biomass is enriched with lignocellulose components responsible for its recalcitrance and biomass produced in this way would be the ideal substrate to be used in this work.

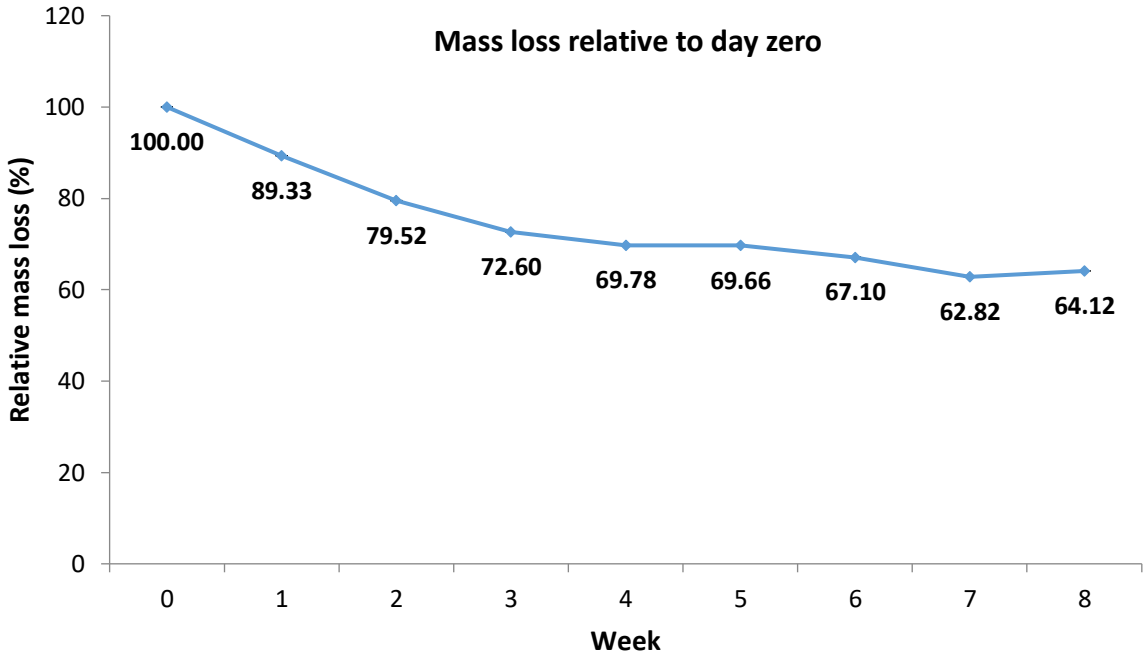


Figure 3.1 Relative biomass degradation through time compared to day zero. Most of the weight loss occurred within the first 3 weeks of incubation. From week 4 onwards, mass loss is much slower and no more significant degradation is observed (Leadbeater *et al*., 2018).

We decided to produce depleted biomass through the incubation of saltmarsh grass *Spartina anglica* with saltmarsh sediment inoculum for 10 weeks prior to harvesting. A longer period was chosen to guarantee that very recalcitrant biomass would be acquired. After this period, the remaining biomass was retrieved through several washes (Materials and Methods, section 2.2), including one wash with SDS to assure that no microorganisms remained attached to the biomass. A compositional analysis of this depleted biomass was performed to confirm the presence of remaining polysaccharides and to determine the relative abundance of its components. Figure 3.2 shows the results obtained for this analysis as well as the relative composition of the original biomass (Leadbeater *et al*., 2018).

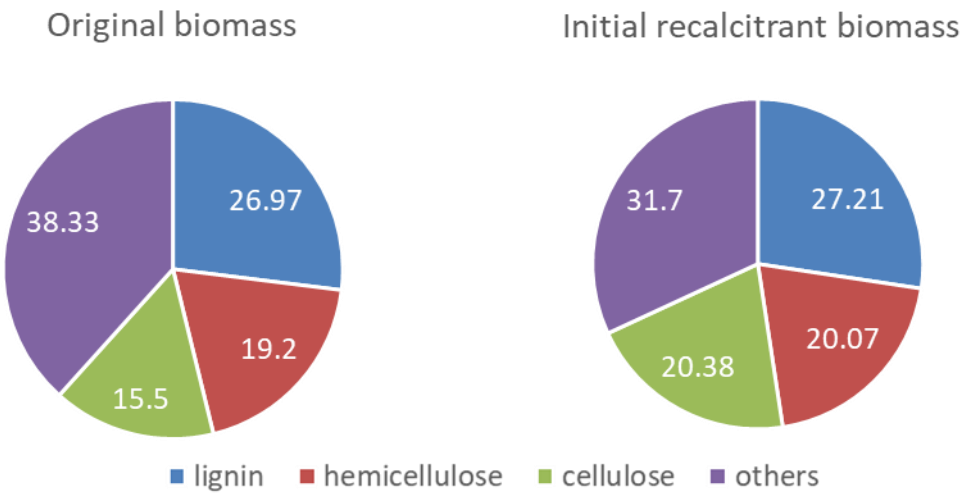


Figure 3.2 Compositional analysis of original *Spartina anglica* and depleted biomass produced during a 10 week incubation period of *Spartina anglica* with saltmarsh sediment. The pie chart shows that 67% lignocellulose is still present in the initial recalcitrant biomass and that it contains 40% remaining polysaccharides (cellulose and hemicellulose). Lignin, hemicellulose and crystalline cellulose were measured using the acetyl bromide, TFA and Anthrone methods, respectively.

As can be seen in figure 3.2, 67% of the depleted biomass is composed of lignocellulose, with lignin representing the largest fraction. A significant amount of other materials is observed, which is likely to be due to inorganic components (such as ashes, for example), proteins and/or soluble sugars that could have been removed during the several wash steps. Ash content in biomass refers to the non-organic matters, as mineral and inorganic materials, that in the case of saltmarsh grasses could be due the presence of silica, iron and sulphur for example. Interestingly, the remaining polysaccharides are composed in equal parts of hemicellulose and cellulose. Although these polysaccharides potentially could be converted into sugar, they remain because they were either not accessible to hydrolases due to the lignin barrier or other compositional and structural features of the biomass, or due to the absence of suitable enzymes for their degradation. Aiming to find enzymes that could help to overcome these barriers for degradation, an experiment was set up using the depleted biomass as the only carbon source and seeded with inoculum taken from saltmarsh sediment. Flasks containing the depleted biomass, minimal media (enriched with 1 µM of Mn) and fresh sediment taken from saltmarsh were incubated for a total of 8 weeks on a shaker. After this period, compositional analysis of the final remaining biomass was performed (Materials and Methods section 2.3). The slow mass loss observed (only 22% during 8 weeks of incubation) confirms the recalcitrance of the biomass. Taking the observed mass loss into account, the compositional analysis shows that all three

components decreased during the 8 weeks but that the decrease in the hemicellulose and lignin fractions was greater than that of the cellulose component (figure 3.3).

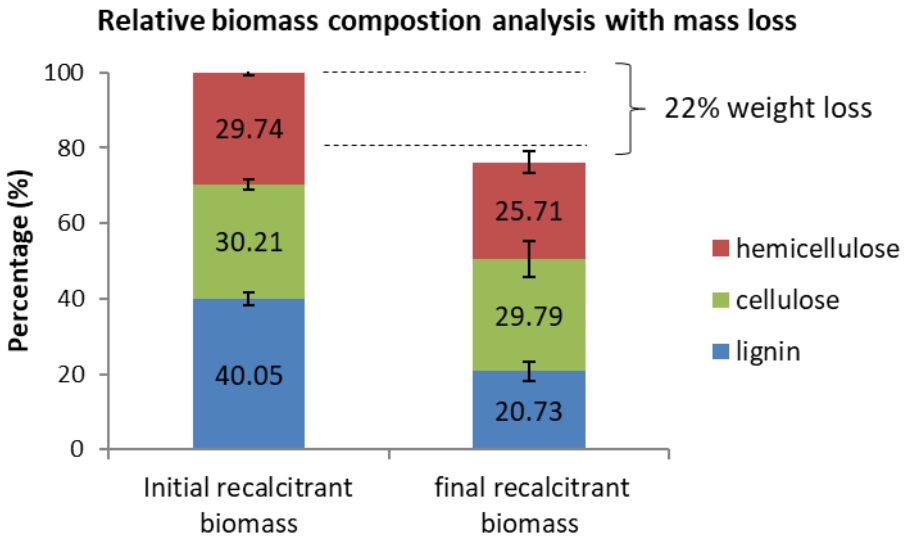


Figure 3.3 Relative lignocellulose content of initial and final recalcitrant biomass taking into account the weight loss observed during the 8 weeks of incubation. Lignin, hemicellulose and crystalline cellulose were measured using the acetyl bromide, TFA and Anthrone methods, respectively. Data are averages of five assays, and the bars represent standard errors.

In figure 3.3 it is evident that lignin is the component that has experienced the greatest decrease during these 8 weeks of incubation, losing almost 50% of its content. Since the biomass used was already very recalcitrant to degradation, the microorganisms growing on it had to produce enzymes able to modify and/or degrade lignin in order to access the remaining polysaccharides. There is also a decrease in hemicellulose content, while the decrease in cellulose was slower. This may indicate that the remaining cellulose is either itself inherently recalcitrant, or that the lignin and hemicellulose need to be mobilised before the cellulose can be accessed. In order to better understand which components of the hemicellulose could be aiding in the recalcitrance of biomass, in the figure 3.4 the variation of each component of the hemicellulose was analysed at the start and end of the enrichments. As the initial biomass has already been degraded to yield the recalcitrant starting material for this analysis, the amount of some monosaccharides is already small even in the initial biomass. Xylose was found to be the most abundant monosaccharide followed by arabinose. Since glucuronoarabinoxylan are the main constituents of hemicelluloses in grasses, it was not surprising to find they are present in larger amounts when compared to mannose and glucose.

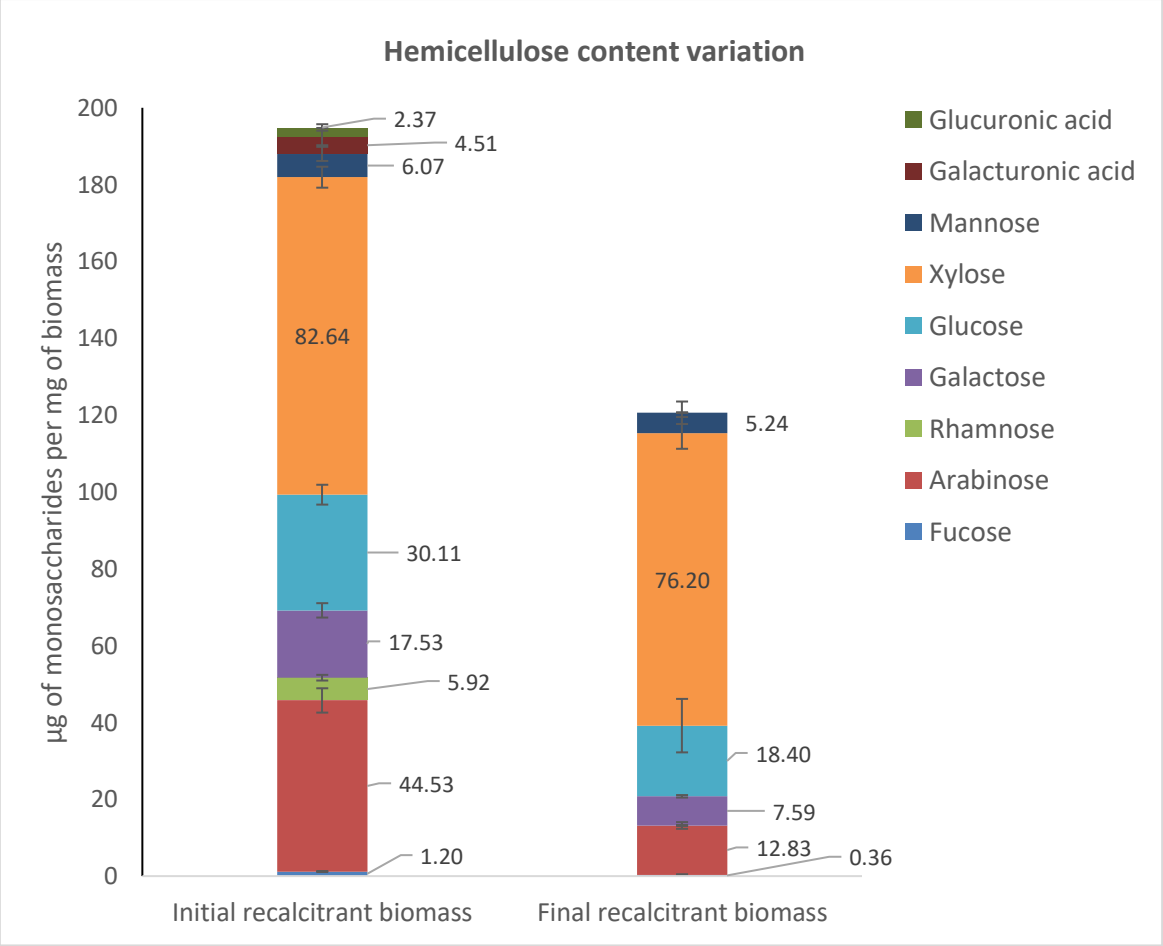


Figure 3.4 Hemicellulose composition of the initial biomass and remaining material after 8 weeks of incubation with saltmarsh sediment. Data are averages of five assays, and the bars represent standard errors.

As shown in the figure 3.4, there was a decrease in all the monosaccharide components of the hemicellulose during the 8 weeks of incubation but the most evident differences are for glucuronic acid (GlcA), galacturonic acid (GalA) and rhamnose, sugars that even though were present in small amounts in the initial recalcitrant biomass, have been completely degraded during the 8 weeks of incubation. GalA and rhamnose are monosaccharides that can be found in pectins while GlcA is typically found in the side chains of the glucuronoarabinoxylan (GAX), which is the main constituent of hemicellulose of grasses. These results suggest that even if pectins are present in small amounts in the secondary cell wall of grasses, they might play important role in the structure and recalcitrance of the plant to degradation, and that either they need to be removed first in order for the other polysaccharides be accessible by microorganisms, or sugars belonging to the pectin are among the most easily accessible sugars remaining in the biomass and thus are completely consumed.

Another very pronounced difference was observed for the arabinose content, which showed over 70% of degradation over the time, varying from 44.53 µg/mg to 12.83 µg/mg. As mentioned before, GAX is the main constituent of hemicellulose of grasses and these results suggest that the enzymes produced by the microorganisms growing on the recalcitrant biomass were acting preferentially on the GAX side chains (evident not only by the variations in arabinose but also from the total degradation of GlcA) instead of the main xylan backbone (xylose only had 7.79% of degradation over the total time of incubation). Galactose, which is a sugar that can be found in side chains of xyloglucans, pectins, galactomannans or glucogalactomannans, also had a considerable reduction during incubation, showing 56.70% of degradation, varying from 17.53 µg/mg to 7.59 µg/mg, while mannose experienced a lower degradation, only 13.67%, varying from 6.07 µg/mg to 5.24 µg/mg. Since galactose can be found in the side chains of galactomannans and/or galactoglucomannans and because it had a higher degradation than the one experienced for mannose, these results also suggest that degradation of the side chains of galactomannans and/or galactoglucomannans were preferred. Finally, 38.89% of the glucose fraction of the hemicellulose, which typically is present in mixed-linkage glucan was degraded (30.11 µg/mg to 18.40 µg/mg).

These results suggest that side chains of hemicellulose can potentially contribute to the recalcitrance of the biomass and for the biomass to be degraded further, the side chains must be degraded by the microbes, making hemicellulose more linear and thus more accessible for other enzymes. Once the side chains of hemicellulose are removed it is expected that the polysaccharides present in the backbone would be the preferred monosaccharide to be consumed. However, the results show that xylose was not greatly consumed during incubation suggesting that much of the remaining xylan is inaccessible.

In order to investigate and to screen for the potential CAZymes involved in the degradation of this depleted biomass, techniques of meta-transcriptomics and meta-proteomics were performed for the microorganisms that grown in the recalcitrant biomass and these experiments and results are described in the next chapter (chapter 4).

Chapter 4 Selection of putative CAZymes through combined proteomic and transcriptomic analysis informed by microbial community profiling

4.1 Introduction

In this work, the aim was to identify halotolerant enzymes that can degrade the most recalcitrant components of biomass. In the previous chapter I explained how complex microbial cultures were grown on biomass that has already been depleted by previous incubation with microbial cultures for a 10 week period. This chapter will focus on the strategies and techniques that were used in order to identify and select putative enzymes related to degradation of this recalcitrant biomass.

Given the complexity of lignocellulose biomass, there is not a unique specific enzyme that is able to degrade it into fermentable sugars. Distinct classes of carbohydrate-active enzymes (CAZymes) act synergistically, each one on a specific bond in order to deconstruct lignocellulose [131, 132]. Glycoside hydrolases (GHs) are the main class of enzymes related to the degradation of plant biomass and act by cleaving glycosidic linkages in the cellulose and hemicellulose components. However, other classes of enzymes that modify and/or break down the lignin network are needed to allow these GHs to attack their substrates effectively. Carbohydrate esterases (CEs) and auxiliary activity enzymes (AAs) are two classes of CAZymes that play important roles in biomass decomposition. While CEs are responsible for the removal of ester groups from carbohydrates, AAs can modify other components of lignocellulose, e.g. changing lignin integrity and/or cellulose crystallinity.

The identification of novel CAZymes has been empowered in recent years by the development of meta-omics techniques [133-135] coupled to high-throughput sequencing platforms [136]. While meta-genomics provides information about the microbial community living in a specific environment and their genomic sequences, meta-transcriptomics reveals the genes being transcribed by that community and their sequences, whereas meta-proteomics allows the identification of the proteins being produced. Meta-omics techniques have been used to successfully identify lignocellulose-related enzymes from different environments and samples [137-140]. Due to the redundancy of the genetic code, microbial community meta-proteomic studies make use of mass spectrometry-based peptide sequencing and are only effective if there is a closely related set of nucleotide sequences that can be translated and searched to identify the peptide sequences. Coupling the analyses of meta-transcriptomes and meta-proteomes provides a powerful tool to accomplish this as nucleotide

sequencing is focused on the expressed transcriptome, enriching it for coding sequences that may correspond to the associated proteome. This combination of approaches can potentially provide an effective way to identify new enzymes, especially when focussed on the appropriate protein types.

Because of the insoluble nature of lignocellulose, microorganisms that are able to digest it usually need to secrete enzymes to break it down into transportable units such as sugars and oligosaccharides. In order to focus this study on lignocellulose degrading enzymes, a meta-secretomic approach was employed, harvesting extracellular proteins from the community of microorganisms that were growing on the depleted biomass to create a proteomic library. Concomitantly, total RNA and genomic DNA were extracted from that community of microbes and used to create a transcriptome library and a microbial community profile, respectively. This chapter details all the steps involved in the creation of these libraries, provides an overview of the bacteria community living in the recalcitrant biomass and explains how the transcriptome library was used as a database to perform searches of the proteome.

4.2 Aims of the chapter

In the previous chapter I explained the approach used in order to force microorganisms to produce enzymes that could be related to the degradation of recalcitrant biomass. In this chapter I will explain how I have performed the identification and selection of putative CAZymes produced from these microorganisms using combined techniques of meta-transcriptomics and meta-proteomics. To this end, this chapter will be focused on DNA preparation for the meta-genomics analysis, RNA preparation for meta-transcriptomics analysis, protein preparation for meta-proteomics analysis, bacterial community profile; and annotation and selection of putative targets.

4.3 Results and discussion

4.3.1 Combined genomic DNA and total RNA extraction

The extraction of total RNA and genomic DNA (gDNA) from the microorganisms growing on the final recalcitrant biomass fractions was performed concomitantly as explained in the Materials and Methods section 2.4.1. RNA extraction from these samples proved to be particularly challenging and time consuming due to the low microbial abundance on the very recalcitrant biomass. After several trials and modifications in the methodology, I observed that bead beating time was the critical parameter in order to have a satisfactory result (which was obtained by two cycles of 1.5min). Figure

4.1 shows the influence of bead beating times during the RNA extraction on RNA integrity, where images A, B and C reflect the results of 2 cycles of 1.5 min, 2 cycles of 2.5 min and 2 cycles of 3.5 min of beating time, respectively. Usually, longer beating times are desirable for a more efficient extraction of RNA from higher microorganisms, such as fungi and eukaryotes. However, RNA in samples submitted to a longer bead beating time (figure 4.1, C) are degraded. Moreover, even though it was possible to extract RNA from these samples using the shortest bead beating time, the amount of RNA extracted was low (light bands for 16S/18S and 23S/28S) and thus several extractions were performed and pooled together in order to collect enough RNA to perform ribosomal RNA depletion and RNA sequencing (next steps for the RNA preparation).

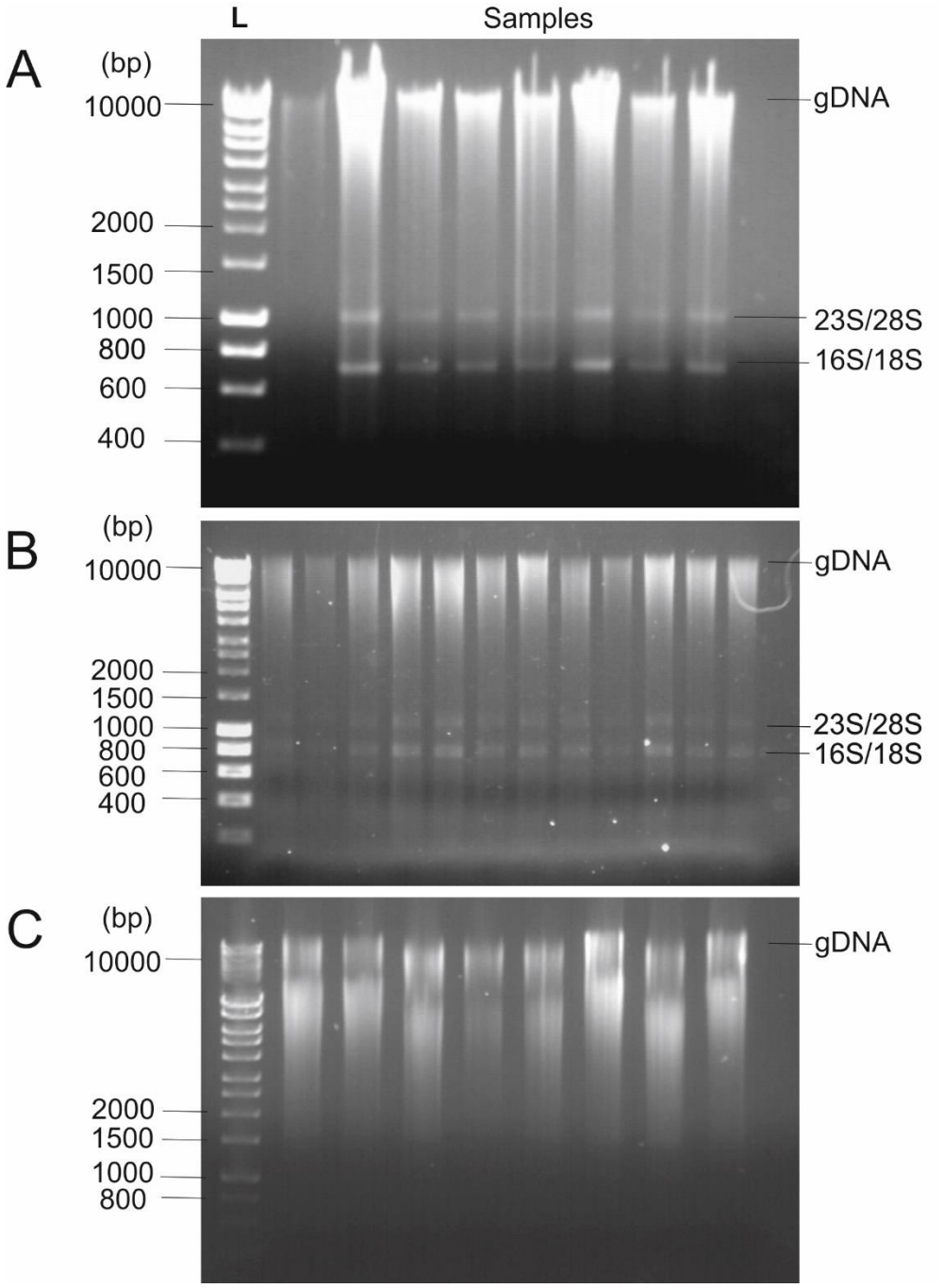
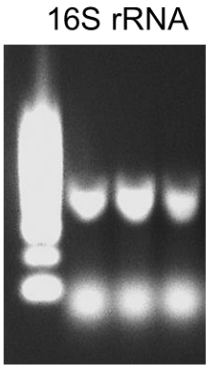


Figure 4.1 Genomic DNA (gDNA) and total RNA extraction showing the influence of bead beating time. L – Bioline Hyper Ladder 1Kb. **A**, **B** and **C** represent bead beating times of 2 cycles of 1.5min, 2 cycles of 2.5 min and 2 cycles of 3.5min, respectively. Best results were obtained using beads beating time of 2x 1.5 min and several samples were submitted to RNA extraction in order to get enough total RNA to perform the next steps.

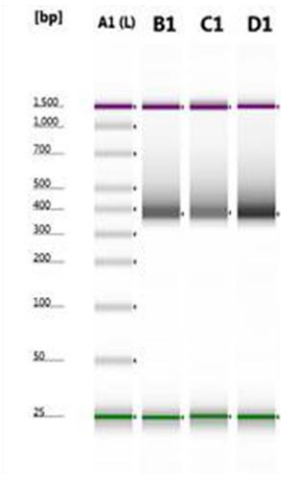
4.3.2 DNA preparation for meta-genomics and DNA sequencing

For the meta-genomics analysis, samples containing gDNA/total RNA were first treated with RNase A, cleaned and concentrated (Materials and Methods section 2.4.2.1). The genomic DNA obtained was then subject to PCR using universal primers to amplify identifiable regions of ribosomal 16S rRNA (Materials and Methods section 2.4.2.2). Products of PCR were applied and excised from agarose gel (figure 4.2 A), and after cleaning (Materials and Method section 2.4.2.3), the sizes of the amplicons were confirmed using a Tapestation (figure 4.2 B). Amplicons were then submitted to an index PCR reaction (Materials and Methods section 2.4.2.4) and the new sizes were again confirmed using a Tapestation (figure 4.2 C). PCR products were quantified, normalized and samples pooled (Materials and Methods section 2.4.2.5). Finally, this library was sequenced using Illumina MiSeq (Materials and Methods section 2.4.2.6) and the results obtained were further analysed and used for the creation of a microbial community profile (section 4.3.6).

A

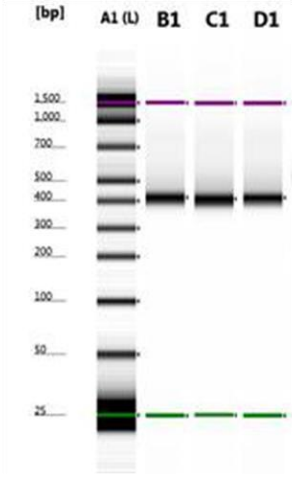


B



Well	Sample	Size (bp)
A1	Ladder	
B1	16s1	382
C1	16s2	386
D1	16s3	381

C



Well	Sample	Size (bp)
A1	Ladder	
B1	16s1	419
C1	16s2	411
D1	16s3	418

Figure 4.2 Steps of DNA preparation for meta-genomics. **A:** Amplicon PCR for the 16S rRNA; **B:** Tapestation analysis after cleaning of PCR amplicon; **C:** New Tapestation analysis after the index PCR reaction. The increase in size obtained for each sample in C (when compared to B) confirms that the index PCR reaction has occurred satisfactorily.

4.3.3 RNA preparation for meta-transcriptomics and RNA sequencing

In order to create a transcriptomic library to be used as a sequence database for the proteomic searches, we performed the sequencing of messenger RNA (mRNA) extracted and purified from the microbial consortium. To this end, the mixture containing total RNA and gDNA obtained in the previous step (section 4.3.1), was treated with RNAse -free DNase (Materials and Methods section 2.4.4.1) for complete removal of gDNA. The remaining RNA was concentrated (Materials and Methods, section 2.4.4.2), quantified and its quality was analysed using a Bioanalyzer. Samples with satisfactory quality, RIN (RNA Integrity Number) ≥ 7, were pooled together and depleted of ribosomal RNA (rRNA), which was confirmed by a new Bioanalyzer analysis (Materials and Methods, section 2.4.4.3). Finally, the enriched mRNA was sequenced using the Illumina platform at the Next Generation Sequencing Facility (NGS) at the University of Leeds (Materials and Methods, section 2.4.4.4). Results are presented in figure 4.3. Note that for the Bionalyzer analysis the origin of the RNA being analysed has to be indicated and here eukaryotic organisms were selected as the source. Therefore, the rRNA bands are labelled as 18S and 28S (figure 4.3), although they could equally be 16S and 23S from prokaryote organisms, as the analysis was performed on total RNA extracted from a whole community of microorganisms that grew on the final recalcitrant biomass.

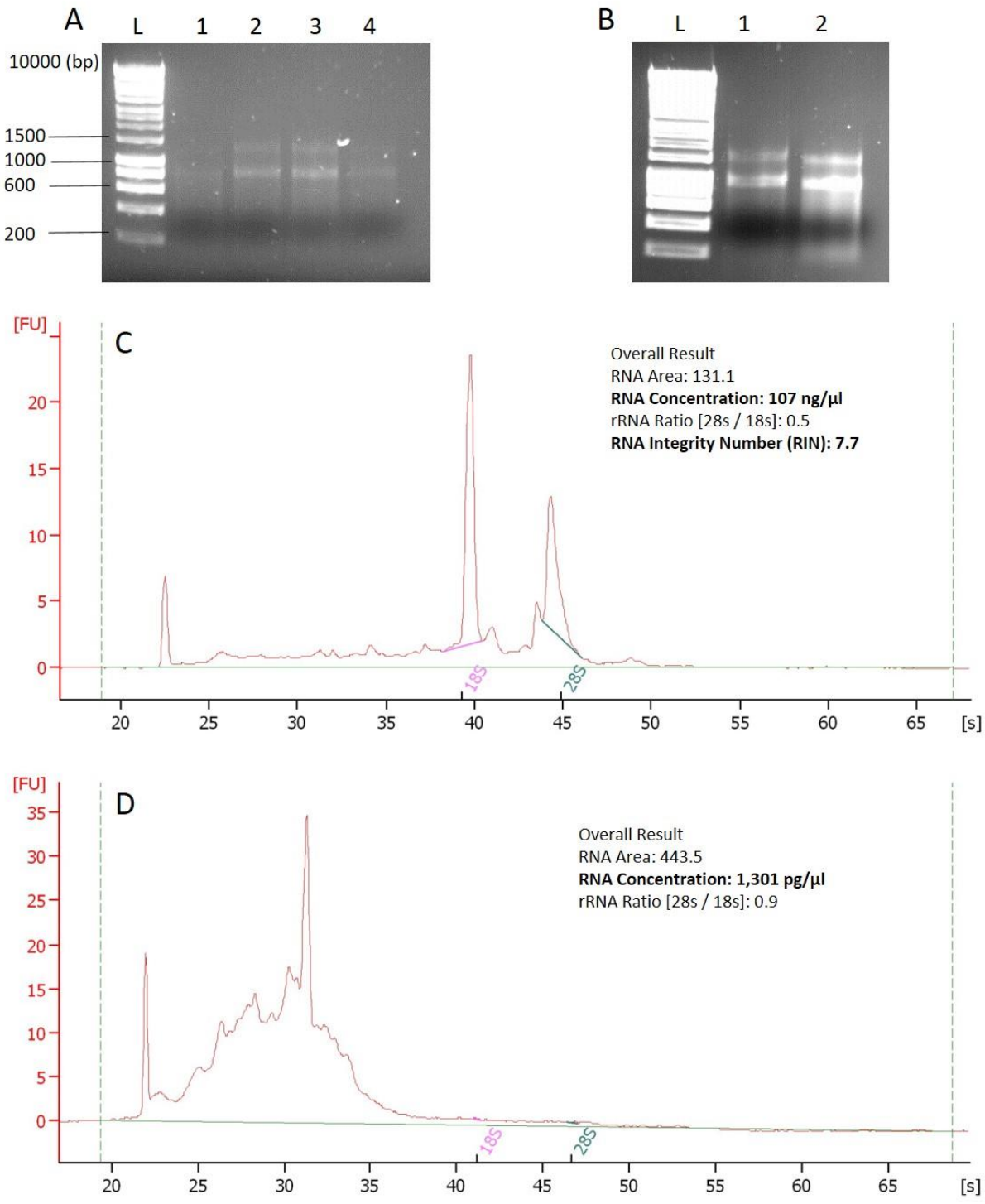


Figure 4.3 Steps of the RNA preparation for meta-transcriptomics analysis. **A:** four different samples of total RNA after treatment with DNase. **B:** samples in A combined, cleaned and concentrated. L: Bioline HyperLadder 1 Kb. **C:** result from the Bioanalyzer for one of the RNA samples present in B and its overall quality (RIN = 7.7), which indicates that the sample is suitable to perform the next experiments. **D:** same sample from C but after ribosomal RNA depletion. The presence of peaks corresponding to 18S and 28S in C and its absence in D shows that the depletion was performed satisfactorily.

4.3.4 Protein extraction and extracellular protein purification

To characterise the extracellular proteins produced by the microbial community, extracellular proteins were labelled with biotin, using EZ-Link Sulfo-NHS-SS-Biotin, a biotinylating reagent that cannot cross cell membranes. After labelling and label quenching (to prevent labelling of intracellular proteins as cells lyse), the whole culture was extracted with SDS, followed by protein recovery and affinity purification of labelled proteins on a streptavidin column (Materials and Methods, section 2.4.5). In this approach, secreted proteins that are either found in the culture medium (supernatant), or attached to the biomass itself (bound fraction) are targeted. Each step of the protein extraction was analysed by SDS-PAGE. To confirm that sufficient protein was obtained from the meta-secretome, one aliquot of each extract was analysed after pooling and concentration, resulting in a visible smear after staining with Coomassie blue (figure 4.4). After affinity purification the biotinylated proteins were pooled and concentrated. These samples were applied to a very short SDS-PAGE run (just enough for the proteins to get into the gel), stained with Coomassie blue and the band containing all proteins together was excised and sent for trypsinolysis and LC-MS/MS analysis at the Bioscience Technology Facility at the University of York, which provided the proteomics results.

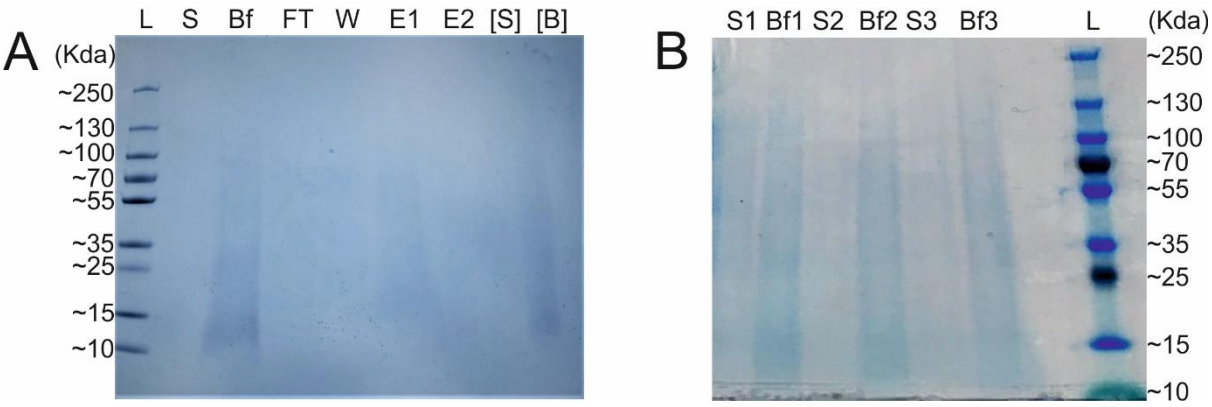


Figure 4.4 Protein extraction and affinity purification of biotinylated proteins from the biomass. **A**: each step of the protein extraction for one biological extraction. L is PageRuler Plus Prestained Protein Ladder; S is proteins from the supernatant post precipitation; Bf is proteins from the bound fraction post precipitation and prior to the application on the Streptavidin column; FT is the flow through from the application of Bf onto the column; E1 and E2 are the 1st and 2nd elution from the column after the addition of DTT, respectively; [S] and [B] are extracted proteins from the supernatant and bound fraction after concentration. **B:** three biological replicates of the concentrated proteins from the supernatant (S1, S2 and S3) and from the bound fractions (Bf1, Bf2 and Bf3).

4.3.5 Protein annotation

In this project, a shotgun proteomic approach was performed, where the extracted proteins were subjected to trypsinolysis prior to separation of the peptide products by HPLC and identification by 2-dimensional mass spectrometry. The redundancy of the genetic code and ambiguity in identifying certain amino acids in peptides means that annotating such data requires a high quality nucleotide sequence database to search against, which contains the coding sequences that correspond to the proteins. The unusual nature of the proteomic approach (using saltmarsh inoculum on depleted biomass) makes it likely that few if any of the coding sequences will be present in public databases. Because of this it was necessary to search the proteomics against a corresponding transcriptomic database. For this purpose, the enriched mRNA was sent for sequencing using Hi-Seq Illumina 3000 technology (Materials and Methods, section 2.4.4.4). The sequence data obtained were assembled (Materials and Methods, section 2.4.4.5) by Dr Yi Li (University of York), using Trinity software, which provided a transcriptomics database composed of approximately 1.8 million contiguous sequences. In order to use these data as a reference database for MASCOT searches of the proteome, the transcriptome sequences were filtered for reads longer than 500 bp using Python software and then transformed into open reading frames (ORFs) using the web server EMBOSS getorf (filtered for a minimum of 300 nucleotides). This new file was then optimized by removing all sequence duplicates (again using Python) before being used as reference for the MASCOT searches of the proteome (Materials and Methods, section 2.4.6.1). MASCOT searches were performed and filtered to require an individual peptide expect score of 0.05 or better, which identified a total of 1953 protein hits in total.

The aim was to identify CAZymes and the strategy involved searching for proteins with similarity to known lignocellulose-active enzymes including glycoside hydrolases (GHs), LPMOs, polysaccharide lyases (PLs), carbohydrate esterases (CEs) and auxiliary activities (AA) enzymes, as lignin peroxidases and laccases for example. The principal approaches were to use dbCAN [124] (a specialised web server and database for automated CAZymes annotation) and in parallel, BlastP searches with annotation to non-redundant databases (in Genbank) and selection of the top three hits (in Linux platform), which were further inspected manually for similarity to known CAZymes. The annotation of the protein hits identified by MASCOT searches using dbCAN provided only 70 potential CAZymes and after eliminating glycosyltransferases (GTs) and duplicates, only 42 hits remained. Considering the size of the transcriptome database (~1.8 millions of sequence), it was suspected that

something could be wrong with these results, and that not all the proteins present in the proteome were being identified. Dr Daniel Leadbeater (CNAP, University of York, personal communication) had also noted this issue before and realized that MASCOT searches with small files proved to have better results than using one unique file. This is likely to happen due to how MASCOT searches are performed, in that the program always creates three "decoy" databases for each database provided. In order to avoid false positives, MASCOT will only return a positive result if the search performed has no hit in any of the decoy databases. This means that the bigger the database provided, the higher the likelihood of a hit in the decoy database to happen and thus more chances of no positive hits returned. Based on this, and in order to improve the power of the searches, the transcriptome file was split into small files containing no more than 25000 sequences. New MASCOT searches of the proteome were performed against each of these individual small files and all the data was pooled together, giving rise to a significant increase in the number of protein hits (total of 6592 hits). These were annotated using dbCAN and BlastP, which returned a total of 216 putative CAZymes. A table containing the top 100 hits proteins according to their abundance in the proteome is presented next (table 4.1).

Table 4.1 Top 100 hit proteins according to their abundance in the proteome (Mol), their presence in the supernatant (SN) and/or bound fraction (BF) and their annotation according to the NCBI non redundant database. (ORF; open reading frame, SN; supernatant, BF; bound fraction, Mol %; molar percentage)

ORF	SN	BF	Annotation NCBInr	Mol (%)
TRINITY_DN261091_c1_g1_i1_4 [3182 - 21]	✓	✓	TonB-dependent receptor [Sphingopyxis baekryungensis]	11.90227
TRINITY_DN261201_c0_g2_i2_4 [2973 - 28]	✓	×	TonB-dependent receptor [Sphingorhabdus sp. M41]	7.654142
TRINITY_DN236941_c0_g1_i2_1 [837 - 37]	✓	×	flagellin [Caldithrix abyssi]	6.715434
TRINITY_DN258779_c1_g6_i1_1 [437 - 3]	✓	×	TonB-dependent receptor [Sphingobium sp. SYK-6]	5.714893
TRINITY_DN259155_c1_g1_i3_1 [146 - 742]	×	✓	amino acid ABC transporter substrate-binding protein [Ruegeria sp. 6PALISEP08]	5.511811
TRINITY_DN257048_c4_g9_i6_1 [101 - 1123]	✓	×	flagellar motor protein MotB [Sphingorhabdus sp. M41]	5.393362
TRINITY_DN254491_c0_g2_i12_2 [3168 - 85]	✓	✓	TonB-dependent receptor [Teredinibacter turnerae]	4.612825
TRINITY_DN258648_c0_g1_i1_2 [1063 - 2]	✓	×	TonB-dependent receptor [Sphingorhabdus sp. M41]	4.561007
TRINITY_DN259155_c1_g1_i1_1 [196 - 762]	✓	✓	amino acid ABC transporter substrate-binding protein [Planktotalea frisia]	4.427604
TRINITY_DN262533_c2_g8_i2_1 [390 - 2459]	✓	×	TonB-dependent receptor [Idiomarina sp. 5.13]	3.941371
TRINITY_DN258648_c0_g1_i2_2 [1153 - 2]	✓	×	TonB-dependent receptor [Sphingorhabdus sp. M41]	3.780189
TRINITY_DN605748_c0_g2_i1_1 [543 - 1]	✓	×	hypothetical protein [Erythrobacter sp. SG61-1L]	3.488914
TRINITY_DN214694_c0_g2_i1_1 [33 - 557]	×	✓	peptide/nickel transport system substrate-binding protein [Celeribacter neptunius]	3.083931
TRINITY_DN235895_c1_g2_i2_2 [497 - 57]	✓	×	TonB-dependent receptor [Cellvibrio sp. pealriver]	3.073527
TRINITY_DN256702_c10_g12_i2_1 [497 - 3]	✓	✓	outer membrane protein/peptidoglycan-associated (lipo)protein [Zhouia amylolytica AD3]	2.965114
TRINITY_DN808627_c0_g1_i1_1 [299 - 619]	✓	×	TonB-dependent receptor [Sphingorhabdus sp. M41]	2.959903
TRINITY_DN261185_c5_g3_i1_1 [26 - 664]	×	✓	hypothetical protein [Devosia sp. H5989]	2.952967
TRINITY_DN262379_c2_g1_i1_1 [201 - 593]	✓	×	Vitamin B12 transporter BtuB precursor [Altererythrobacter atlanticus]	2.927818
TRINITY_DN135867_c0_g1_i1_2 [562 - 2]	✓	✓	beta-tubuli partial [Oxymonadida environmental sample]	2.814578
TRINITY_DN63157_c0_g1_i1_1 [354 - 1]	✓	✓	ribosomal protein S7 [Truepera radiovictrix]	2.78455
TRINITY_DN253181_c0_g1_i3_1 [292 - 1872]	✓	×	hypothetical protein [Sphingorhabdus sp. M41	2.639492
TRINITY_DN244283_c3_g5_i4_1 [770 - 3]	✓	✓	amino acid ABC transporter substrate-binding protein [Wenxinia marina]	2.565082
TRINITY_DN249428_c3_g8_i3_2 [1235 - 3]	✓	×	Vitamin B12 transporter BtuB precursor [Altererythrobacter atlanticus]	2.40647
TRINITY_DN262426_c0_g4_i4_2 [643 - 23]	✓	×	TonB-dependent receptor [Alteromonas sp. V450]	2.352278

TRINITY_DN251149_c1_g2_i2_1 [420 - 1442]	✓	✓	TonB-dependent receptor [Asticcacaulis sp. AC460]	2.315651
TRINITY_DN236855_c0_g1_i2_1 [704 - 189]	×	✓	L-glutamine-binding protein /L-glutamate-binding protein /L-aspartate-binding protein /L-asparagine-binding protein [Cribrihabitans marinus]	2.25823
TRINITY_DN606267_c0_g1_i1_1 [99 - 503]	×	✓	amino acid ABC transporter substrate-binding protein [Roseovarius nanhaiticus]	2.214652
TRINITY_DN256080_c0_g2_i1_2 [248 - 1966]	✓	×	hypothetical protein [Sphingorhabdus sp. M41]	2.197618
TRINITY_DN230172_c0_g1_i1_1 [94 - 423]	✓	✓	hypothetical protein AMJ58_12470 [Gammaproteobacteria bacterium SG8_30]	2.171384
TRINITY_DN242205_c0_g1_i9_1 [102 - 539]	✓	✓	histidine kinase [Marinagarivorans algicola]	2.061031
TRINITY_DN255543_c2_g4_i16_1 [629 - 3]	✓	×	alpha tubulin (fragment) [Trypanosoma brucei gambiense DAL972]	2.012878
TRINITY_DN149829_c0_g1_i1_1 [116 - 511]	✓	×	hypothetical protein [Geobacter sulfurreducens]	1.980194
TRINITY_DN258910_c2_g1_i1_1 [518 - 3]	✓	✓	TonB-linked outer membrane protein SusC/RagA family [Zobellia uliginosa]	1.956824
TRINITY_DN200492_c0_g1_i1_1 [419 - 36]	✓	×	TonB-dependent receptor [Alteromonadales bacterium BS08	1.890095
TRINITY_DN257745_c5_g5_i4_1 [1152 - 118]	✓	✓	hypothetical protein [Hellea balneolensis]	1.889426
TRINITY_DN192193_c0_g2_i1_1 [111 - 557]	✓	×	di-heme cytochrome c peroxidase [Alcanivorax jadensis T9]	1.885899
TRINITY_DN251290_c2_g1_i1_2 [768 - 3137]	✓	✓	TonB-dependent receptor [Sphingobium sp. SYK-6]	1.880472
TRINITY_DN261185_c5_g4_i1_1 [98 - 1114]	×	✓	amino acid ABC transporter substrate-binding protein [Devosia insulae]	1.878742
TRINITY_DN248710_c0_g1_i1_1 [963 - 1]	×	✓	sugar ABC transporter substrate-binding protein [Hoeflea sp. BAL378]	1.864506
TRINITY_DN849725_c1_g1_i1_1 [520 - 2]	✓	✓	polysaccharide biosynthesis protein [Ilumatobacter coccineus]	1.848111
TRINITY_DN249096_c0_g3_i7_1 [583 - 95]	✓	✓	acetolactate synthas large subunit biosynthetic type [Sporocytophaga myxococcoides]	1.823528
TRINITY_DN244271_c0_g2_i4_1 [817 - 32]	✓	✓	peptidoglycan-binding protein [Marinagarivorans algicola]	1.738141
TRINITY_DN259342_c0_g4_i1_1 [1082 - 3]	✓	✓	hypothetical protein [Gilvimarinus chinensis]	1.696627
TRINITY_DN258674_c3_g13_i10_1 [1741 - 80]	✓	×	hypothetical protein [Altererythrobacter atlanticus]	1.654894
TRINITY_DN257988_c0_g1_i4_1 [109 - 1767]	×	✓	glycosyl hydrolase family 5_53 domain-containing protein [Alteromonadaceae bacterium Bs02]	1.597601
TRINITY_DN258692_c0_g1_i1_1 [1206 - 169]	×	✓	D-xylose ABC transporter substrate-binding protein [Shinella sp. HZN7]	1.596729
TRINITY_DN261451_c0_g1_i3_2 [824 - 1384]	×	✓	pilus assembly protein PilN [Marinagarivorans algicola]	1.590089
TRINITY_DN233843_c0_g1_i2_1 [66 - 668]	×	✓	branched-chain amino acid ABC transporter substrate-binding protein [Litoreibacter arenae	1.58862
TRINITY_DN261119_c0_g2_i1_1 [84 - 599]	×	✓	hypothetical protein [Olleya marilimosa]	1.571752
TRINITY_DN261185_c4_g1_i2_1 [19 - 837]	×	✓	hypothetical protein [Devosia sp. H5989]	1.561179
TRINITY_DN257048_c4_g9_i3_1 [101 - 1126]	✓	×	flagellar motor protein MotB [Sphingorhabdus sp.M41]	1.56022

TRINITY_DN254802_c0_g2_i1_1 [850 - 38]	✓	×	hypothetical protein [Leptolyngbya valderiana]	1.554264
TRINITY_DN243337_c0_g1_i2_2 [595 - 2]	×	✓	branched-chain amino acid ABC transporter substrate-binding protein [Pseudodonghicola xiamenensis]	1.553099
TRINITY_DN238138_c0_g1_i2_1 [16 - 525]	×	✓	sugar ABC transporter substrate-binding protein [Thalassospira lucentensis]	1.551827
TRINITY_DN610709_c0_g1_i1_1 [170 - 505]	✓	✓	ketoacyl-ACP synthase III [Woeseia oceani]	1.550021
TRINITY_DN831186_c0_g1_i1_1 [64 - 561]	×	✓	amino acid ABC transporter substrate-bindnig protein [Hoeflea sp. BRH_c9	1.537394
TRINITY_DN255543_c2_g2_i1_1 [560 - 3]	✓	✓	PREDICTED: tubulin alpha-8 chain-like partial [Sarcophilus harrisii]	1.533669
TRINITY_DN260672_c2_g14_i3_1 [1314 - 301]	✓	×	general L-amino acid transport system substrate-binding protein [Rhodobacteraceae bacterium HLUCCA08]	1.533543
TRINITY_DN256515_c0_g1_i1_1 [1012 - 1659]	✓	✓	cell envelope biogenesis protein OmpA [Saccharophagus degradans]	1.528857
TRINITY_DN257254_c2_g5_i2_1 [531 - 1]	✓	✓	TonB-dependent receptor [Sphingobium sp. SYK-6]	1.52594
TRINITY_DN254728_c0_g3_i2_1 [122 - 643]	✓	✓	DEAD/DEAH box helicase [Joostella marina]	1.524082
TRINITY_DN251642_c1_g2_i1_1 [336 - 794]	✓	×	TonB-dependent receptor [Porphyrobacter cryp	1.512357
TRINITY_DN253429_c2_g2_i2_2 [753 - 67]	×	✓	hypothetical protein [Kiloniella spongiae]	1.497675
TRINITY_DN261148_c0_g1_i10_1 [91 - 1899]	✓	×	hypothetical protein [Robiginitomaculum antarcticum]	1.492613
TRINITY_DN262333_c0_g3_i1_1 [366 - 2891]	✓	✓	TonB-dependent receptor [Erythrobacter sp. SG61-1L]	1.479026
TRINITY_DN260672_c2_g7_i13_1 [915 - 40]	×	✓	hypothetical protein [Devosia sp. H5989]	1.46551
TRINITY_DN251924_c0_g3_i1_2 [1157 - 39]	✓	✓	TonB-dependent receptor [Porphyrobacter cryptus]	1.444891
TRINITY_DN243485_c0_g2_i2_1 [701 - 3]	✓	×	flagellar motor protein MotB [Erythrobacter sp. SD-21]	1.440907
TRINITY_DN225253_c0_g1_i1_1 [329 - 3]	✓	✓	hypothetical protein [Cellvibrio mixtus]	1.436519
TRINITY_DN258831_c4_g1_i3_2 [1724 - 4699]	✓	×	TonB-dependent receptor [Alteromonadaceae bacterium Bs12]	1.432029
TRINITY_DN260090_c0_g1_i3_2 [59 - 1522]	×	✓	peptide/nickel transport system substrate-binding protein [Rhodobacteraceae bacterium HLUCCO07]	1.430047
TRINITY_DN262258_c0_g6_i3_1 [78 - 617]	✓	✓	hypothetical protein HLUCCO07_06720 [Rhodobacteraceae bacterium HLUCCO07]	1.404247
TRINITY_DN172411_c0_g1_i1_2 [718 - 29]	✓	×	flagellar motor protein MotB [Erythrobacter atlanticus]	1.3892
TRINITY_DN255771_c0_g4_i2_2 [515 - 3]	✓	✓	elongation factor 1-alpha [Leishmania mexicana MHOM/GT/2001/U1103]	1.370966
TRINITY_DN254941_c12_g14_i1_1 [480 - 79]	✓	×	elongation factor Tu [Pseudomonas alcaliphila]	1.368513
TRINITY_DN212757_c0_g1_i1_1 [106 - 549]	✓	✓	hypothetical protein [Marinagarivorans algicola]	1.364688
TRINITY_DN245749_c0_g2_i4_1 [128 - 574]	✓	✓	Porin subfamily protein [Phyllobacterium sp. CL33Tsu]	1.360996

TRINITY_DN262426_c0_g4_i5_2 [1478 - 3]	✓	×	hypothetical protein [Kordiimonas lipolytica]	1.335768
TRINITY_DN255099_c1_g5_i1_1 [27 - 1415]	✓	×	hypothetical protein SAMN02745824_0225 [Sphingorhabdus marina DSM 22363]	1.327416
TRINITY_DN256100_c0_g1_i3_1 [206 - 760]	×	✓	ABC transporter substrate-binding protein [Labrenzia alba]	1.321467
TRINITY_DN1106727_c0_g2_i1_1 [610 - 2]	✓	×	SMC-Scp complex subunit ScpB [Teredinibacter turnerae]	1.320129
TRINITY_DN259363_c0_g2_i1_1 [259 - 1089]	×	✓	nucleoside-binding protein [Pseudooceanicola nitratireducens]	1.314845
TRINITY_DN210776_c21731_g1_i4_1 [1188 - 1]	✓	✓	hypothetical protein [Methylophilus sp. Q8]	1.312634
TRINITY_DN1096856_c1_g1_i1_1 [129 - 503]	×	✓	hypothetical protein [Gilvimarinus polysaccharolyticus]	1.297573
TRINITY_DN143845_c0_g2_i1_1 [608 - 99]	×	✓	ABC transporter substrate-binding protein [Ahrensia sp. 13_GOM-1096m]	1.291028
TRINITY_DN152519_c1_g1_i1_4 [3280 - 35]	✓	✓	hypothetical protein [Altererythrobacter atlanticus]	1.274744
TRINITY_DN223409_c0_g1_i1_1 [83 - 571]	✓	✓	hypothetical protein [Teredinibacter sp. 1162T.S.0a.05]	1.257131
TRINITY_DN258164_c0_g1_i2_4 [2104 - 524]	×	✓	peptide/nickel transport system substrate-binding protein [Lutimaribacter saemankumensis]	1.242643
TRINITY_DN258253_c0_g6_i2_1 [424 - 29]	×	✓	hypothetical protein [Tangfeifania diversioriginum]	1.225486
TRINITY_DN262426_c0_g4_i3_2 [2000 - 3]	✓	×	hypothetical protein [Kordiimonas lipolytica]	1.216709
TRINITY_DN243251_c0_g2_i1_3 [1459 - 1998]	×	✓	ATP synthase subunit alpha [Marinagarivorans algicola]	1.211068
TRINITY_DN258867_c0_g1_i3_1 [812 - 3]	✓	✓	SusC/RagA family TonB-linked outer membrane protein [Maribacter sp. Hel_I_7]	1.20005
TRINITY_DN104327_c0_g1_i1_1 [191 - 568]	✓	×	hypothetical protein [Erythrobacter sp. SG61-1L]	1.187542
TRINITY_DN254567_c0_g1_i2_1 [104 - 1156]	×	✓	C4-dicarboxylate ABC transporter [Ruegeria sp. ZGT118]	1.186027
TRINITY_DN616652_c0_g1_i1_1 [438 - 1]	✓	×	NitT/TauT family transport system permease protein [Sulfitobacter delicatus]	1.185864
TRINITY_DN151497_c0_g2_i1_1 [84 - 926]	✓	×	flagellar motor protein MotB [Sphingopyxis macrogoltabida]	1.169245
TRINITY_DN157908_c2_g1_i1_2 [725 - 3]	×	✓	amino acid ABC transporter substrate-binding protein [Sulfitobacter donghicola]	1.165288
TRINITY_DN260072_c0_g3_i4_1 [172 - 1410]	×	✓	branched-chain amino acid ABC transporter substrate-binding protein [Sulfitobacter sp. AM1-D1]	1.162512
TRINITY_DN258831_c4_g3_i2_2 [1817 - 4954]	✓	×	TonB-dependent receptor [Teredinibacter turnerae]	1.159641
TRINITY_DN148346_c0_g1_i1_1 [499 - 2]	✓	✓	isocitrate dehydrogenase (NADP(+)) [Herbaspirillum autotrophicum]	1.156846

Table 4.1 shows, among other things, the closest related sequences available in the database. It is important to mention that some of the organisms represented (*Leishmania mexicana*, for example) are highly unlikely to be present in saltmarshes. This happens because the methodology employed only takes into account sequence similarity and, since saltmarshes are a very diverse and poorly studied environment, the source organisms may not be represented in the database. Likewise, among the top 100 hits, a large number of “hypothetical proteins" were identified, which could lead to new enzyme discoveries. Interestingly, nearly all the top hits are either transporter proteins or, in case of the TonB, are receptors associated with transport by mechanisms still not elucidated [141] and these results show that the labelling approach could also be a useful tool for the identification of transporters, thus aiding in studies related to them. Moreover, as was expected with the labelling approach, most of the top 100 hits are cell surface proteins, which shows the effectiveness of the labelling approach. The effectiveness of this technique for detecting biomass bound protein is even more pronounced when comparing the origin (bound fraction and/or supernatant) of the 216 putative CAZymes identified in this work (figure 4.5).

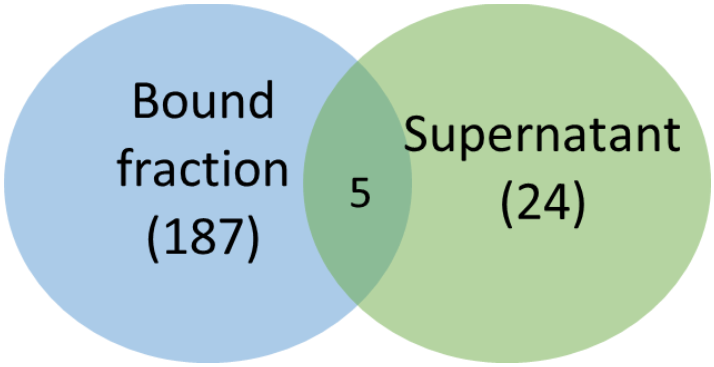


Figure 4.5 Annotation of the 216 putative CAZymes identified in this work. In total 192 putative CAZymes were identified from the bound fraction and 29 from the supernatant. Among them, only 5 were common to both, bound fraction and supernatant, from which 4 are CE8 and one is PL6.

As shown in figure 4.5, the majority (187) of the putative CAZymes identified in this study were only from the labelled (bound) fraction reinforcing the power and effectiveness of the labelling approach for identifying potentially biomass-bound CAZymes. Most meta-secretome studies typically focus only on the proteins secreted to the supernatant and thus are probably missing large amount of possible candidates. The five putative CAZymes identified in both fractions belong either to the CE8 or PL6 families. According to the CAZy database, CE8 are exclusively pectin methylesterases, thus

related to degradation of pectins and PL6 are typically alginate lyases, which are enzymes related to the degradation of alginate (a polysaccharide originated from seaweeds) [142]. Among the 24 putative CAZymes identified only in the supernatant are GH3s, GH23, GH103, CE8s, CE10s, AA6 and PL1. GH3s are typically β-glucosidases, which cleaves cellobiose and other cello-oligosaccharides into glucose; GH23 and GH103 are typically active in peptidoglycan and could indicate bacterotrophic activity among the community; CE10s are a class of enzyme currently removed from the CAZy database as it has not shown yet active in polysaccharides; AA6 are benzoquinone reductase and are believed to be involved in degradation of aromatic compounds [143]; and PL1 are typically pectate or pectin lyases and thus are involved in the degradation of pectins. A summary of the 216 putative CAZymes annotated by both platform, dbCAN and BlastP is shown in figure 4.6.

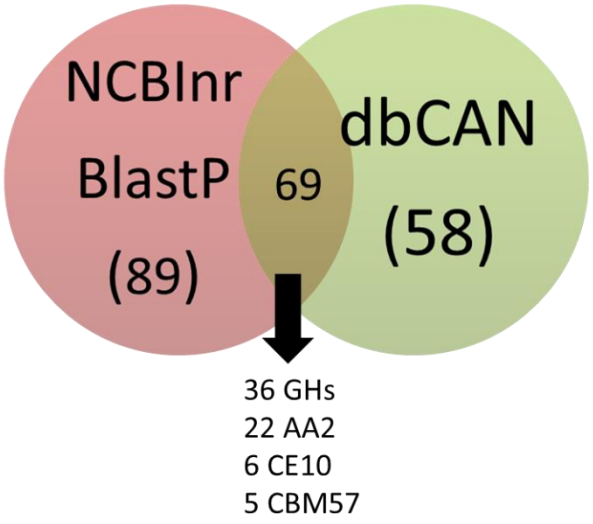


Figure 4.6 Venn diagram showing the results obtained for the protein annotation using dbCAN and BlastP platforms. In total, 216 putative CAZymes were identified from which 89 only appeared in BlastP annotation and 58 only in dbCAN. The 69 targets common to both annotation are as follows: 36 glycoside hydrolases (GHs); 22 Auxiliary activity enzymes from family 2 (AA2); 6 carbohydrate esterases from family 10 (CE10); 5 carbohydrate-binding modules from family 57 (CBM57).

As shown in figure 4.6, a total of 158 and 127 putative CAZymes were identified from BlastP and dbCAN annotations respectively; and 69 of them are common to both annotation platforms. Among the 69 common annotated proteins, 36 are GHs of different families, which are mainly responsible for the degradation of cellulose and hemicellulose main chains; 22 are AA2s, which typically are peroxidases or catalases and thus have an important role in the modification of lignin; 6

are CE10s, which according to the CAZy database currently represent a class of carbohydrate esterases whose members are active on non-carbohydrate substrates; and 5 are CBM57s, which are domains attached to different glycosidases that are enzymes responsible for the conversion of cellobiose into its final sugar, glucose. Furthermore, targets annotated only by dbCAN appear in the BlastP annotation as a "hypothetical protein", likely because only the top hits are taken into account for the annotation as manual inspection of the Blast results revealed. Among the 89 targets identified only by BlastP, 31 are similar to known CAZymes that for some reason were not identified using dbCAN, and the remaining targets are possible candidates with lignolytic activity (mostly superoxide dismutases [144] and peroxidases) that are not yet classified as CAZymes. A pie chart summarising all the different classes of putative CAZymes identified in this study is shown in figure 4.7.

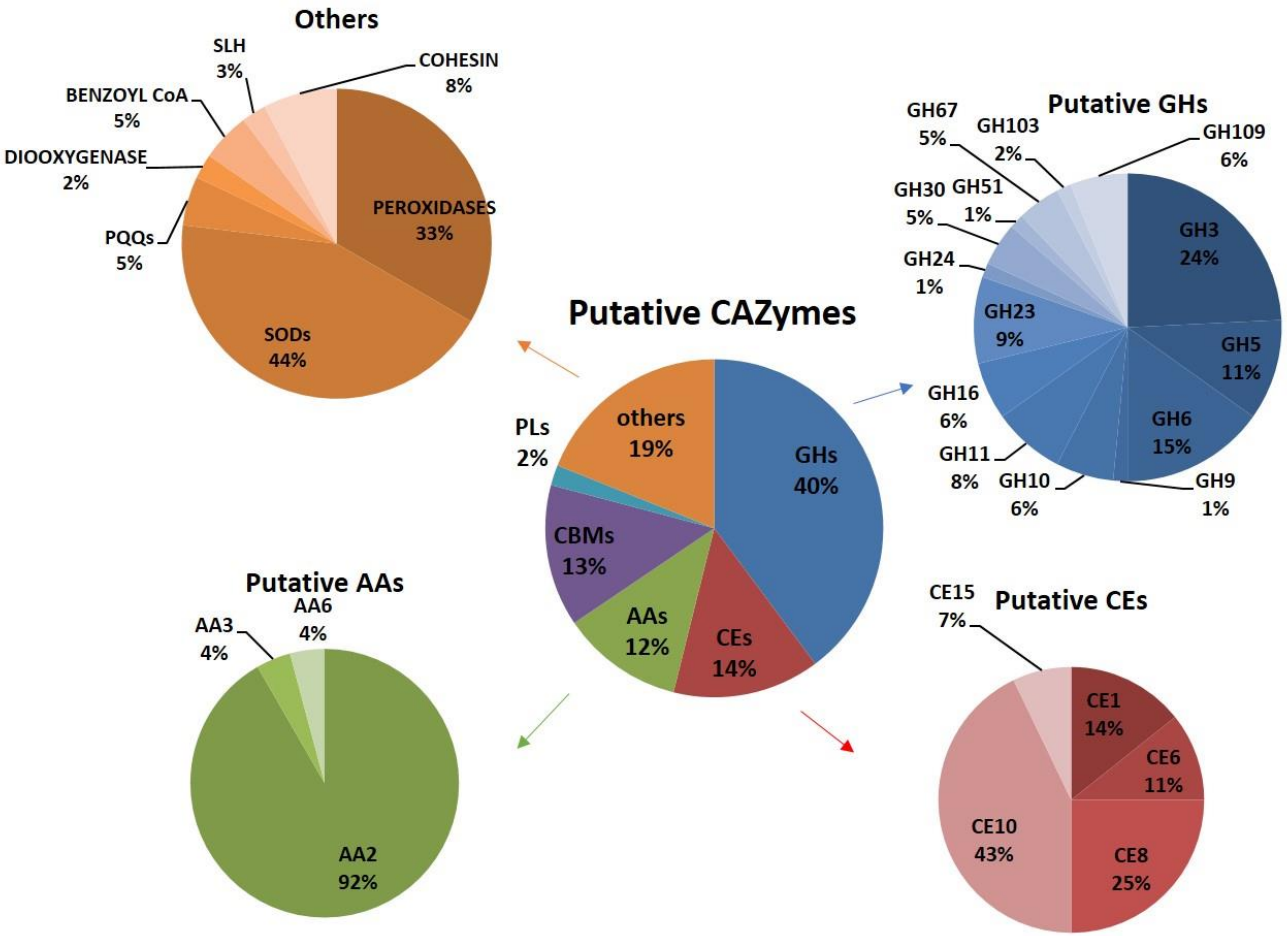


Figure 4.7 Pie chart containing all the putative CAZymes identified by dbCAN and BlastP. GHs represents the majority of the CAZymes identified (40%), followed by CEs (14%) and AAs (12%). "Others" represents the class of enzymes potentially related to lignocellulose degradation, but not yet classified as CAZymes. AA; auxiliary activity enzymes, GH: glycoside hydrolases, CE; carbohydrate esterases, SOD: superoxide dismutases, PQQ; Pyrroloquinoline quinone, and SLH: S-layer homology.

As apparent in figure 4.7, GHs are the most abundant class of putative CAZymes identified in this work (40%), which is expected as they are directly related to the degradation of cellulose and hemicellulose and their conversion into sugars. Interestingly, in this study we observed a notable abundance of CEs and AAs. The presence of these two classes of enzymes in this proportion (together they represent ~25% of all CAZymes), might suggest their importance as accessories enzymes in the degradation of the recalcitrant biomass. CEs are typically enzymes involved in deacetylation and disruption of ester linkages between lignin and polysaccharides, assisting in the degradation of lignocellulose [42, 43]. AAs are usually oxidases that can be related to the degradation/modification of lignin and cellulose, although no lytic polysaccharide monooxygenases were evident. In addition, a considerable high percentage of putative peroxidases and superoxide dismutases (SOD) among the 'others' fraction is observed, enzymes which can potentially be involved in lignin degradation/modification [144, 145].

In table 4.2 the top 25 putative CAZymes are listed according to their abundance in the proteome. Among them, seven were identified as potential CAZymes by dbCAN but not BlastP, and 6 were identified as such by BlastP, but not dbCAN. From the GHs present in the table, GHs 3, 5, 6, 9 and 16 are typically cellulases. GH5, 9 and 16 are often endoglucanases that act by cleaving the internal bonds in cellulose, making their ends accessible to cellobiohydrolases (usually GH6s) that processively release cellobiose [132]. GH3s usually encode glucosidases releasing glucose from cellobiose. GHs 10 and 11 are typically xylanases required to hydrolyse glucuronoarabinoxylan, the main hemicellulose in grass biomass [17]. GHs 23 and 103s are typically active on peptidoglycan present in bacterial cell walls and their presence may indicate bacterotrophic activity in the microbial community. Among the CEs identified, CE1s are typically feruloyl esterases (FAEs) or acetyl xylan esterases (AXE). FAEs are enzymes that act by cleaving ester bonds between arabinosyl residues present in hemicellulose and ferulic acid (linked either to lignin or another chain of hemicellulose) [23], and AXE acts removing the acetyl groups from hemicellulose side chains [146], and thus, they are enzymes that assist in the degradation of lignocellulose. CE8s are pectin methylesterases and might be associated with the pectin degradation observed in the biomass degradation data presented in the previous chapter. CE10s are a varied family of esterases that as mentioned before, according to the CAZy database, are currently classified as enzymes not related to carbohydrate degradation but their presence in the extracellular proteome suggests they may be involved in lignocellulose degradation in this case. CBM57s are Carbohydrate-Binding Modules found attached to various glucosidases and in the case of the contig identified in this table, were not associated to any other catalytic domain. As

CBMs are typically associated to other domains, this might either indicate an erroneous classification or a strong candidate that might be associated to a still unknown domain.

Table 4.2 Top 25 putative CAZymes according to their abundance in the proteome. E-value; expect value.

seqID	dbCAN annotation		BlastP annotation		Mol (%)
	Subject ID	e-value	Subject ID	e-value	
TRINITY_DN257988_c0_g1_i4_1	GH5	9.6E-40	glycosyl hydrolase family 5_53 domain-containing protein [Alteromonadaceae bacterium Bs02]	0	1.597601
TRINITY_DN241582_c0_g2_i3_1	GH6	2.3E-27	cellobiohydrolase [Teredinibacter turnerae]	5E-69	1.094571
TRINITY_DN144659_c0_g1_i1_1	GH103	2.1E-33	lytic transglycosylase [Congregibacter litoralis]	2E-38	0.665033
TRINITY_DN257539_c0_g1_i1_1	GH23	1.7E-23	hypothetical protein [Bacillus wakoensis]	3E-32	0.582568
TRINITY_DN241582_c0_g2_i5_1	GH6	3.3E-27	cellobiohydrolase [Teredinibacter turnerae]	8E-93	0.577413
TRINITY_DN50757_c0_g1_i1_1	GH11	1.9E-26	1","4-beta-xylanase [Luteimonas sp. J29]	8E-33	0.576786
TRINITY_DN195563_c0_g2_i1_1	---	---	glycosyl hydrolase [Synechococcus sp. WH 5701]	5E-07	0.576786
TRINITY_DN160714_c0_g3_i1_2	---	---	peroxidase [Gilvimarinus chinensis]	3E-108	0.519029
TRINITY_DN259889_c1_g2_i3_1	CE8	0.0000019	hypothetical protein"," partial [Gemmobacter nectariphilus]	3E-175	0.404147
TRINITY_DN154011_c0_g1_i1_1	GH16	3E-23	glycoside hydrolase family 16 [Teredinibacter sp. 1162T.S.0a.05]	1E-69	0.372733
TRINITY_DN219547_c0_g3_i1_2	CE1	2.2E-18	Poly(3-hydroxybutyrate) depolymerase [Verrucosispora sediminis]	9E-91	0.338237
TRINITY_DN259889_c1_g2_i1_2	CE8	4.2E-11	hypothetical protein"," partial [Gemmobacter nectariphilus]	2E-140	0.333595
TRINITY_DN257369_c0_g1_i1_1	GH6	2E-45	hypothetical protein [Marinimicrobium agarilyticum]	0	0.333106
TRINITY_DN157814_c0_g2_i1_1	GH3	1.4E-37	beta-glucosidase [Teredinibacter sp. 1162T.S.0a.05]	2E-77	0.332576

TRINITY_DN191910_c0_g2_i2_1	CE8	0.00003	Cadherin domain protein [Pirellula sp. SH-Sr6A]	2E-90	0.331981
TRINITY_DN161239_c0_g2_i1_1	GH10	7.4E-21	hypothetical protein [Marinomonas spartinae]	4E-37	0.331602
TRINITY_DN199787_c0_g4_i1_1	GH9	1.1E-09	hypothetical protein [Teredinibacter sp. 1162T.S.0a.05]	2E-69	0.302767
TRINITY_DN259889_c1_g2_i5_2	CE8	0.000021	hypothetical protein"," partial [Gemmobacter nectariphilus]	2E-143	0.29476
TRINITY_DN248762_c2_g1_i1_1	AA2	2.2E-13	catalase/hydroperoxidase HPI(I) [Sulfuricaulis limicola]	6E-96	0.255867
TRINITY_DN160550_c1_g1_i1_1	---	---	endoglucanase [uncultured bacterium]	2E-16	0.246602
TRINITY_DN260535_c0_g1_i4_3	---	---	superoxide dismutase [Oceanibulbus indolifex]	4E-139	0.224939
TRINITY_DN224411_c0_g1_i1_2	CE10	1.8E-17	esterase [Aestuariibacter aggregatus]	6E-167	0.21728
TRINITY_DN1073222_c0_g1_i1_1	---	---	carbohydrate esterase [Cellvibrio sp. OA-2007]	5E-36	0.201845
TRINITY_DN141852_c0_g1_i5_1	---	---	superoxide dismutase [Polaribacter sp. MED152]	3E-106	0.201845
TRINITY_DN140152_c0_g1_i1_1	CBM57	2.6E-29	alpha-N-arabinofuranosidase [Alteromonadales bacterium BS08]	1E-44	0.183888

4.3.6 Bacterial community profile

The results of DNA sequencing obtained from section 4.3.2 were analysed using bioinformatics tools (see Materials and Methods, section 2.4.3 for more details) for the construction of a bacterial community profile for the microorganisms that grew on the recalcitrant biomass. This community profiling was performed only for the final time point (8 weeks of incubation) and thus no analysis over time was conducted here. Instead, the main purpose of this experiment was to compare the community living on this very recalcitrant biomass with the putative CAZymes identified by the data of meta-transcriptomics and meta-proteomics. For the same reason, because the annotation given for those putative CAZymes identified were always from prokaryote origins, only analysis of 16S rRNA was performed and the results obtained are presented in figure 4.8.

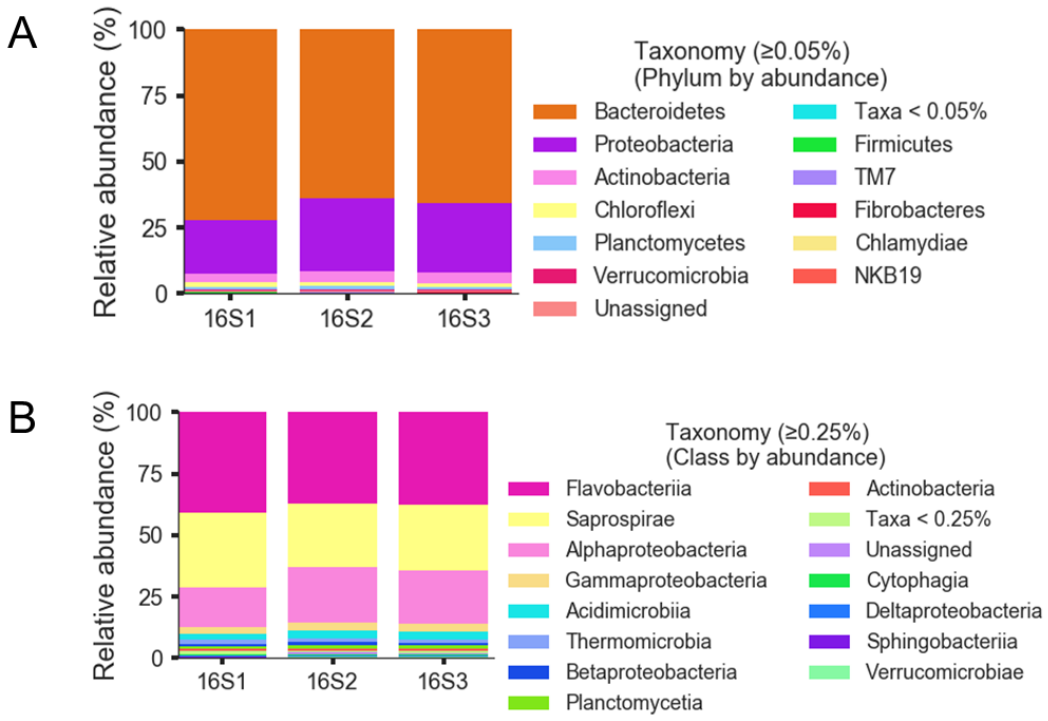


Figure 4.8 Bacterial community profile obtained from the data analysis of 16S rRNA. **A:** representation by phyla. **B:** representation by Class.

The results obtained for the elucidation of the bacterial profile (figure 4.8) show that there were 9 bacterial phyla (*Bacteroidetes, Proteobacteria, Actinobacteria, Chloroflexi, Planctomycetes, Verrucomicrobia, Firmicutes, Fibrobacteres* and *Chlamydiae*) and two phyla candidates (*TM7* and

NKB19) identified among the microorganisms that grew on the recalcitrant biomass, from which *Bacteroidetes* and *Proteobacteria* were the dominant phyla.

Bacteroidetes were the most abundant phylum identified in this study with organisms belonging to the *Flavobacteriia* class and to the *Saprospirae* class being the two most abundant. *Bacteroidetes* are a phylum of bacteria that contains a variety of organisms from anaerobic to aerobic environments and can be found in all ecosystems. The CAZymes produced by *Bacteroidetes* are typically arranged in a polysaccharide utilisation loci (PUL). The PUL is a set of genes linked to each other that typically are organised around a *sus*C/D gene pair, which are sequences encoding for transporters responsible for bringing polysaccharides into the periplasm of the cells, where the CAZymes can act free of competition [147]. Because of this organisation, *Bacteroidetes* are very well known for their production of CAZymes and have an evolutionary advantage in the degradation of complex polysaccharides compared to other microorganisms. Although *Bacteroidetes* were undoubtedly the most abundant phylum in the 16S rRNA analysis (over 70% of representatives), this was the second most abundant phylum present in the annotation of the putative CAZymes, behind *Proteobacteria.*

Proteobacteria were the second most abundant phylum identified in this study, with *Alphaproteobacteria* and *Gammaproteobacteria* being the two classes most abundant. *Proteobacteria* are gram-negative bacteria and they currently represent the most studied phylum of bacteria. It is in this phylum that is believed that mitochondria has evolved from, for its symbiosis with *Alphaproteobacteria* [148] and this phylum includes one of the most studied bacteria present in the world, *Escherichia coli* [149]. *Proteobacteria* are divided in six classes: *Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, Gammaproteobacteria,* and *Zetaproteobacteria,* and CAZymes belonging to all, but *Epsilonproteobacteria,* these classes were annotated in this study (table 4.4). *Alphaproteobacteria* are known for their higher plasticity, being found in diverse areas in the world either living alone as parasites or living in symbiosis [149]. They are also known for their abundance in marine ecosystems [150], but can also be found in soils and (in lower amounts) freshwater. *Gammaproteobacteria* are equally found in diverse ecosystems and, although in comparatively lower amounts, they are also abundant in marine ecosystems [149, 151].

The results for the likely phylogenetic origin of the 216 putative CAZymes identified by combined proteomics and transcriptomics analysis was compared with the microbial community profile. In this case, *Bacteroidetes* and *Proteobacteria* were also the most abundant phyla producing CAZymes, but with inverted results: 75% of the putative CAZymes identified belong to the *Proteobacteria* phylum, while nearly 15% belong to the *Bacteroidetes* phylum. The relatively low

amount of putative CAZymes identified belonging to the *Bacteroidetes,* when compared to its abundance among the microbial community, suggests that there may be several yet unknown CAZymes belonging to this phylum in this environment and they were not identified because the approach that we have used is based on sequence homology. On the other hand, the large abundance of putative CAZymes identified for the *Proteobacteria* might reflect the extensive knowledge and studies related to this phylum. The remaining putative CAZymes identified are distributed according to their phyla as shown in the table 4.3.

Table 4.3 List of the 216 putative CAZymes identified by combined meta-proteomics and meta-transcriptomics approaches, according to their phyla classification.

Phyla	Number of putative CAZymes identified	Types of putative CAZymes identified
Proteobacteria	162	See table 4.4
Bacteroidetes	32	See table 4.4
Actinobacteria	6	GH6, CE1, CE10, CE15, CBM6 and others
Firmicutes	6	GH23s and others
Chloroflexi	3	GH3 and GH5
Cyanobacteria	2	Others
Planctomycetes	2	AA2 and CE8
Verrucomicrobia	1	Other
Gemmatimonadetes	1	GH3
Unknown	1	AA2

In table 4.3 we can have a general idea of the distribution of the putative CAZymes identified among their phyla. CAZymes identified belonging to either *Proteobacteria* or *Bacteroidetes* phylum will be discussed below. Among the remaining phyla, *Actinobacteria* presented the most diverse type of CAZymes, varying from cellulases (GH6) to different families of esterases (CEs) and peroxidases (others). CE15 are glucuronoyl esterases, which are enzymes that cleave ester bonds connecting glucuronoyl residues of the hemicellulose and lignin [152]. Their presence associated with the presence of CE1, which as mentioned before are typically acetyl xylan esterases (AXE) or feruloyl

esterase (FAE) suggest that this phylum might produce enzymes related to the cleavage of the links between hemicellulose and lignin. All the GH23s identified in this study came from the *Firmicutes,* suggesting that members of this phylum survived in the culture by acting on peptidoglycan of other microorganisms instead of using biomass. *Chloroflexi* and *Gemmatimonadetes* only produced cellulases, suggesting that they either used the freed cellulose produced by other microorganisms or they also produce other CAZymes not identified in this study. *Planctomycetes* produced an interesting combination of putative CAZymes: AA2 that are catalases and/or peroxidases related to lignin modification; and CE8 that as mentioned before are enzymes active in pectins, suggesting that this phylum could be a potential candidate for production of lignocellulose-active enzymes related to the degradation of recalcitrant biomass. In the *Cyanobacteria* and *Verrucomicrobia* phyla, only putative enzymes not yet classified as CAZyme were identified. Also, an interesting observation is that although putative CAZymes were identified belonging to *Cyanobacteria* and *Gemmatimonadetes* phyla, none of them were recognised in the 16S rRNA profile, suggesting that either they are among the unassigned fraction in the community profile or they were wrongly annotated. Finally, the unknown phylum is due a putative AA2 annotated as from "uncultured bacterium".

Table 4.4 presents the distribution of the putative CAZymes belonging to *Bacteroidetes* and *Proteobacteria* phyla*,* according to their classes. Among the *Bacteroidetes*, the majority of CAZymes identified are from *Flavobacteriia* class (37.5%), while 31% are from unknown class and the majority of the CAZymes belonging to the *Proteobacteria*, are from *Gammaproteobacteria* class (60%) followed by *Alphaproteobacteria* (28%).

Table 4.4 Distribution of putative CAZymes belonging to Bacteroidetes and Proteobacteria phyla, according their class classification

Bacteroidetes		
Class	**Number of CAZymes identified**	**Type of CAZymes identified**
Flavobacteriia	12	GH3, GH109, CE8, CE10, AA3 and others
Unknown	10	CE8, CE10 and others
Cytophagia	4	CE8, AA2 and others
Sphingobacteria	3	GH109 and others
Saprospiria	2	CE10 and others

Chitinophagia	1	GH24
Proteobacteria		
Class	**Number of CAZymes identified**	**Type of CAZymes identified**
Gammaproteobacteria	98	GH3, GH5, GH6, GH9, GH10, GH11, GH16, GH67, GH103, CE1, CE6, CE10, AA2, CBM60, CBM57 and others
Alphaproteobacteria	45	GH3, GH5, GH51, CE1, CE8, CE10, AA2, AA6, PL1, PL6 and others
Deltaproteobacteria	14	GH3, GH11, GH30, CE1, CE15 and PL9
Betaproteobacteria	3	AA2 and others
Zetaproteobacteria	2	AA2

Although the *Bacteroidetes* were highly abundant in the microbial community, not as many putative CAZymes were identified as originating from this phylum. Also, as we can see in table 4.4, only a few CAZymes identified in this phylum are typically associated with lignocellulose degradation (GH3, CE8 and AA2).

The majority of the CAZymes identified in this study were unquestionably from *Proteobacteria*, especially the *Gammaproteobacteria* class. In this class, a wide variety of CAZymes were identified, from cellulases (GHs 3, 5, 6, 9, 16) and hemicellulases (GHs 10, 11 and 67), to diverse accessory enzymes (CE1, CE6, AA2) suggesting that this class of bacteria is well equipped for lignocellulose deconstruction. The second and third most abundant classes of this phylum, *Alphaproteobacteria* and Deltaproteobacteria, also provided a variety of CAZymes identified. In this case also including families of enzymes potentially related to the degradation of pectin, such as CE8, PL1 and PL9. *Betaproteobacteria and Zetaproteobacteria* classes were identified as producing peroxidases, suggesting they might have an important role in lignin degradation/modification.

Overall, the results obtained from the 16S rRNA shows a prevalence of two phyla among the community, which were also the biggest producers of CAZymes. However, it is difficult to make a deeper analysis of this data going further into genus and species for example, because the majority of the assignment returns as unknowns, which reflects how underexplored the saltmarsh environment

is. Due to the potential for discovery of new CAZymes presented by this environment, we decided to select some putative CAZymes for further studies, which are presented in the next section.

4.3.7 Selection of putative CAZymes for further study

Results presented in figures 4.6 and 4.7, and in table 4.2 were analysed for the selection of enzymes for further study. Selection criteria included abundance in the proteome, e-values and the identity percentage given by the annotations; and the apparent completeness of the sequence evidenced by the presence or absence of a stop codon to and at the end of the sequences. This analysis was performed using online bioinformatics tools of translation from Expasy (https://web.expasy.org/translate/). For the remaining targets, the presence or absence of a predicted signal peptide was investigated using webserver online tools of SignalP (http://www.cbs.dtu.dk/services/SignalP-3.0/). Because the mechanisms that some microorganisms use to secrete proteins are not completely understood and because secretion is not only restricted to the presence of a signal peptide, some targets without predicted signal peptides were also selected when the annotation suggested an interesting activity. Figure 4.9 shows examples for sequences that were selected or excluded and the reason for this choice. In total, 37 targets were identified and selected as putative CAZymes potentially involved in the degradation of lignocellulosic biomass, of which 28 have a signal peptide predicted. Among these targets are 17 glycoside hydrolases (GHs), 11 carbohydrate esterases (CEs), 4 auxiliary activity enzymes (AAs), one polysaccharide lyase (PL), two peroxidases and two putative metal depended hydrolase. Furthermore, from the 37 targets, only two have a putative CBM attached: target 13, a putative GH11 has a CBM60 also annotated, which according to the CAZy database are typically found associated with xylanases; and target 22, a putative CE1, has a CBM6 also annotated, which are CBMs whose function have been demonstrated as binding glucan, xylan and/or amorphous cellulose. The presence of this CBM associated with a CE, might indicate erroneous annotation or potentially a new activity for either the CE1 or for the CBM6. Table 4.5 contains all the selected targets, their annotation, e-values, identity, molarity in the proteome and additional information relevant to the next chapter. The cDNA sequence of each target was retrieved and used for cloning and heterologous expression, which will be detailed in the next chapter.

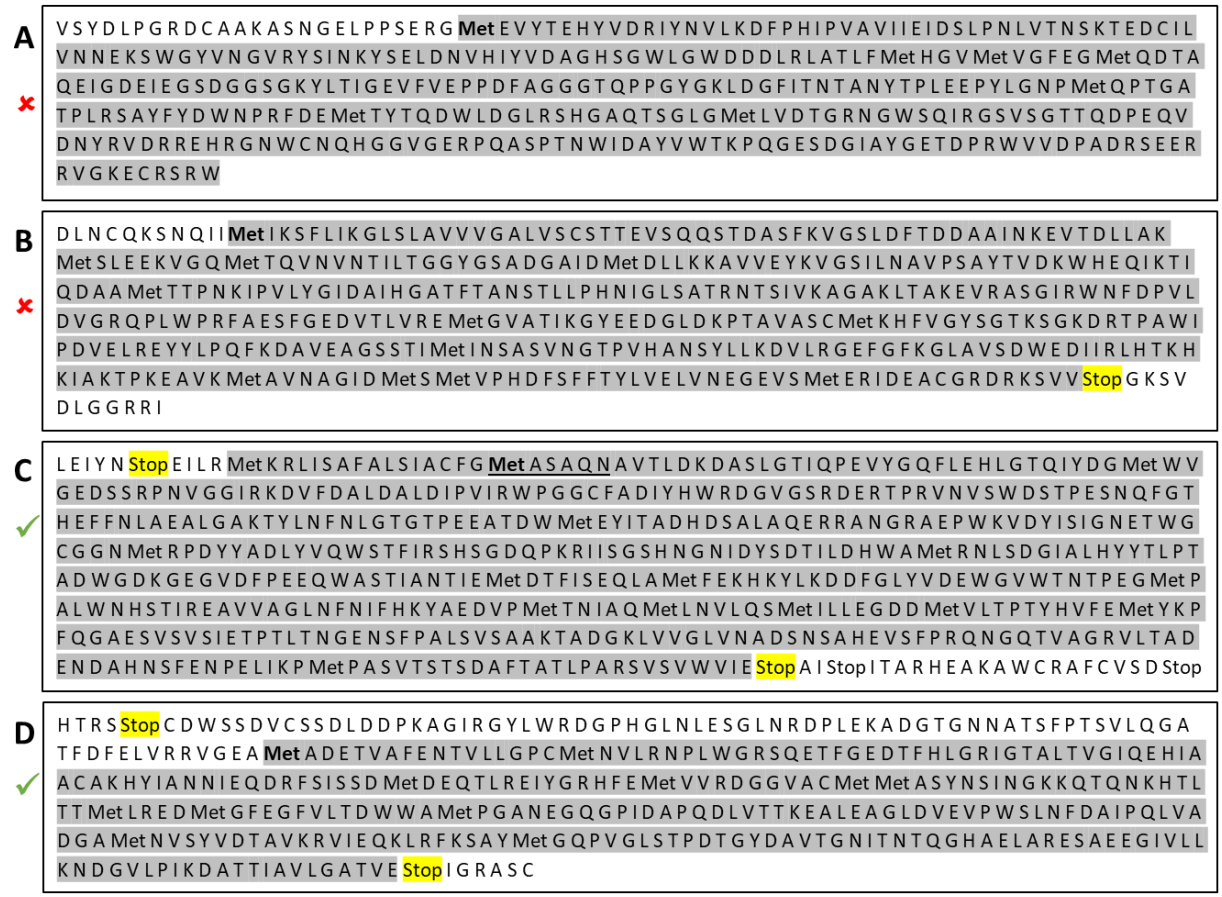


Figure 4.9 Examples of sequences that were selected or excluded from the annotations. Met is the abbreviation for the first methionine present in the sequence and Stop highlighted in yellow is the stop codon for each sequence. **A** and **B** are examples of sequences that were not selected: A does not have any stop codon in the sequence and although B has a stop codon in the end, the absence of it in the beginning of the sequence cannot assure that the methionine selected is actually the first one in that given sequence. **C** and **D** are examples of sequences that were selected: both sequences have stop codons prior and post the sequence of interest, suggesting that they are a complete ORF. C is an example of a sequence with a predicted signal peptide (underlined) and D is an example of sequence that was selected despite lacking a signal peptide.

Table 4.5 List of final 37 putative CAZymes obtained from the dbCAN and NCBInr annotations. The table shows the ID, its e-value and identity (for the NCBInr annotation), the molarity of each target in the proteome (Mol), the presence or absence of an identified polyserine sequence (PSL), the amount of predicted disulphide bonds and codon rare and the success or failure of cloning into expression vector. Targets presented in grey were excluded from the cloning attempts for reasons explained in the next chapter (section 5.3.1).

	dbCAN Annotation		NCBInr annotation			Signal peptide predicted?	Mol (%) in the proteome	Cloned in cytoplasmatic vector?	PSL?	Disulphide bounds predicted	Rare codon prediction (CAI*)
	Subject ID	e-value	Subject ID	Identity	e-value						
1	GH5	9.60E-40	glycosyl hydrolase family 5_53 domain-containing protein [Alteromonadaceae bacterium Bs02]	61-64%	0	✓	1.597601	✓	✓	3	0.72
2	GH6	4.40E-44	hypothetical protein [Marinimicrobium agarilyticum]	55-68%	2E-45	✓	0.333106	×	✓	---	---
3	GH5	3.60E-21	Cellulase (glycosyl hydrolase family 5) [Asticcacaulis taihuensis]	45%	7E-80	✓	0.166288	✓	×	1	0.73
4	CE6	3.40E-24	hypothetical protein [Gynuella sunshinyii]	52-57%	9E-95	✓	0.043862	×	✓	5	0.67
5	CE10	2.50E-14	Carboxylesterase type B [Bacteroidetes bacterium OLB9]	41-53%	0	✓	0.043782	✓	×	3	0.60
6	GH10	5.60E-75	hypothetical protein [Teredinibacter turnerae]	59%	2E-180	✓	0.043783	✓	✓	1	0.68

7	PL9	8.20E-06	hypothetical protein [Candidatus Desulfofervidus auxilii]	31-46%	4E-64	✓	0.039626	✓	×	4	0.66
8	GH51	4.10E-115	alpha-N-arabinofuranosidase [Parvularcula oceani]	55-62%	0	✓	0.028304	✓	×	1	0.71
9	GH3	1.60E-65	glycoside hydrolase family protein [Hyphomonas johnsonii MHS-2]	57-58%	0	✓	0.040001	✓	×	3	0.70
10	AA2	6.30E-19	peroxidase"," partial [OM182 bacterium BACL3 MAG-121001-bin29]	66-68%	6E-98	✓	0.130004	×	×	---	0.70
11	GH3	4.20E-68	beta-glucosidase [Wenyingzhuangia fucanilytica]	61-62%	0	✓	0.045870	×	×	4	0.61
12	CE1	4.20E-26	hypothetical protein [Teredinibacter turnerae]	56%	6E-86	✓	0.090005	✓	✓	6	0.68
13	GH11	9.40E-62	hypothetical protein [Teredinibacter turnerae]	72%	2E-164	✓	0.070058	×	✓	2	0.73
14	GH3	5.30E-64	hypothetical protein [Hyphomonas chukchiensis]	58-59%	0	✓	0.030456	✓	×	4	0.69
15	GH5	1.00E-46	hypothetical protein [Erythrobacter longus]	43-49%	4E-112	✓	0.084270	✓	×	1	0.71
16	CE10	5.80E-15	Carboxylesterase type B [Bacteroidetes bacterium OLB9]	50-71%	5E-81	✓	0.130179	×	×	---	---
17	---	---	putative metal-dependent hydrolase	78%	1E-55	✓	0.336909	×	×	---	---

			[Rhodobacteraceae bacterium HLUCCO07]								
18	GH3	2.90E-65	beta-glucosidase [Wenyingzhuangia fucanilytica]	64-66%	4E-155	✓	0.083680	×	×	---	---
19	---	---	putative metal-dependent hydrolase [Rhodobacteraceae bacterium HLUCCO07]	78-81%	1E-140	✓	0.540000	×	×	1	0.75
20	CE10	2.30E-15	hypothetical protein [Maricaulis sp. W15]	50%	0	✓	0.040658	✓	×	1	0.66
21	GH6	1.50E-61	hypothetical protein [Marinimicrobium agarilyticum]	60-68%	0	✓	0.140783	✓	✓	2	0.70
22	CE1	2.20E-18	Poly(3-hydroxybutyrate) depolymerase [Verrucosispora sediminis]	50-68%	9E-91	✓	0.338237	✓	×	4	0.70
23	GH10	7.40E-21	hypothetical protein [Marinomonas spartinae]	60-61%	4E-37	×	0.331602	×	×	---	0.70
24	GH109	3.20E-12	oxidoreductase [Sphingobacteriales bacterium BACL12 MAG-120802-bin5]	63-65%	6E-66	×	0.131588	×	×	1	0.67
25	CE15	1.90E-59	hypothetical protein [Streptomyces xinghaiensis]	49-51%	2E-73	×	0.080415	×	×	1	0.67
26	CE15	1.20E-44	hypothetical protein BE04_50575 [Sorangium cellulosum]	65-76%	4E-148	✓	0.065794	✓	×	4	0.67

27	GH3	2.60E-50	glycosyl hydrolase [Sandaracinus amylolyticus]	40-61%	4E-51	×	0.056609	×	×	1	0.69
28	GH3	1.40E-56	glycosyl hydrolase [Sandaracinus amylolyticus]	33-43%	2E-80	✓	0.051173	✓	×	4	0.67
29	CE10	8.80E-10	Carboxylesterase type B [Bacteroidetes bacterium OLB8]	37-52%	4E-126	✓	0.05457	✓	×	2	0.55
30	AA2	5.00E-15	catalase/peroxidase HPI [Zetaproteobacteria bacterium CG1_02_55_237]	86-87%	0	×	0.080758	×	×	1	0.74
31	GH67	4.90E-255	alpha-glucuronidase [Gynuella sunshinyii]	61-63%	0	✓	0.044271	×	×	1	0.67
32	---	---	peroxidase [Saccharophagus degradans]	90-91%	5E-130	×	0.310456	✓	×	1	0.71
33	---	---	peroxidase [Neptuniibacter caesariensis]	84-85%	3E-86	×	0.040457	×	×	1	0.67
34	CE1	1.50E-48	esterase [Asticcacaulis excentricus]	63-64%	9E-148	✓	0.025871	✓	×	1	0.70
35	AA3	5.30E-50	Choline dehydrogenase [Tenacibaculum sp. MAR_2009_124]	66-67%	0	×	0.041240	✓	×	3	0.63
36	AA6	2.60E-52	flavodoxin [uncultured bacterium]	83-86%	3E-124	×	0.140010	×	×	---	0.74

37	CE8	1.90E-06	hypothetical protein"," partial [Gemmobacter nectariphilus]	41-44%	3E-175	✓	0.404147	×	×	---	0.75

*CAI: codon adaptation index. An optimum CAI is the one equal to 1, but CAI > 0.8 can be considerable for the expression system.

Chapter 5 Cloning and heterologous protein production of selected putative CAZymes

5.1 Introduction

The previous chapter described the identification of 216 putative CAZymes in the secreted metaproteome from a community of marine microorganisms growing on recalcitrant biomass from a saltmarsh grass. From these 216, 37 candidates were selected for further analysis based on their abundance and putative activities. The annotation and selection of targets was made by taking into account their similarities with known sequences available in public database. However, even though this is a powerful method for selection of proteins of interest, the use of this approach does not guarantee that the selected targets will have the expected function since sequence similarity is not a guarantee of enzyme function and certainly not of specific enzyme characteristics. The prediction of protein activity is even more uncertain in the cases were small similarities were encountered, which could lead to mistaken annotation [153]. Therefore, in order to investigate the actual function of these putative CAZymes in the degradation of lignocellulose, it is necessary to experimentally assess the biochemical activity of these enzymes.

Heterologous protein expression is a commonly used technique for this purpose. In this technique, a host organism is chosen to express a protein of interest that is not normally produced by this organism. The use of this technique allows us to investigate the characteristics of target proteins without the need for protein extraction from the original host microorganism, which in this case would be problematic as this study was performed with a community of microorganisms. Different expression systems are available commercially nowadays and the right choice usually takes into account the main characteristics of the protein of interest, as for example the presence/absence of disulphide bridges, codon bias, the original host, post translation modifications (as glycosylation, for example) among others. Moreover, the host organism can also vary from more simple cells, as for example bacteria and yeast, to more complex organisms as fungi, insects and mammals. Bacterial recombinant expression system remains one of the most attractive systems to date due to its ease, simplicity, low cost, efficiency and potential to produce high levels of recombinant proteins [154-156]. In this study, because the annotation referred always to a prokaryote microorganisms, the host organism chosen was *Escherichia coli.*

5.2 Aims of the chapter

This chapter describes the cloning and sequencing of full length target sequences, as well as their subcloning into a suitable vector for expression in *E. coli.*

5.3 Results and discussion

5.3.1 Sequence analysis and preparation for cloning

Based on the results obtained by the comparison of proteomic and transcriptomic data, a list containing a total of 37 putative CAZymes was selected for cloning (table 4.5). After examination of the amino acid sequences of each of these targets, it was observed that targets 2GH6, 16CE10, 17hydrolase and 18GH3 (in grey in the table 4.5) were probably truncated versions of targets 21GH6, 5CE10, 19hydrolase and 11GH3, respectively and it was decided to only work with the longer forms. It was also observed that among the 37 selected targets, 7 of them (1GH5, 2GH6, 4CE6, 6GH10, 12CE1, 13GH11 and 21GH6) had as a particular feature, the presence of a polyserine chain in their peptide sequence. Although the occurrence of such repetitive sequences is not yet very well understood, its presence is usually associated to a region of linker between different domains of the protein, which hypothetically confers higher flexibility to the protein improving the interactions between protein and substrate, hence improving protein activity [157, 158]. Howard *et al.,* [159] investigated the presence of a polyserine linker (PSL) in 46 genes of the marine bacterium *Microbulbifer degradans* and found that all 46 genes are either proteins related to the degradation of carbohydrates or have a similar sequence to known carbohydrate degrading enzymes. In their studies, they showed that the PSLs present in those genes were responsible for connecting different functional domains of the proteins. Interestingly, in this work, a second domain related to CAZymes (for the targets with PSLs) was only identified for target 13GH11, where its sequence also encodes a putative CBM60, which are typically carbohydrate binding modules associated with xylanases. The analysis of the sequences for the targets 4CE6 and 6GH10 revealed the presence of malectin domain. Malectin is a membrane-anchored protein of the endoplasmatic reticulum that has revealed high similarity with CBMs of prokaryotes and is believed to be involved in N-glycosylation [160]. Although glycosylation occurs more often in eukaryotes, it is already known that bacteria are able to perform glycosylation [161] and the study of these targets could potentially help to understand their mechanisms to achieve it. In the sequences for the remaining 4 targets with PSLs (1GH5, 2GH6, 12CE1 and 21GH6) only a single domain was identified, suggesting that these linkers could be connected to a yet unknown functional domain revealing even more the potential for novelty that might be obtained from this present work. Although

these investigations could potentially lead to new findings, because of the lack of time, it was decided to focus mainly on the characterization of the putative CAZymes identified in this study, but we believe that a more careful study of these polyserine sequences might potentially return interesting results.

5.3.2 Cloning

In order to confirm the veracity of the assembled sequences it was necessary to amplify, clone and sequence the genes of interest and finally assess their activity by recombinant expression. Since the annotations given by BlastP always returned a microorganism that is prokaryote, and because the extraction of RNA from these samples was difficult, all the PCR reactions were performed using gDNA and 60% of the predicted CAZymes annotated (20 out 33 total targets) have been satisfactorily cloned. First forward and reverse primers exterior to the gene of interest were designed, which allowed the flexibility to design nested primers interior to this first sequence for nested PCR (see Materials and Methods, section 2.5.3 for more details). In the cases where the first PCR reaction had apparently failed, nested primers were used for a second PCR reaction using the product of the first reaction as a template. It is important to mention that this strategy was only possible because genes of interest were firstly cloned into a cloning vector, giving the versatility to design primers in any external area of the gene and not necessarily for the first amino acid of the protein of interest. After confirmation of the cloning by Sanger sequencing these genes were subcloned into an expression vector. This approach of first cloning into cloning vector has also been adopted because it provided the option to test different expression vectors. This technique proved to be particularly effective for the cases where the first reaction of PCR apparently failed or in cases where too many non-specific bands were present, as for clone 21GH6 and 8GH51 respectively, for example (figure 5.1 and 5.2).

As a first step in the cloning process, the mixture of total RNA/gDNA extracted previously (Materials and Methods, section 2.4.1) was first treated with RNAse A and the gDNA remaining was cleaned and concentrated using Genomic DNA Clean and Concentrator (Materials and Methods, section 2.4.2.1). Following this, PCR reactions were performed using gDNA as a template and the external cloning primers for each target (Materials and Methods, section 2.5.3). In the cases where no amplification product was observed or nonspecific bands were obtained, a new PCR reaction was performed using the first PCR product as a template with nested primers. PCR products were separated in agarose gel electrophoresis (1%), cleaned and purified (Materials and Methods, sections 2.5.4 and 2.5.5). These products were submitted to ligation and transformation using StrataClone Blunt PCR Cloning Kit (Materials and Methods, section 2.5.6). A few colonies of each target were

selected and submitted to a colony PCR (Materials and Methods, section 2.5.7) in order to identify positive clones, which had their plasmid DNA extracted and sent for Sanger sequencing (Materials and Methods, section 2.5.8). The results of the first PCR reactions are shown in the figure 5.1. All products of PCR, except the ones where a single band of amplification was observed, were submitted to the nested PCR, with results shown in figure 5.2. Finally, in figure 5.3 a few examples of the results obtained for the colony PCRs are shown.

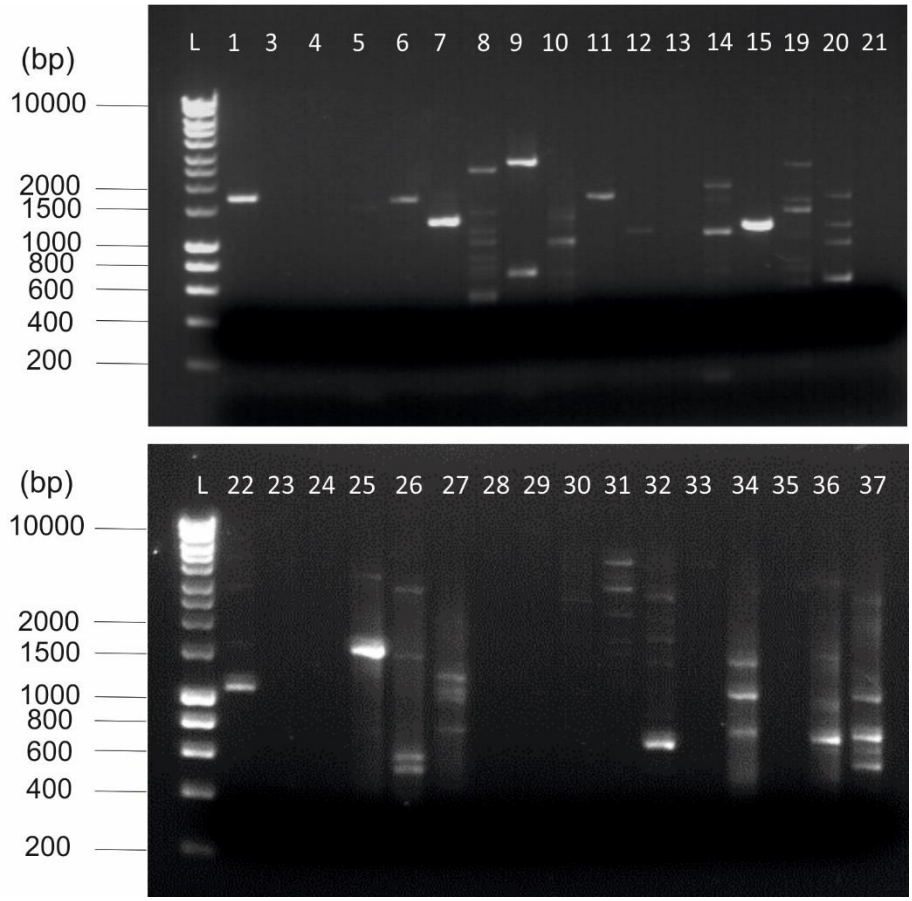


Figure 5.1 Products of PCR reactions obtained for each of the selected targets. The PCR reaction was performed for Tm = 60 °C, using gDNA as a template and the specifics external primers. **L**: HyperLadder I from Biolines. The expected theoretical sizes for each of the targets are as follow: **1** – 1GH5: ~1705 bp; **3** – 3GH5: ~1188 bp; **4** – 4CE6: ~1706 bp; **5** – 5CE10: ~1729 bp; **6** – 6GH10: ~1657 bp; **7** – 7PL9: ~1286 bp; **8** – 8GH51: ~1649 bp; **9** – 9GH3: ~2599 bp; **10** – 10AA2: ~721 bp; **11** – 11GH3: ~2210 bp; **12** – 12CE1: ~1102 bp; **13** – 13GH11: ~1300 bp; **14** – 14GH3: ~2617 bp; **15** – 15GH5: ~1166 bp; **19** – 19Hydro: ~826 bp; **20** – 20CE10: ~2060 bp; **21** – 21GH6: ~1998 bp; **22** – 22CE1: ~1221 bp; **23** – 23GH10: ~473 bp; **24** – 24GH109: ~536 bp; **25** – 25CE15: ~871 bp; **26** – 26CE15: ~1089 bp; **27** – 27GH3: ~915 bp; **28** – 28GH3: ~2616 bp; **29** – 29CE10: ~1733 bp; **30** – 30AA2: ~2241 bp; **31** – 31GH67: ~1959 bp; **32** – 32Peroxidase: ~705 bp; **33** – 33Peroxidase: ~540 bp; **34** – 34CE1: ~1102 bp; **35** – 35AA3: ~1913 bp; **36** – 36AA6: ~742 bp; **37** – 37CE8: ~5126 bp;

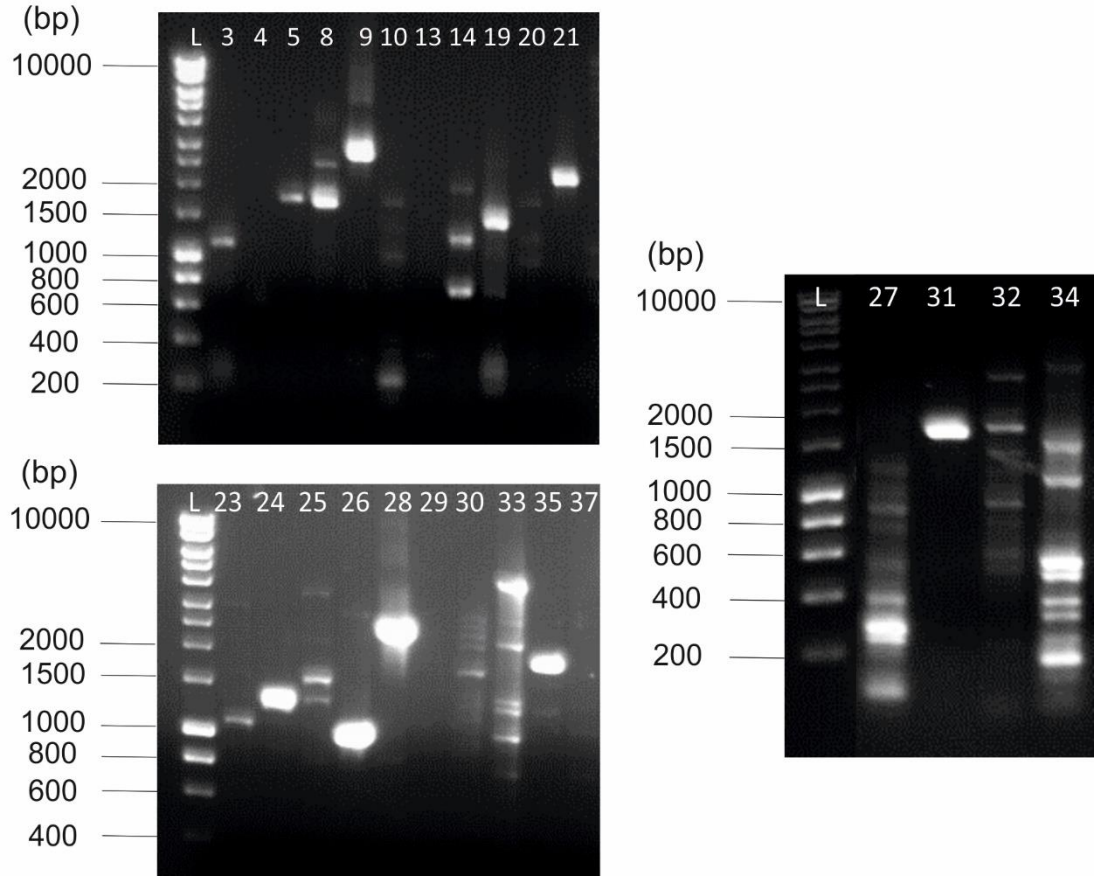


Figure 5.2 Products of Nest PCR reactions. The PCR reaction was performed for Tm = 60 °C, using 1 µL of the product of PCR from the previous reaction (figure 5.1) as a template and the specific nest primers. **L**: HyperLadder I – Biolines; **3** – 3GH5: ~1188 bp; **4** – 4CE6: ~1706 bp; **5** – 5CE10: ~1729 bp; **8** – 8GH51: ~1649 bp; **9** – 9GH3: ~2599 bp; **10** – 10AA2: ~721 bp; **13** – 13GH11: ~1300 bp; **14** – 14GH3: ~2617 bp; **19** – 19Hydro: ~826 bp; **20** – 20CE10: ~2060 bp; **21** – 21GH6: ~1998 bp. **23** – 23GH10: ~473 bp; **24** – 24GH109: ~536 bp; **25** – 25CE15: ~871 bp; **26** – 26CE15: ~1089 bp; **27** – 27GH3: ~915 bp; **28** – 28GH3: ~2616 bp; **29** – 29CE10: ~1733 bp; **30** – 30AA2: ~2241 bp; **31** – 31GH67: ~1959 bp; **32** – 32Peroxidase: ~705 bp; **33** – 33Peroxidase: ~540 bp; **34** – 34CE1: ~1102 bp; **35** – 35AA3: ~1913 bp; **37** – 37CE8: ~5126 bp; - The use of this strategy (nest primers + PCR product of the first reaction) is particularly interesting in cases when the first reaction apparently doesn't work (targets 3GH5, 5CE10, 21GH6, 23GH10, 24GH109, 28GH3, 33Peroxidase and 35AA3) and/or in case where many unspecific bands are present (targets 8GH51, 9GH3, 19Hydro, 26CE15, 27GH3, 31GH67, 32Peroxidase and 34CE1).

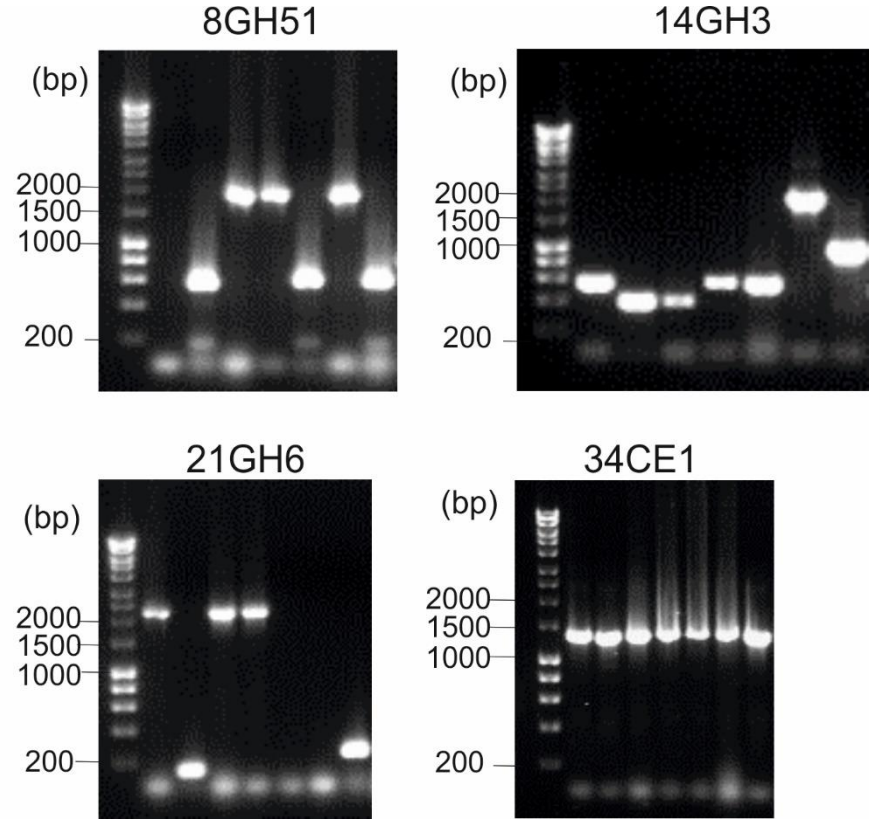


Figure 5.3 Products of colony PCR for the targets numbered 8GH51, 14GH3, 21GH6 and 34CE1. Seven colonies of each target were selected and submitted to colony PCR. As it is observed, three colonies selected for targets 8GH51 (~1649 bp), one colony for the target 14GH3 (~2617 bp), three colonies for the target 21GH6 (~1998 bp) and all the colonies for the target 34CE1 (~1102 bp) were positives (confirmed by sequencing).

In total, 20 targets were satisfactorily cloned, from which nine are glycoside hydrolases (1GH5, 3GH5, 6GH10, 8GH51, 9GH3, 14GH3, 15GH5, 21GH6 and 28GH3); seven are carbohydrate esterases (5CE10, 12CE1, 20CE10, 22CE1, 26CE15 29CE10 and 34CE1); two are auxiliary activities enzymes (35AA3 and 36AA6); one is polysaccharide lyase (7PL9) and one is peroxidase (32Peroxidase). All these clones were used as template for the subcloning steps into expression vector.

5.3.3 Subcloning into expression vector pet52b+

Targets that were satisfactorily cloned from the previous step were used as a template and subcloned into a chosen expression vector. As mentioned before, it was decided to work with bacterial system of expression as a host organism and *Escheria coli* was selected as the host cells for the trials of expression. For the expression vector, pet52b+ (from Novagen) was chosen because it is a well

known and available vector used in our lab, it is compatible with bacterial expression and it has tags on its sequences, which aid towards the purification of the proteins expressed, as well as help in the identification of the target proteins under analysis on Western Blot. Moreover, since pet52b+ is a cytoplasmatic expression system, primers were designed eliminating the native signal peptide (in case of its presence) and ensuring that the sequence of the protein would be in frame with the remaining vector sequence. Subcloning was performed using Infusion technology with the expression vector pet52b+ (see Materials and Methods, section 2.5.11 for more details). All the steps of PCR reactions for linearization of the vector and for insert preparation are shown in figure 5.4. Products of PCR were cleaned, purified and submitted to ligation, which was performed using 1:2 of inserts and linearized vectors in a mix containing the In-Fusion HD enzyme following the manufacturer's instructions. Ligation was transformed into Stellar competent cells (from Clontech) and a new colony PCR reaction was performed for a few selected colonies (data not shown). Positive results were sent for sequencing, which confirmed the success of the subcloning for all targets but clone 36AA6. Because of time constraints, it was decided to carry over the work with the positive clones obtained and to leave 36AA6 behind.

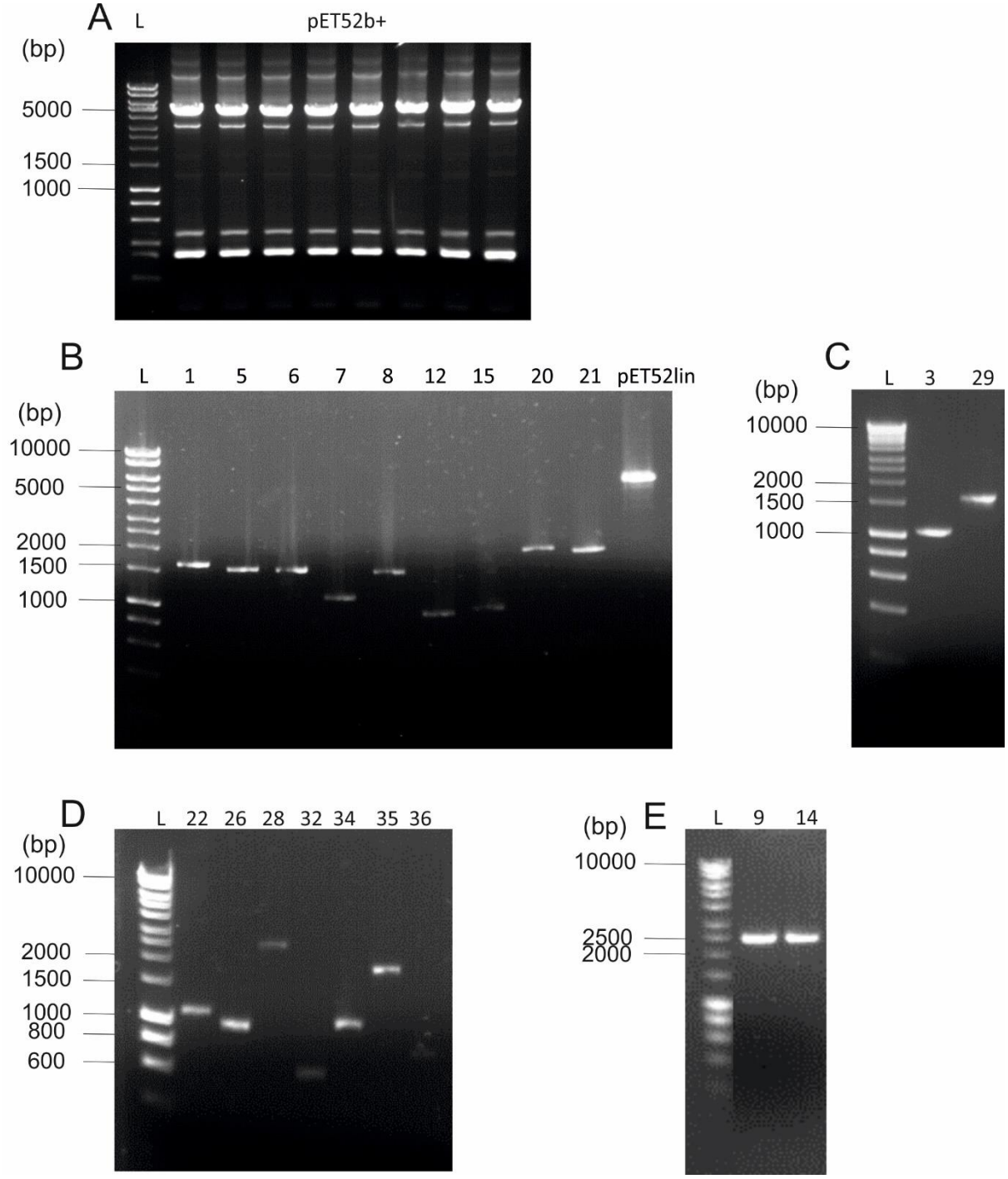


Figure 5.4 Products of PCR obtained for the subcloning steps**. A** – PCR products for the linearization of the vector pet52b+; **B, C, D** and **E –** Products of PCR for the linearization of each target after cleaning and purification. **L**: HyperLadder I – Biolines; **pET52lin**: pet52b+ linearized after cleaning steps ~5000 bp; **1** – 1GH5: ~1705 bp; **3** – 3GH5: ~1188 bp; **5** – 5CE10: ~1729 bp; **6** – 6GH10: ~1657 bp; **7** – 7PL9: ~1286 bp; **8** – 8GH51: ~1649 bp; **9** – 9GH3: ~2599 bp; **12** – 12CE1: ~1102 bp; **14** – 14GH3: ~2617 bp; **15** – 15GH5: ~1166 bp; **20** – 20CE10: ~2060 bp; **21** – 21GH6: ~1998 bp; **22** – 22CE1: ~1221 bp; **26** – 26CE15: ~1089 bp; **28** – 28GH3: ~2616 bp; **29** – 29CE10: ~1733 bp; **32** – 32Peroxidase: ~705 bp **34** – 34CE1: ~1102 bp; **35** – 35AA3: ~1913 bp; **36** – 36AA6: ~742 bp; Target 36AA6 was the only target where the subcloning has failed.

5.3.4 Recombinant protein production

Recombinant protein expression for the targets in this study proved to be challenging. Expression trials were performed for all 19 targets that were satisfactorily subcloned into expression vector pet52b+, but only five of them had visible soluble protein when analysed by Western Blots (WB). In order to select which strains of *E.coli* to be used, predictions of possible formation of disulphide bridges was made using DIANNA web server (http://clavius.bc.edu/~clotelab/DiANNA/) and it was observed that potential disulphide bonds formation vary among the targets from none up to six (table 4.5). Also, analysis of the DNA sequences was performed using Rare Codon Analysis Tool from GenScript webserver (https://www.genscript.com/tools/rare-codon-analysis), which revealed that the sequences analysed had high amounts of rare codons present (table 4.5, none of the CAI were suitable for the host cells). Taking these predictions into account, it was decided to use Rosetta-gami 2 (DE3) cells, from Novagen as the expression host, because these cells have a less reducing cytoplasm to encourage disulphide bridges formation and are enhanced to express proteins containing codons rarely used in *E. coli*. Therefore, expression trials were performed for a variety of conditions (see Materials and Methods, section 2.6.2 for more details). After the final time of expression, samples were lysed and an aliquot of soluble and insoluble fractions was analysed by WB (Materials and Methods, section 2.6.4); however no satisfactory results were obtained for any of the targets in any of the conditions and medium tested (data not shown). Expression was therefore tried in selected ArcticExpress (DE3) cells from Agilent Technologies because they grow and express at lower temperature, 30 °C and 5-20 °C respectively, which can improve protein folding and also because they express chaperones that can improve the solubility of the target protein. New expression trials were performed using LB (and 0.5 mM of IPTG) and auto induction (AI) mediums at 16 °C overnight. Cells were once again lysed and soluble and insoluble fractions were analysed by WB. Results of WB obtained for the AI medium, which had slightly higher levels of expression than LB, are shown in figure 5.5.

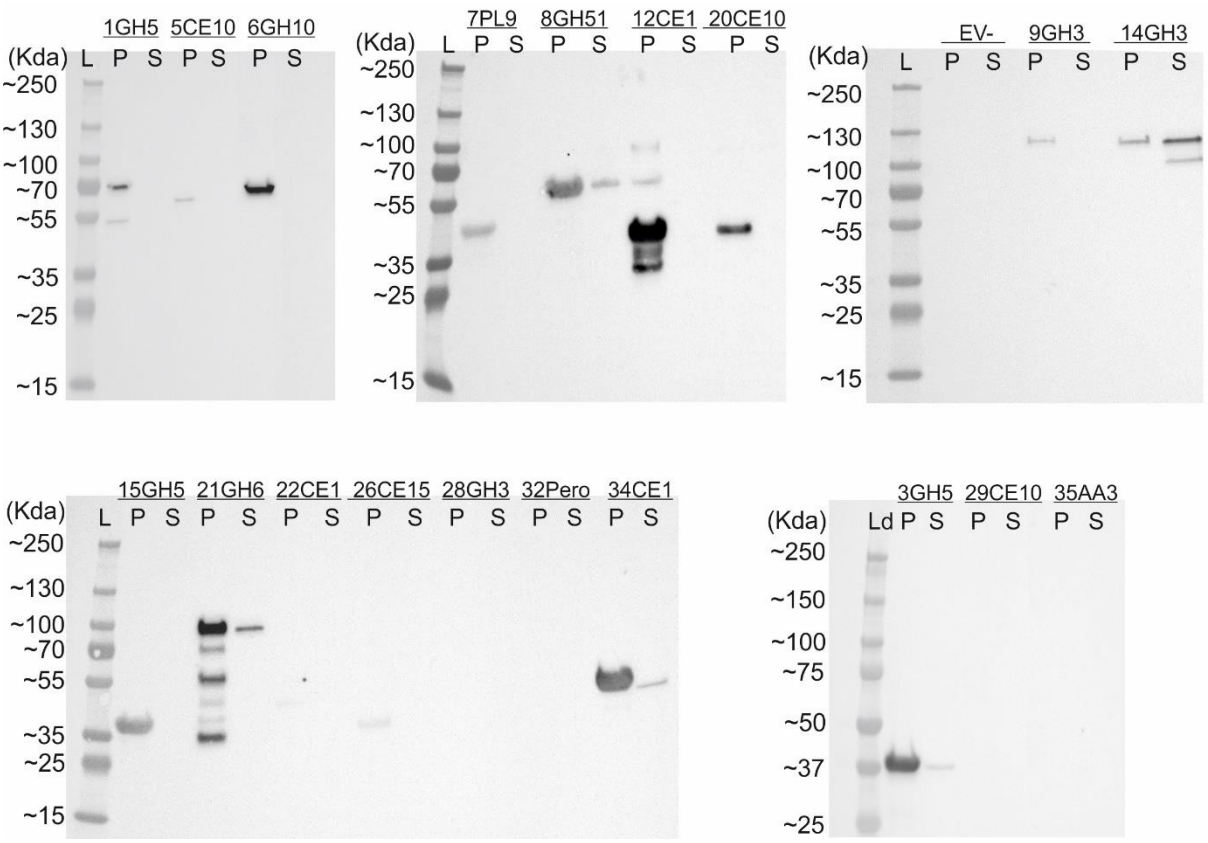


Figure 5.5 Chemiluminescent immunoblot results for the protein expression using AI medium. **L:** PageRuler Plus Prestained Protein Ladder from Thermo Scientific; **Ld:** Precision Plus protein standards from Bio-Rad; **EV-**: empty vector, pet52b+; P and S are the insoluble and soluble fraction of each target respectively. The expected sizes for each target tested are as following: **1GH5** ~60KDa, **3GH5** ~36.2KDa, **5CE10** ~57KDa, **6GH10** ~57KDa, **7PL9** ~42KDa, **8GH51** ~57KDa, **9GH3** ~88KDa, **12CE1** ~33KDa, **14GH3** ~88KDa, **15GH5** ~39KDa, **20CE10** ~72KDa, **21GH6** ~67KDa, **22CE1** ~42KDa, **26CE15** ~37KDa, **28GH3** ~93KDa, **29CE10** ~56.9KDa, **32Pero** ~25KDa, **34CE1** ~37KDa, and **35AA3** ~68.3KDa.

Results obtained by WB (figure 5.5) analysis reveal that although ArcticExpress (DE3) cells have performed better than Rosetta-gami 2 (DE3) (presence of soluble expression for targets 3GH5, 8GH51, 14GH3, 21GH6 and 34CE1), most of the targets (1GH5, 5CE10, 6GH10, 7PL9, 20CE10, 9GH3, 15GH5 and 22CE1) only had visible bands for the insoluble fraction and targets 28GH3, 29CE10, 32Pero and 35AA3 did not display any levels of expression (repeating the results previously observed for Rosetta-gami 2 cells). Figure 5.5 also reveals that target 20CE10 had a single band in the insoluble fraction that is lower and differs from the expected size (~72KDa), which could indicate proteolysis. It was also observed that targets 12CE1 and 21GH6 had several bands present in the WB analysis for the insoluble fraction, which could also be due to proteolysis, s both targets have a polyserine chain in their

sequence, this could potentially indicate an artefact caused by the polyserine chain. Finally, targets where soluble expression was observed (3GH5, 8GH51, 14GH3, 21GH6 and 34CE1) were expressed in bigger volumes (500 mL of medium) and subject to protein purification.

5.3.5 Protein purification

In order to assess the enzymatic activity of the soluble proteins expressed, it was first necessary to purify these proteins. The expression vector used in this work, pet52b+, has a sequence that encodes a streptavidin II in the N-terminus of the protein of the interest, which makes it possible to use affinity chromatography for protein purification. Thus targets 3GH5, 8GH51, 14GH3, 21GH6 and 34CE1 were expressed in 500 mL of AI medium at 16 ℃ overnight, followed by lysis and purification. Protein purification was carried out by affinity chromatograph using StrepTrap HP column (GE Healthcare Life Science) and ÄKTA start (GE Healthcare Life Science) following the manufacturer's instructions and the protein elution was performed using the same buffer of lysis with the addition of 2.5mM of d-desthiobiotin (Sigma-Aldrich) (see Materials and Methods, section 2.7.1 for more details). Although soluble protein (in the expression trials) was observed in WB for targets 3GH5 and 21GH6, these targets failed in the purification (data not shown). Once target 3GH5 exhibited only a slight visible band in WB for the soluble fraction and considering that WB analysis is very sensitive, it is likely that not enough soluble protein was obtained for its purification. On the other hand, even though target 21GH6 had a more intense band for the soluble fraction when analysed by WB, the reason for failing in the purification might be that the conformation of the enzyme could obscure the streptavidin tag, making it not accessible to interact with the StrepTrap column, or it may be that the polyserine chain might interfere with the purification. An aliquot of each step of the purification (lysate, flow through and elution) for targets 8GH51, 14GH3 and 34CE1 was taken and analysed by SDS-PAGE (Materials and Methods, section 2.6.3), which is shown in figure 5.6.

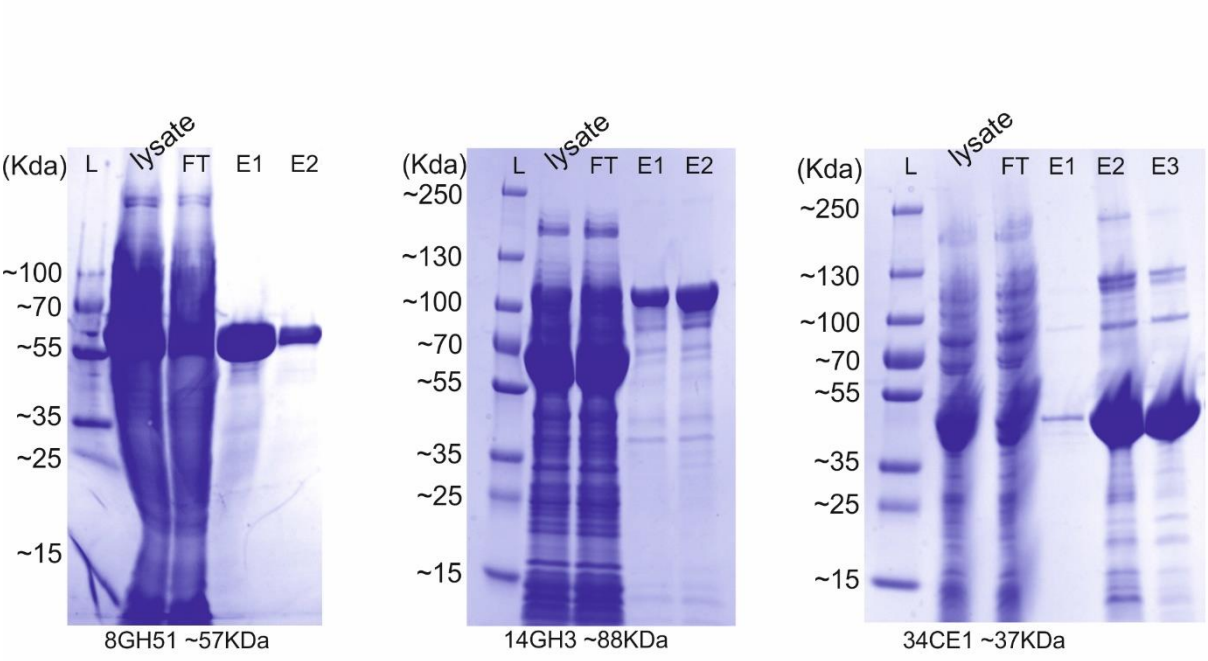


Figure 5.6 SDS-PAGE for each step of the purification by affinity chromatograph. **L:** PageRuler Plus Prestained Protein Ladder from Thermo Scientific; **lysate:** aliquot of the sample prior to application in the column; **FT:** flow through after the application of the lysate into StreTrap HP column; **E**1, **E2** and **E3** represents each of the elution collected. The expected size for each protein is as shown on the figure.

Once elution obtained for clone 34CE1 still had the presence of intense contaminants bands, this clone was submitted to a further clean up step using gel filtration chromatography (Materials and Methods, section 2.7.3). Elution fractions for target 34CE1 were concentrated (Materials and Methods, section 2.7.2), pooled together and the results obtained are shown in figure 5.7.

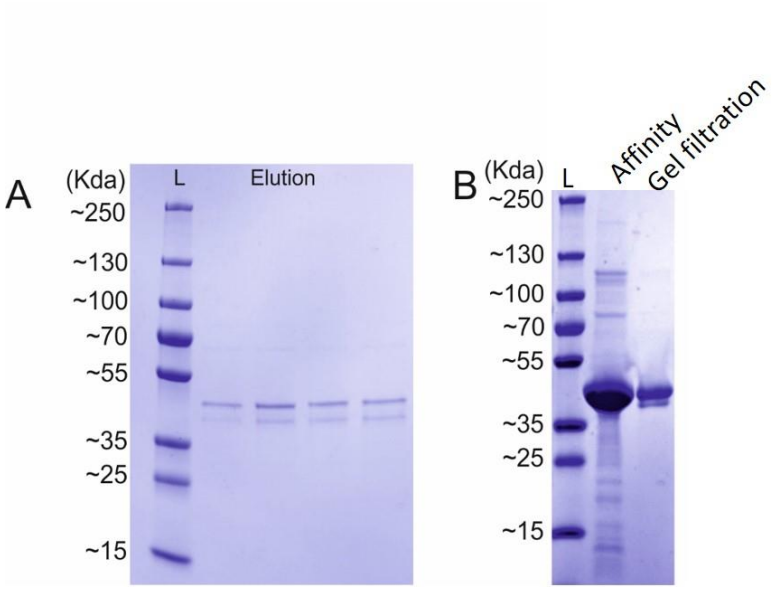


Figure 5.7 SDS-PAGE for steps of purification for target 34CE1. **L:** PageRuler Plus Prestained Protein Ladder from Thermo Scientific; **A** shows examples of the elution obtained for 34CE1 after the gel filtration. These samples were concentrated and pooled together. **B** shows a comparison from purification obtained by affinity chromatography and gel filtration after samples were concentrated. The expected size for this protein is ~37KDa.

As a final step in this project, targets 8GH51, 14GH3 and 34CE1 were tested for a range of different substrates and activity characterization (pH, temperature and salt tolerance) was performed on the substrate where positive activity was observed. All the steps and results obtained for these experiments are detailed in the next chapter.

Chapter 6 Enzyme characterisation, influence of seawater and influence of salt concentration

6.1 Introduction

The previous chapter explained the process for cloning and recombinant expression of the selected targets identified in this work. After considerable challenges encountered during recombinant protein production, three proteins were obtained in a soluble form. The putative activity given for each of these targets was based on similarity with known sequences available in public database and this work would not be complete without the experimental investigation of the actual activities exhibited by these proteins. Clones 8GH51, 14GH3 and 34CE1 are the targets being investigated in this chapter and according to their annotation they are putative glycoside hydrolases from family GH51 and from family GH3, and a putative carbohydrate esterase from CE1, respectively.

Glycoside hydrolases (GH) are a wide group of enzymes that catalyse the hydrolytic cleavage of glycosidic bonds and have been assigned to more than 100 different families based on sequence similarity [39]. According to the CAZy database, GHs from family 51 are a relatively small group and enzymes belonging to this group are usually endoglucanase, endoxylanases, β-xylosidases or arabinofuranosidases. In contrast, GH3 is a larger family that typically comprises β-glucosidases, β-xylosidases, arabinofuranosidases or exo-glucanases. On the other hand, carbohydrate esterases (CE) are a smaller group of enzymes that according to CAZy database are currently divided into 16 different classes. These enzymes remove ester linked substitutions from polysaccharides and could potentially help towards lignocellulose degradation by acting on the bonds between side chains of hemicellulose and lignin. Such an action might serve to increase the accessibility of lignocellulose-active GHs to their substrates [23, 70, 162] without being directly active in the polysaccharide chain degradation. CEs from family 1 are typically acetyl xylan esterases or feruloyl esterases, which act by removing acetyl and ferulic acid (and/or coumaric acid) groups from the side chains of the xylan backbone [70, 163]. This chapter will describe the activity tests that have been performed for these targets in order to understand their role in the degradation of lignocellulose and in order to assess their salt tolerance against NaCl and seawater.

6.2 Aims of the chapter

The main aim of this chapter is to characterise the activity of three recombinant enzymes from the saltmarsh biomass degrading community and assess their responses to temperature, pH and salinity. This data will help identify the roles of these enzymes in biomass degradation and indicate their potential industrial relevance.

6.3 Results and discussion

Based on the annotation provided by dbCAN and BlastP, a range of different model substrates were selected and activity tests were performed for the putative GH51, GH3 and CE1. Preliminary tests were performed by incubation of each substrate with each individually purified protein at 30 ℃ overnight aiming to detect which case would return positive results (table 6.1). The final activity was measured spectrophotometrically upon the release of 4-nitrophenyl (pNP) (see Materials and Methods, section 2.8.1 for more details). Once the most appropriate model substrate for each enzyme had been identified, a more detailed characterisation (pH, temperature and salt tolerance) was performed (Materials and Method, sections 2.8.3 to 2.8.5) and the results obtained are presented below.

Table 6.1 Preliminary enzymatic activity tests performed for each target using pNP substrates. ✗ no activity was detected for that substrate, ✓ and ✓✓ indicates low and high activity, respectively.

Substrate tested	GH51	GH3	CE1
pNP-α-L-arabinofuranoside	✓✓	✗	✗
pNP-α-L- rhamnopyranoside	✗	✗	✗
pNP-β-L- fucopyranoside	✗	✗	✗
pNP-α-D- xylopyranoside	✗	✗	✗
pNP-β-D- xylopyranoside	✗	✓	✗
oNP-β-D- xylopyranoside	✓	✗	✗
pNP-α-D- manopyranoside	✗	✗	✗
pNP-β-D- manopyranoside	✗	✗	✗
pNP-β-D- glucopyranoside	✗	✓✓	✗
pNP-β-D- galactopyranoside	✗	✗	✗
pNP-α-D- galactopyranoside	✗	✗	✗
pNP Acetate	✗	✗	✓✓

6.3.1 Characterisation of a putative GH51 - clone 8GH51

Preliminary activity tests (table 6.1) showed that putative GH51 showed highest activity on 4-nitrophenyl-α-L-arabinofuranoside (pNP-Ara) and lower activity on ortho-nitrophenyl β-D-xylopyranoside (oNP-βXyl), suggesting that the enzyme is an arabinofuranosidase (AFase). Subsequent experiments were performed with pNP-Ara. For the determination of the optimum pH, a range from pH 3 to 10 (McIlvaine buffer for the range of pH 3-7; Tris-HCl for the range of pH 7-9; and Glycine-NaOH for pHs 9 and 10) was tested (Materials and Methods, section 2.8.2). For the determination of optimum temperature, a range from zero to 80 ℃ was tested using the optimum pH obtained previously (Materials and Methods, section 2.8.3). In order to investigate the influence of seawater on the enzyme activity, the same range of temperatures were tested using artificial seawater in place of the buffer (Materials and Methods, section 2.8.4). Finally, to test the salt tolerance of the enzyme, the same range of temperatures was used against different concentrations of buffered sodium chloride (NaCl) (Materials and Methods, section 2.8.5). The results of these experiments are presented below where “relative activity” is the enzymatic activity obtained related to the maximum activity observed. In figure 6.1 it is shown that the putative AFase GH51 has an optimum pH for activity between 6 and 7.

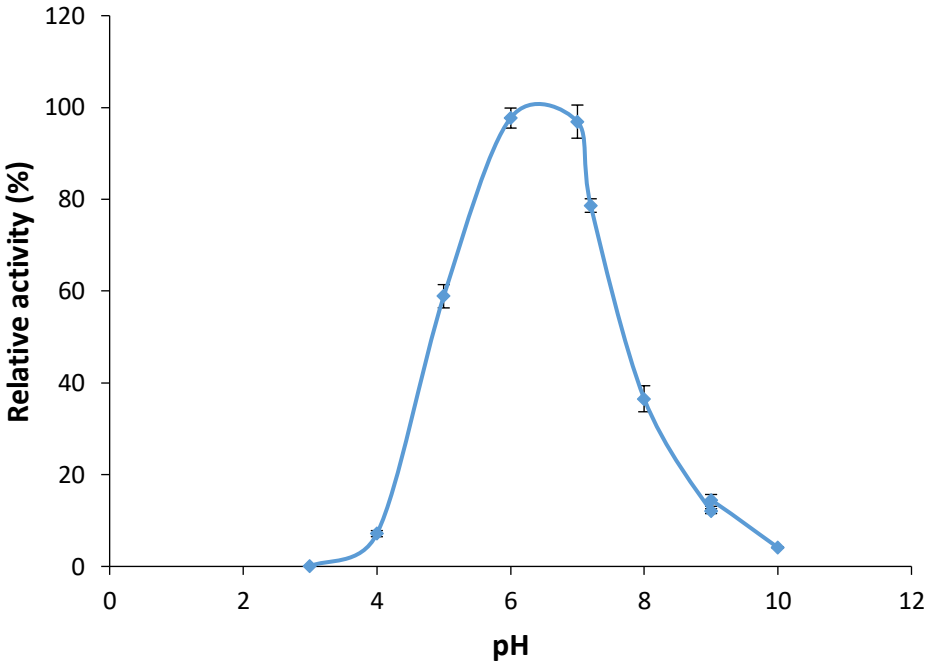


Figure 6.1 Determination of optimum pH for the putative AFase GH51. Activity tests were performed using pNP-Araf as substrate and the reaction was performed at 30 ℃ for 30 min. Data are averages of five assays, and the bars represent standard errors.

The results in figure 6.2 reveal that the temperature optimum in low salt conditions is around 40 ℃, whereas in seawater there is a double peak of activity at 30 ℃ and 50 ℃, with a small trough at 40 ℃. It is also evident that activity was significantly higher in seawater at all temperatures except 40 ℃. This may suggest that the enzyme is either more active, or more stable, in seawater. A curious observation was that this enzyme remains active at low temperature, with almost 40% of the maximal activity seen at 0 ℃ in seawater. This low temperature activity may reflect the environment from which the microbial inoculum used in this study was obtained. Saltmarshes in Northern England can experience periods of low temperature during the winter and are less protected from these extremes than pelagic microbes. Thus, there may be benefits from having enzymes able to operate at low temperatures in the saltmarsh sediment environment. This low temperature tolerance might prove to be of some biotechnological interest.

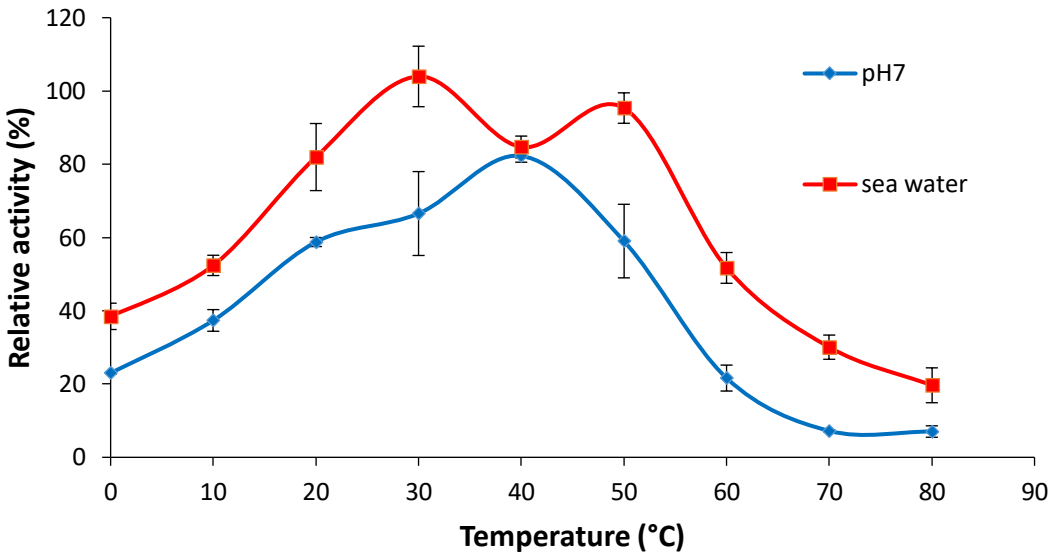


Figure 6.2 Determination of optimum temperature and influence of seawater on the activity of the putative AFase GH51. Activity tests were performed using pNP-Araf as substrate and the reaction was performed in the presence of the buffer at pH 7 (or seawater), for 30 min. Data are averages of five assays, and the bars represent standard errors.

The effects of increasing NaCl concentration (figure 6.3) show that the enzyme is generally more active in higher salt conditions, but the double peak of activity seen in seawater is not evident in NaCl. Seawater typically contains around 0.6 M NaCl, along with many other ionic species and it is likely that some of these are influential on protein activity. Another interesting observation in figure 6.3 is that AFase GH51 exhibits tolerance to NaCl concentrations well above those seen in seawater. Even concentrations of NaCl as high as 3M are not inhibitory. Indeed, the highest activity is seen in 2 or 3 M compared to lower concentrations.

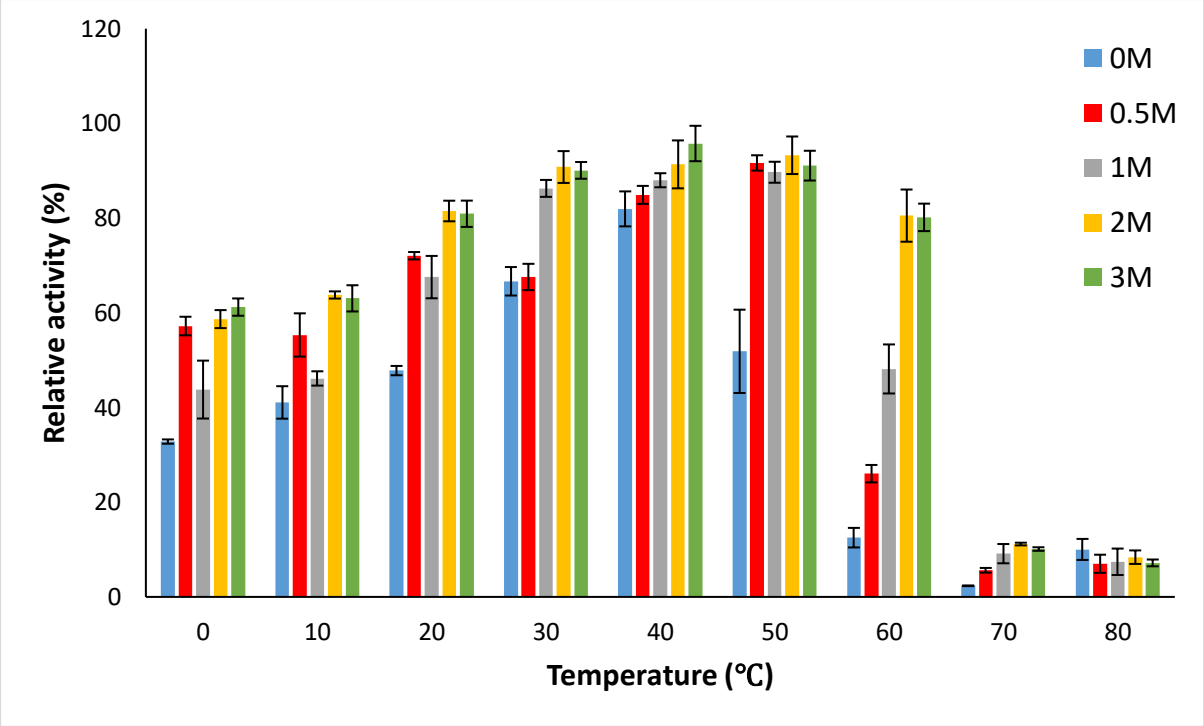


Figure 6.3 Influence of NaCl on the activity of the putative AFase GH51. Activity tests were performed using pNP-Araf as substrate and the reaction was performed in the presence of buffered NaCl at different concentrations (0M to 3M), for 30 min. Data are averages of five assays, and the bars represent standard errors.

In order to further investigate the action of AFase GH51, the purified enzyme was incubated with arabinoxylan (AX), unwashed AX, debranched arabinan and gum arabic. AX are the main constituents of hemicellulose in grass cell walls and arabinose residues are typically present as 1,2 and 1,3 linkages to the xylan chain. In contrast, arabinans are typically present in the side chains of rhamnogalacturonan I (RG I) and are typically linked in the 1,5 positions; and in gum arabic the arabinogalactans contain arabinosyl residues typically linked by 1,3 and 1,4 positions. The amount of arabinose present in each of these substrates varies from one substrate to another, which means that a quantitative comparison of activity on these different substrates is difficult but can serve to indicate the target linkages for enzyme activity. Enzyme and substrates were incubated at 30 ℃ overnight and the products of incubation were analysed by HPAEC. Standards of arabino-oligosaccharides and xylo-oligosaccharides were used for comparison. The results presented in figure 6.4 show that there was release of arabinose from all samples, except gum arabic. Neither xylose nor any other arabino/xylose-oligosaccharides were released for any of the substrates tested.

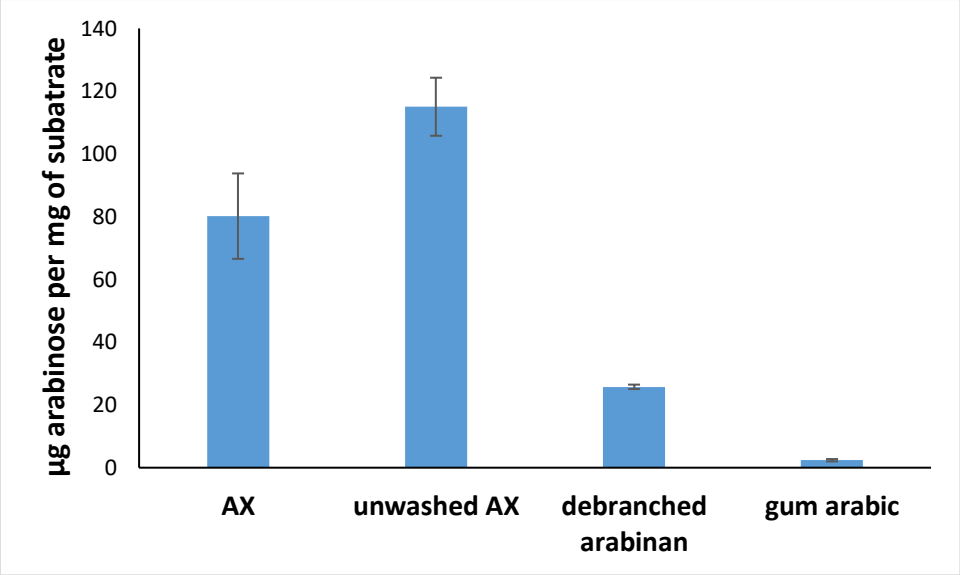


Figure 6.4 Analysis on HPAEC of products released after incubation of AFase GH51 with different substrates. Arabinose residues were released for all cases except gum arabic. Data are averages of three assays, and the bars represent standard errors.

These results suggest that AFase GH51 is active on both the AX and RG I substrates, suggesting it could be active on 1,2, 1,3 and 1,5 linkages. The lack of activity on gum arabic likely indicates a preference for linkages to a xylan rather than galactan backbone. Also, although the arabinan used for this experiment was debranched, the relatively low amount of arabinose being released might be that the enzyme is acting on minor amounts of arabinose residues still present as side chains in the positon 1,3 instead of the main chain of arabinose connected by 1,5 linkages. It is also important to mention that the only difference between AX and unwashed AX concerns how the stock solutions were prepared: the AX has been previously precipitated with ethanol and then washed to remove oligos and monos that may be present due to any spontaneous break down reaction before being resuspended in dH_2O and unwashed AX were directly resuspended in dH_2O. The fact that AFase GH51 is more active in the unwashed AX might indicate that the enzyme works better on smaller oligosaccharides than on longer polysaccharides.

6.3.2 Characterisation of a putative GH3 - clone 14GH3

A similar range of exploratory experiments was carried out to assess the biochemical activity of recombinant 14GH3 protein. Preliminary activity tests (table 6.1) showed that this target is most active on 4-nitrophenyl β-D-glucopyranoside (pNP-Glc) and slightly active in 4-nitrophenyl β-D-

xylopyranoside (pNP-Xyl), suggesting that the enzyme is a β-glucosidase (βglu). For this reason pNP-Glc was selected as the substrate for further enzymatic characterisation. Determination of optimum pH, optimum temperature, seawater influence and salt tolerance were performed as described for the GH51, replacing the substrate for pNP-Glc and the results obtained are present below. In figures 6.5 we can see that putative βglu GH3 has an optimum pH for activity between pH 6 and 7.

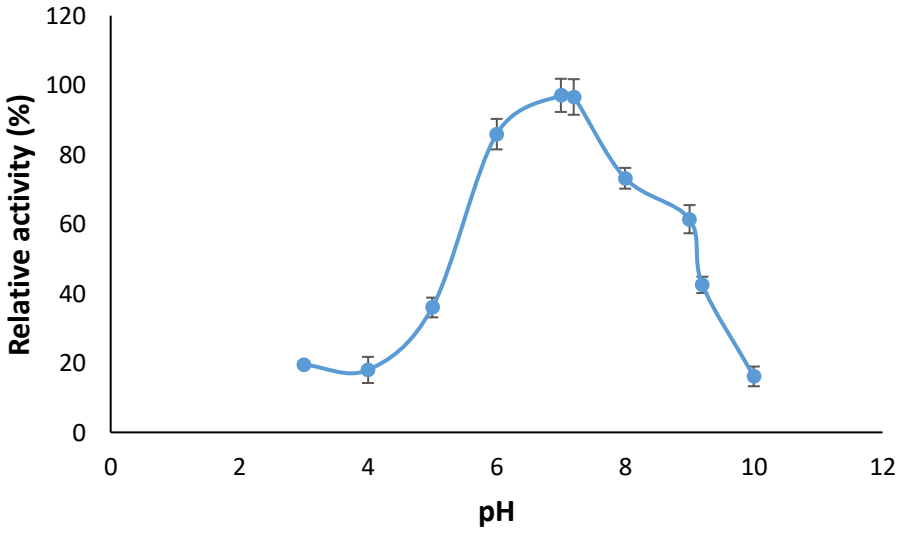


Figure 6.5 Determination of optimum pH for the putative βglu GH3. Activity tests were performed using pNP-Glc as substrate and the reaction was performed at 30 ℃ for 30 min. Data are averages of five assays, and the bars represent standard errors.

Data presented in figure 6.6 shows that βglu GH3 had much higher activity in seawater than in low salt buffered solution. Activity in seawater was more than double that evident in low salt conditions at 40 ℃, and almost 10 times greater at 50 ℃, whereas there was little difference between the two conditions at 10 and 20 ℃ or at 60 ℃. Interestingly, the βglu GH3, shows a similar level of activity at low temperature to that seen for the AFase GH51. The recombinant GH3 enzyme retained ~50% of its maximal activity at zero degrees Celsius in seawater.

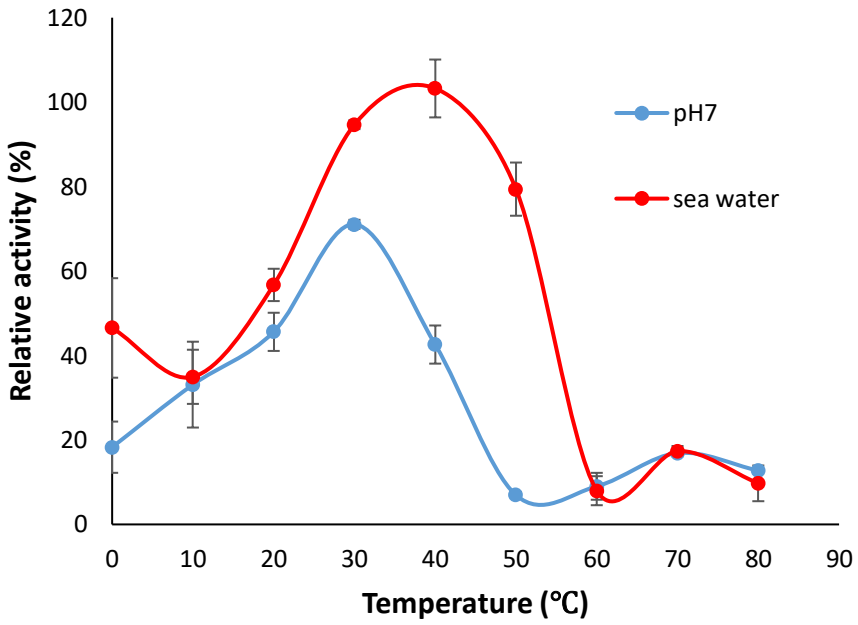


Figure 6.6 Determination of optimum temperature and influence of seawater on the activity of the putative βglu GH3. Activity tests were performed using pNP-Glc as substrate and the reaction was performed in the presence of the buffer at pH 7 (or seawater), for 30 min. Data are averages of five assays, and the bars represent standard errors.

Salt tolerance for the βglu GH3 (figure 6.7) is again evident in the results with NaCl, with the highest observed activities at 0.5, 1, 2 and 3M.

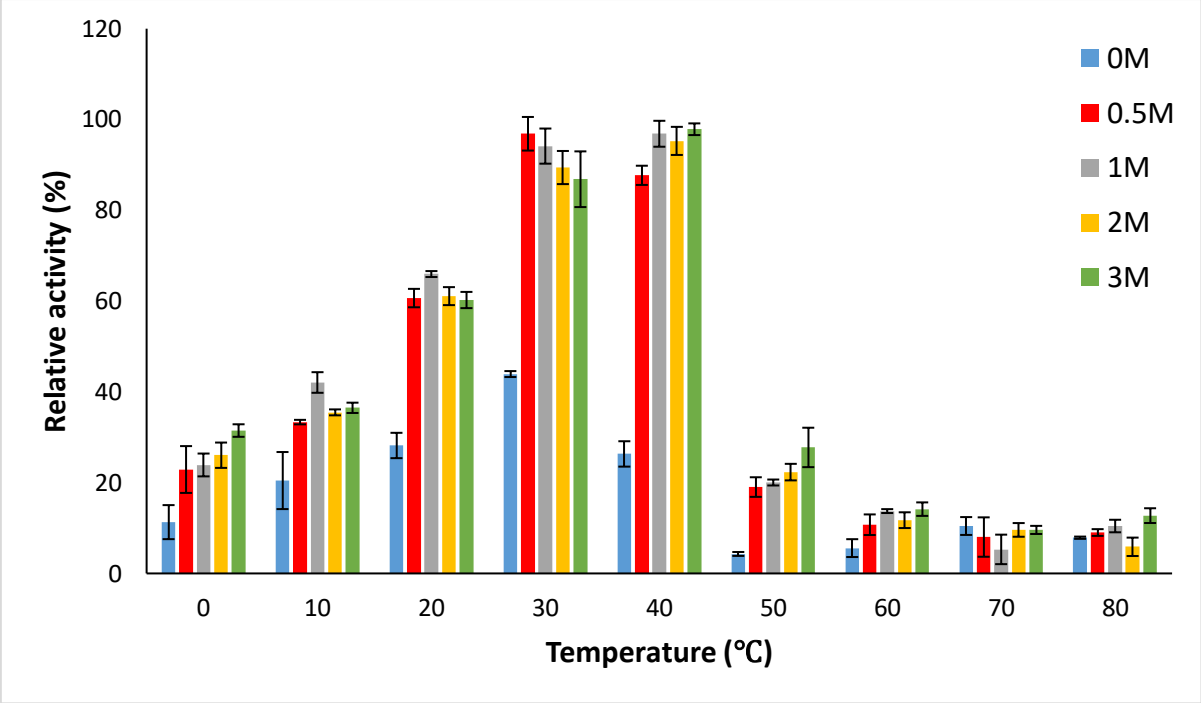


Figure 6.7 Influence of NaCl on the activity of the putative βglu GH3. Activity tests were performed using pNP-Glc as substrate and the reaction was performed in the presence of buffered NaCl at different concentrations (0.5M to 3M), for 30 min. Data are averages of five assays, and the bars represent standard errors.

Due to constraints of time only the activity tests mentioned above have been carried out for the putative βglu GH3. However if time permitted, experiments of incubation of this enzyme with different polysaccharides and oligos of different linkages, such as β 1,4 glucans; pachyman (β 1,3 glucan); and lichenan (β 1,3:1,4 glucan) for example, followed by analysis of the products released, would have been performed in order to better investigate βglu GH3's function in the deconstruction of biomass.

6.3.2 Characterisation of a putative CE1 - clone 34CE1

A similar range of experiments were performed for the putative CE1. Preliminary activity tests were performed (table 6.1), which revealed activity on 4-nitrophenyl Acetate (pNP-Ace). This is a standard, non-specific colorimetric substrate for esterase activity and was selected to perform the characterisation for the CE1. Once again, determination of optimum pH and temperature, as well as seawater influence and salt tolerance against NaCl were performed, now replacing the substrate for pNP-Ace Figure 6.8 shows that the putative CE1 has a more alkaline activity profile than the two GHs

tested, evident by its optimum pH between 7 and 8, which is compatible with general ester disruption that is favoured by high pH.

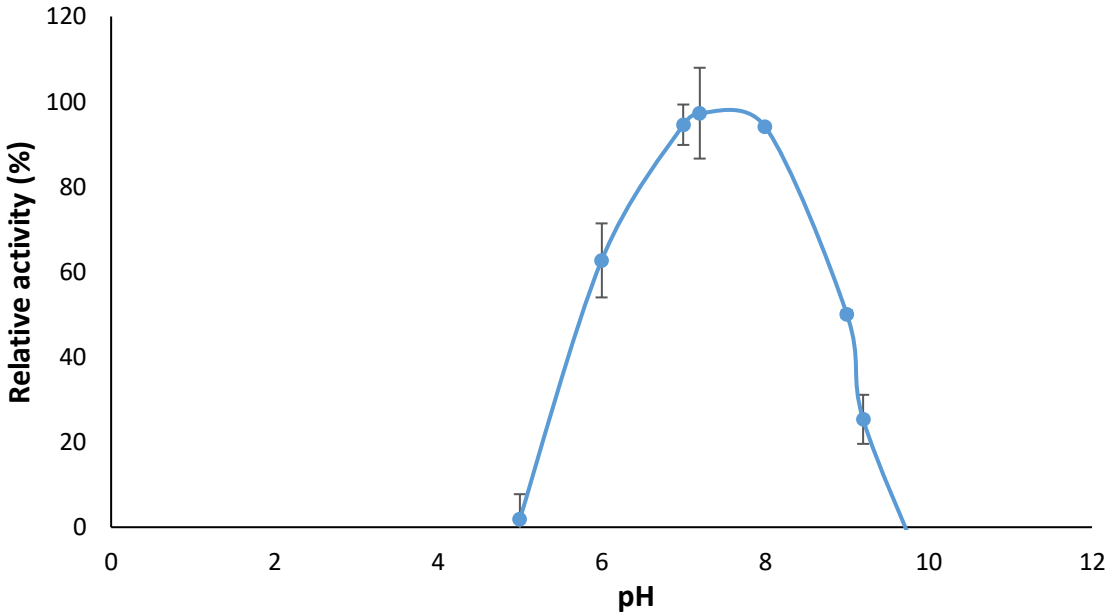


Figure 6.8 Determination of optimum pH for the putative CE1. Activity tests were performed using pNP-Ace as substrate and the reaction was performed at 30 ℃ for 30 min. Data are averages of five assays, and the bars represent standard errors.

The temperature profile of activity for this enzyme (figure 6.9) shows a much lower optimum (between 20 and 30 ℃) than seen with the other two enzymes. Preference of seawater also is apparent for this enzyme, for example activity of the enzyme at 20 ℃ was 40% higher in seawater than in simple buffered solution. An effect of stabilization of the protein by seawater is observed since the putative CE1 has ~70% of activity at 40 ℃, but it is less than a third of this value in the simple buffered solution. The same phenomenon of significant activity at low temperature was seen with this enzyme with ~40% of maximal activity seen at 0 ℃.

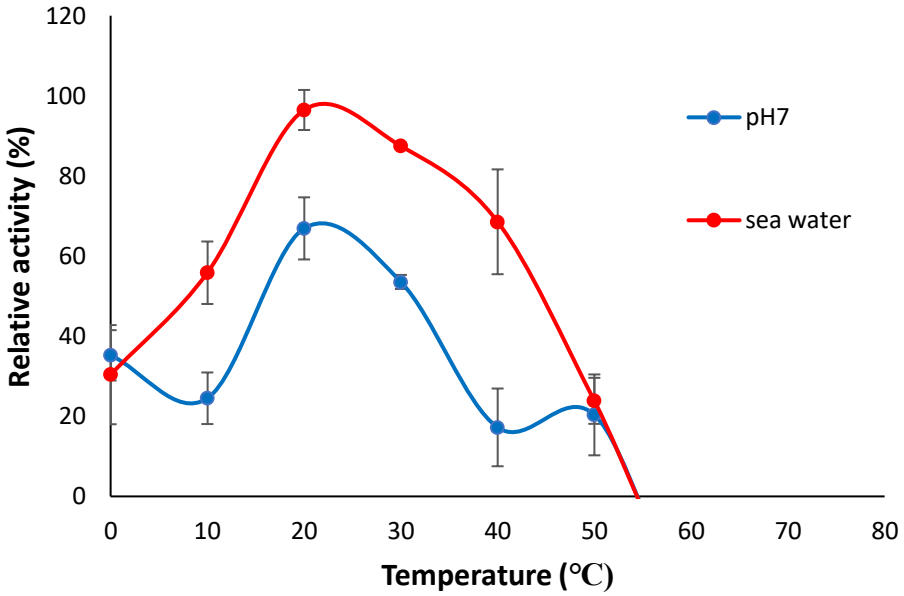


Figure 6.9 Determination of optimum temperature and influence of seawater on the activity of the putative CE1. Activity tests were performed using pNP-Ace as substrate and the reaction was performed in the presence of the buffer at pH 7 (or seawater), for 30 min. Data are averages of five assays, and the bars represent standard errors.

The putative CE1 demonstrated considerable stability in elevated levels of NaCl (figure 6.10) as was observed for the other two targets. Indeed activity in NaCl concentrations between 0.5 and 3.0 M were generally similar and significantly higher than in the absence of NaCl. Interestingly, the CE1 enzyme shows significantly higher activity in seawater than in similar concentrations of NaCl (0.5 M), most notably at 40 ℃. These findings suggest that other factors/components than NaCl in the seawater are having a stabilising effect on the protein at higher temperatures, perhaps some of the divalent metal ions such as calcium, magnesium or manganese help stabilise enzyme structure in seawater. This could be investigated in subsequent experiments.

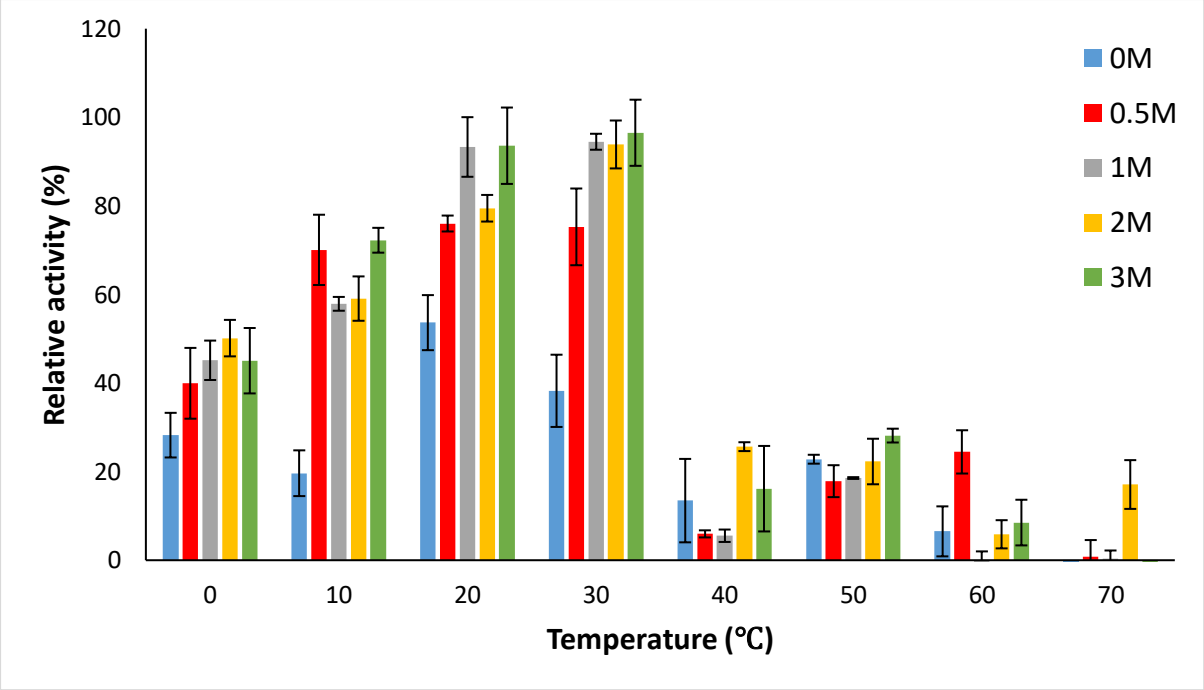


Figure 6.10 Influence of NaCl on the activity of the putative CE1. Activity tests were performed using pNP-Ace as substrate and the reaction was performed in the presence of buffered NaCl at different concentrations (0.5M to 3M), for 30 min. Data are averages of five assays, and the bars represent standard errors.

As mentioned before, all the characterization experiments were performed using the model pNP substrates. However, unlike the glycoside hydrolase substrate specificity experiments where a specific pNP substrate could be used for each enzyme, the use of model substrates with pNP groups are not specific for enzymes belonging to the carbohydrate esterases family. In this case, the assay using pNP-acetate only confirms that the enzyme being tested can cleave ester bonds. In order to understand what type of carbohydrate esterase the putative CE1 is, we performed some more specific tests and upon incubation of this enzyme with the substrate methyl ferulate (MFA), we observed the production of ferulic acid (FA) by HPLC. In figure 6.11 we can see the results of HPLC obtained by the incubation of putative CE1 with MFA. Figure 6.11 A and B refers to standards substrate (MFA) and product (FA) respectively. Fig 6.11 C shows the results obtained for time zero of the incubation, where we can see that a large peak of the substrate MFA is present and just a slight presence of FA as product. Figure 6.11 C, D and E show the change of profile over time (10 and 20 minutes of incubation, respectively) with the consumption of MFA by the CE1 and consequent formation of FA, which is the largest peak present in E. These results show that the putative CE1 is a feruloyl esterase (FAE).

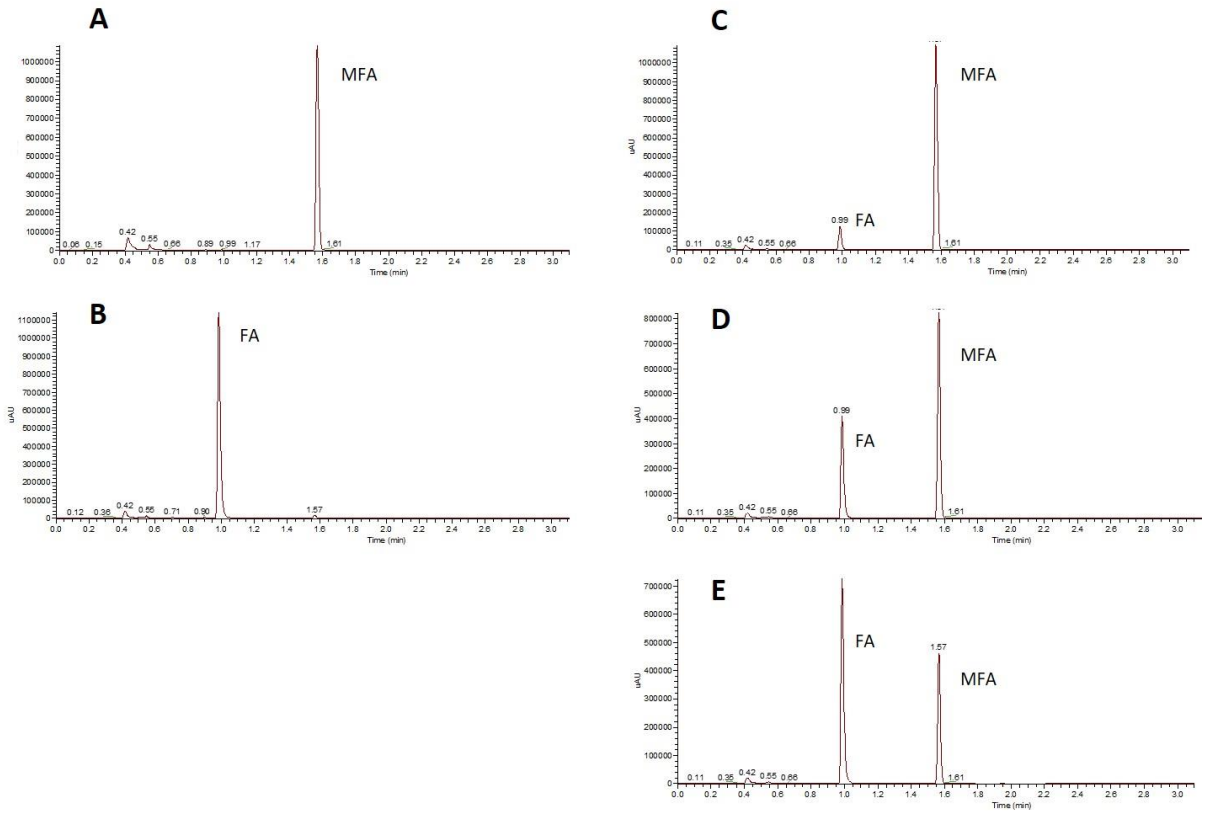


Figure 6.11 Analysis by HPLC of the products obtained after incubation of putative CE1 with methyl ferulate (MFA). **A** is the standard model substrate methyl ferulate (MFA). **B** is the standard for ferulic acid (FA). **C**, **D** and **E** are the products released for time zero, 10 and 20 minutes of incubation, respectively.

These experiments finalise the results obtained in this study until this moment. From the soluble proteins obtained, only AFase GH51 had a more in deep analysis with different polysaccharides. However, further experiments to investigate the action of this enzyme in arabino-oligosaccharides would be also helpful. The recombinant enzyme demonstrated enzymatic activity against arabinoxylans, and to a lesser extent, against debranched arabinans, corroborating with previous studies of AFases belonging to the family GH51 that exhibited activity on a range of different substrates [80]. The ability of this enzyme to release arabinose from these two substrates may indicate that it is specific for 1,3 linked arabinose likely to be present in both substrates. Moreover, AFases seem to have a broad range of pH for activity as different ranges have been reported in the literature (varying from 4 to 9), and they seem to typically have optimum temperatures closer to the optimum temperature of AFase GH51 when in presence of seawater (around 50 ℃) [72, 75, 164], which indicates that seawater is important for the enzymatic activity and stability of this protein. Moreover, not many studies have investigated the halotolerance exhibited by AFases, except for some cases

where bifunctionality (β xylosidase/arabinofuranosidase from families GH3 and GH43) has been reported [165, 166] and to the best of my knowledge, investigation of activity of AFases in seawater has not yet been demonstrated.

Regarding the putative βglu GH3, more experiments are needed in order to better understand the mechanisms and actions of this enzyme. It would be interesting to investigate its action against different cello-oligos for example, (cellotriose, cellotetrose, etc.) as well as how the βglu GH3 would behave in the presence of different concentrations of glucose (its main inhibitor). Finding new β glucosidases tolerant and/or stimulated by glucose is of great interest, and to the best of my knowledge there is no report yet of halotolerant β glucosidases with glucose stimulation. To better understand the real roles of this recombinant enzyme in the lignocellulose deconstruction, it would be helpful to test different polysaccharides with different linkages, as β 1,3; β 1,4 and α 1,4 glucans for example.

The final target identified in this study, the CE1 appears to be a FAE. As mentioned before the cross linking in hemicellulose of grasses is mostly made through ferulic acid (FA), dimers and trimers [19]. Similarly, there can be oxidative cross links between FA side chains on AX and lignin, further contributing to lignocellulose integrity [22]. Thus, FAEs are likely to play a crucial role in the deconstruction of lignocellulose. FAEs are typically more active in releasing FA from methyl substrates, but their substrate preference can vary [167]. For this reason, additional experiments involving the incubation of FAE CE1 with different hydroxycinnamate methyl esters substrates (methyl p-coumarate, methyl caffeate and methyl sinapate), as well as incubation of FAE CE1 with feruloylated AX biomass, followed by analysis of the products released, would help us to understand the activity of this recombinant enzyme. Furthermore, FAEs seem to have a great variety of optimum pH (3 to 10) and temperature (20 to 75 °C) reported [167] and a few cases of halotolerance have already been described [168, 169].

Moreover, experiments to investigate the kinetic parameters (Vmax and Km) of these three recombinant enzymes would be of great importance as it would allow us to evaluate and compare the performance of these enzymes with those described in the literature. Also, the elucidation of the protein structure by X-ray crystallography, would not only allow to perform comparison among these proteins' structure with their equivalents already published, but also it could potentially help us to better understand the mechanisms used by these enzymes regarding salt and cold tolerance for example.

Overall, it is clear that saltmarshes provided a good source to find halotolerant enzymes, as all the three soluble proteins obtained had preference for seawater and exhibited high salt tolerance. These results show that saltmarshes provide a valuable environment for the discovery of new lignocellulose-degrading enzymes with salt tolerance that could be used to create a salt-tolerant saccharification cocktail. Although the experiments of halotolerance have been performed for these enzymes, the second question of this work, whether these enzymes can contribute (or not) in the degradation of recalcitrant biomass, remains unanswered. Unfortunately, time constraints mean that those experiments will have to be placed on hold for now.

Chapter 7 Final discussion

7.1 General discussion

In the face of global environmental challenges (global warming, food and water security, etc.) it is unquestionable that we have to find sustainable replacements for fossil fuels and petrochemicals. In this context, lignocellulose plant biomass emerges as a promising feedstock for the production of bio-based chemicals and biofuels due to its abundance in the world. However, because plants have evolved to resist microbial and enzymatic attack, the conversion of lignocellulose into these products is not yet feasible due the recalcitrant nature of plant biomass. Although researchers and companies have focused on overcoming this by the use of different methods of biomass pretreatment and enzymatic cocktails for saccharification, these approaches are still not economically competitive. In addition, biomass processing uses large amounts of fresh water, which adds to its environmental footprint. With these issues in mind, the main objective of the work presented in this thesis was to try to find alternative halotolerant enzymes that would be able to act on the most recalcitrant components of the biomass. It was hypothesized that if these enzymes were able to degrade the recalcitrant components of lignocellulose, they could aid and potentially increase yields in biomass saccharification, and because they are salt tolerant they could potentially be used together with seawater (which is cheap, accessible and abundant) in the process, saving valuable fresh water for agriculture and human consumption. Halotolerant enzymes are typically found in marine ecosystems and in this work we have mined for them from among microbes isolated from a saltmarsh in northern England. Saltmarshes are constantly flooded by seawater and have salt tolerant terrestrial plant biomass as the main feedstock for the microorganisms living in that area. Because of this, it is likely that a range of different microorganisms living in saltmarshes are halotolerant and able to degrade plant biomass. In addition, saltmarshes are a relatively underexplored ecosystems in terms of biotechnological applications, which could potentially lead to new findings.

The strategy adopted in this project was to grow microorganisms from saltmarsh sediments on residual saltmarsh grass biomass that had already been degraded for a ten week period. Because this very recalcitrant biomass was the only source of carbon for those microorganisms, we expected that only those able to degrade the biomass would survive, providing us with a source of potentially useful enzymes. We observed that during incubation, the weight loss from the depleted biomass was very slow and after eight weeks of incubation there was only a 22% reduction in biomass. It is well known that lignin is often the hardest component of the lignocellulose to degrade and that its degradation tends to be the slowest. Thus, the slower degradation observed in this work suggest the

presence of microorganisms and enzymes among the community growing in the recalcitrant biomass, with the capability to degrade and/or modify lignin. In fact, the compositional analysis of initial and final recalcitrant biomass revealed that lignin was the component with the highest content loss (almost 50%) during incubation. In nature, many microorganisms are able to degrade/modify lignin, but the most efficient are wood-rotting fungi. These fungi are classified as white, brown and soft-rot fungi according to the aspects of the wood being degraded and the characteristics of the remaining lignocellulose after degradation. White-rot fungi are the most efficient lignin degraders and are able to degrade and mineralise all the components of lignocellulose; brown rot fungi typically are able to degrade lignin to a lesser extent by partially modifying it, leaving mostly oxidised lignin in place; and soft-rot fungi typically weakly affect lignin resulting in a soft and crumby residue [170]. The strategies used by these fungi to degrade lignin involves the secretion of a range of oxidases, such as lignin and manganese peroxidases, as well as laccases. These enzymes act synergistically and promote the degradation/modification of lignin by oxidation [170, 171]. Interestingly, in this work no fungi were identified among the CAZyme producers present in the final recalcitrant biomass and the changes observed in lignin content is likely to be due to bacterial activity. Although bacterial degradation of lignin has not been as intensively studied as fungal, there are reports of its occurrence, mainly from member belonging to *Actinomycetes*, *Alphaproteobacteria*, and *Gammaproteobacteria* classes [172]. Interestingly, in the present study, the two main classes of putative CAZymes producers identified were *Alphaproteobacteira* and *Gammaproteobacteria*, suggesting that enzymes belonging to these classes could be responsible for the degradation of lignin observed. Also, among the known enzymes related to lignin modification, enzymes belonging to the AA2 family were recognised in this work but no laccases were seen. In fact, the AA super family was almost exclusively represented by members of the AA2 family. AA2 comprehends a group of peroxidases, such as manganese, lignin and versatile peroxidases that are typically secreted by fungi [173, 174] and enzymes belonging to the AA2 family are among the main enzymes secreted by the wood-rotting fungi. Interestingly, there is no report of any member of this family from bacterial origin on the CAZy database to date suggesting that the putative AA2 identified in this work are either annotated wrongly (and could be from fungi not identified in this work) or may include some potentially interesting new enzymes. In addition, a range of different putative peroxidases and superoxide dismutases (SODs) that are enzymes with potential lignolytic activity [144, 145] were also observed in this present work. These results suggest that the community living on the recalcitrant biomass is either exclusively using peroxidases from the AA2 family in order to promote the degradation of lignin, or may have evolved specialized lignin-degrading enzymes that are annotated as SODs, or perhaps there are novel lignolytic enzymes present in saltmarshes that are not currently described.

Results from comparing the composition of monosaccharides present in the biomass before and after incubation show that monosaccharides belonging to pectin were the ones that were completely degraded. Pectin describes a complex range of polysaccharides, characterised by the presence of galacturonic acid. Both galacturonic acid and rhamnose were completely lost during the incubation of the biomass. It is generally accepted that pectins are usually associated with primary cell walls and are at best minor components of secondary cell walls [6]. Although senesced grass stem biomass (such as that used in this project) is principally composed of secondary cell walls, each secondary cell wall is by necessity surrounded by a primary cell wall on which it is deposited. There are reports in the literature of pectin playing a significant role in determining the saccharification of lignocellulosic biomass [33, 34]. Recent studies conducted by Biswal *et al*., for example, have shown that engineered switchgrass, rice and poplar plants with lower content of homogalacturans (HG) and rhamnogalacturonan II (RG II) showed improved saccharification yields for plants growing in greenhouses and in the field [175]. In addition, studies published by Lionetti, *et al*., have shown that Arabidopsis plants engineered for reduced methylesterification in HG had also an increased efficiency in enzymatic saccharification [176]. In the currently work, although in small abundance, three putative polysaccharide lyases (PLs) related to pectin degradation were identified in the proteome: a PL9 and two PL1s. Enzymes belonging to these families are typically pectate lyases, which are enzymes that degrade pectins with lower degree of esterification [177, 178]. These enzymes are then sub classified according to their substrate preference (poly-galacturonic acids or oligo-galacturonic acids) into exo-pectate lyases (as for example exopolygalacturonases lyases) or endo-pectate lyases (as for example endopolygalacturonases lyases) [179]. Additionally, among the CEs identified in the present work, there was a notable abundance of putative CE8s. According to the CAZy database, CE8s are exclusively pectin methylesterase, which are enzymes that can promote the demethylesterification of homogalacturonans [180] and even though all rhamnose was lost during the incubation, no putative rhamnosidase was evident in the proteome. Therefore, the results observed in this work could indicate that the removal of pectin might be needed for the degradation of lignocellulose and suggest that the synergistic action of CE8s and pectate lyases potentially are involved in the degradation of pectin observed in this study.

Regarding monosaccharides from the hemicellulose fraction, there was a complete loss of glucuronic acid (GlcA) during biomass incubation. This is notable because GlcA residues in xylans have been suggested to serve as points of linkage to lignin and published studies have demonstrated an increase in saccharification by the removal of GlcA [181]. There was also a substantial loss of arabinose during biomass incubation. Arabinose makes up the major substituting monosaccharide in the

complex glucuronoarabinoxylan (GAX) of grass cell walls. In addition, some of the arabinosyl side chains of GAX are decorated with feruloyl esters, which can form crosslinks within and between GAX chains as well as linkages to lignin [22, 23] and have been shown to be important for saccharification [182, 183]. In fact, putative enzymes related to the removal of all the main decorations presents in GAX have been identified in this work: CE1 (typically feruloyl esterase), which are related to the removal of ferulic acid [83]; CE6 (acetyl xylan esterase), which are related to the removal of acetyl groups [70]; CE15 (methyl-glucuronoyl methylesterase), which are responsible by demethylesterification of GlcA [184]; GH51 (typically arabinofuranosidase), which are related to the removal of arabinosyl residues [72]; and GH67 (glucuronidase), which are related to the removal of GlcA [185]. These results imply that the community growing on the recalcitrant biomass is well equipped with enzymes able to remove decorations present on GAX.

Producing active recombinant versions of selected enzymes identified in this study proved problematic, with only three being successfully produced. The difficulties encountered in protein expression were perhaps to be expected as, saltmarshes microbes have been little studied and are adapted to harsh conditions. Additionally, although saltmarshes have been studied for a while in terms of biodiversity, very little is known concerning their potential for biotechnology, making it difficult to find papers and studies to be used as models for the production of recombinant proteins. The three enzymes satisfactorily purified were demonstrated to be an arabinofuranosidase (AFase), a β glucosidase (βglu) and a feruloyl esterase (FAE). The fact that two of these enzymes (AFase and FAE) are active on GAX side chains reinforces the abundance of such apparent enzymes encountered in the proteome However, further studies investigating the synergistic action of these enzymes on saccharification as well as their influence as addictive in cellulose cocktails are needed before any conclusions can be drawn. Enzymatic activity tests performed for these proteins have demonstrated their preference for seawater, their high salt tolerance and their cold stability, reflecting the potential of saltmarsh for the discovery of novel halotolerant lignocellulose-degrading enzymes as well as its potential as a source for cold-active proteins in general.

It is known that the interaction between water molecules and proteins are crucial for maintaining their three dimensional structure, biological activity and solubility, and that at high salt concentrations most known enzymes are inactive. This is likely to happen because in high salt concentrations, ions sequester water molecules, limiting the free molecules available for hydration of the enzyme. In addition, ions promote disruption of the organized layers of water molecules around hydrophobic regions of the protein's surface, as well as disturb of electrostatic interactions between adjacent charged groups [89, 96]. Unlike most enzymes, those showing halotolerance are able to

compete with ions in high concentrations for water molecules, conserving their activity and structure. This ability is believed to be related to a higher content of charged amino acid residues (especially acidic residues), and a lower content of hydrophobic amino acid (compared to smaller amino acids) on the protein's surface when compared to non-halotolerant enzymes [186]. In fact, a general property of known salt tolerant enzymes is the predominance of acidic to basic residues. Apparently, the acidic amino acids on the surface of halotolerant proteins bind to hydrated cations, forming a hydration layer that preserves their structure and activity and prevents aggregation and precipitation of the protein. On the other hand, interactions between opposite charged amino acids near to each other, tend to form salt bridges that are equally important for protein folding and structure [96]. Finally, the low content of hydrophobic amino acid on the protein's surface compared to the smaller amino acids, facilitates proteins hydration and increase their flexibility [187]. Indeed a quick analysis of the protein's parameters using tools of ProtParam from Expasy website (https://web.expasy.org/protparam/), revealed a prevalence of acidic charged amino acids (glutamate and aspartate) over basic charged amino acids (lysine, arginine and histidine) and a lower amount of hydrophobic amino acid (phenylalanine, leucine and isoleucine) compared to smaller amino acid (glycine and alanine) for all three halotolerant enzymes described in this work, except for the AFase GH51, where amounts of hydrophobic and smaller amino acids were equal (table 7.1). Our data suggest that these enzymes could be good candidates for use in biorefineries using seawater. However, more detailed experiments of x-ray crystallography are needed in order to better understand how these amino acids contribute with the protein stability and activity in higher salinity concentrations.

Table 7.1 Composition of charged, hydrophobic and small size amino acids for the three proteins identified in this work

		AFase GH51	**βglu GH3**	**FAE CE1**
Basic charged amino acid	Lysine	2.9%	3.5%	3.6%
	Arginine	3.7%	3.5%	3.9%
	Histidine	2.9%	1.8%	2.1%
Acidic charged amino acid	Glutamate	6.8%	4.7%	6.6%
	Aspartate	7.4%	8.8%	8.1%
Hydrophobic amino acid	Phenylalanine	4.3%	3.8%	3.9%
	Leucine	6.8%	8.7%	6.6%

	Isoleucine	4.9%	4.7%	5.4%
Small amino acid	Glycine	7.6%	10.3%	7.6%
	Alanine	8.4%	10.3%	9.9%

Another curious observation for the three enzymes characterized in this work, is regarding their enzymatic activity at low temperatures. Rather as in the case of high concentrations of salt, cold temperatures also affects the features of proteins due to their interactions with water molecules. In this case, a decrease in temperature causes the water molecules around the protein to become more organized and interact with each other, decreasing the interaction between water and protein, which consequently leads to protein denaturation. It seems likely that both halotolerant enzymes and cold active enzymes maintain stability by encouraging strong interactions with surface water molecules in order to maintain their structure and activity [96]. The mechanisms adopted by cold active enzymes are not yet completely understood but these are believed to be related to high levels of interaction between protein and solvent (the same features exhibited by halotolerant enzymes) and to high levels of flexibility mainly in the active site of the cold active protein [188, 189]. The analysis of the amino acids present in the active site of the proteins identified in this work is not possible, since the protein structure is unknown, however the fact that they demonstrated enzymatic activity at low temperatures, suggests that those features could be present in these proteins. Finally, although temperatures of saccharification are typically conducted between 40 and 50 °C, the cold tolerance exhibited by the proteins identified in this work could be of biotechnological interest for different industrial applications, such as food, detergent, pharmaceutical and textile industries, for example [189].

Furthermore, a general comparison of the results presented in this work with previous work performed in our laboratory with saltmarshes (Leadbeater *et al*., 2019 - under submission) revealed some interesting findings. Among the CAZymes producers for example, both studies show the prevalence of two main phyla: *Proteobacteria* and *Bacteroidetes*; and in both cases *Gammaproteobacteria* was the class with the biggest representatives (60% in the currently study and 39% in Leadbeater's work). Considering that Leadbeater's work was performed *in situ* in saltmarshes, while the present work was performed in shake flasks with controlled temperature and agitation, it is interesting that the same phylogenetic groups were dominant. Not surprisingly, likely in the present work, all the enzymes characterised in the Leadbeater's work also showed halotolerance. One notable difference between these two works was that in the current work the AA super family represented 12% of the proteome, from which AA2 were the majority. In contrast, the AA superfamily was not very

abundant in Leadbeater's work. Likewise, although CE8s appeared in high abundance in this work, this family was not identified in Leadbeater's work. These differences suggest that the use of recalcitrant biomass in this studies had a considerable influence in the enzymes being produced by the communities.

7.2 Future work

The studies conducted in this thesis had as a main objective the discovery of salt tolerant lignocellulose-degrading enzymes able to degrade the most recalcitrant biomass and a total of three targets have been satisfactorily characterised. However, if time wasn't a limit, it would be interesting to investigate how to produce a wider range of enzymes from this system, perhaps by testing a wider range of solubility tags into the expression vector, performing the optimization of rare codons for bacterial expression, or trying alternative expression systems. Also other aspects of the protein expression could be tested, as for example the introduction of seawater and/or NaCl in the medium used for expression.

Among the putative CAZymes identified, GHs (mainly cellulases and hemicellulases) were the most abundant, followed by CEs and AAs. Interestingly, the majority of the CEs identified belong to the CE10 family, which until this moment is a family of enzymes without proven activity on polysaccharides and their presence in such an abundance in this work suggest they might be important for lignocellulose deconstruction. Because of this, it would be interesting to investigate those targets more deeply. Another interesting feature of the putative CAZymes identified in this work was the presence of a polyserine sequences exhibited by some targets. The presence of these repetitive amino acid sequences is usually associated as a linker between different domains of the protein. The fact that some of these polyserines are present in sequences annotated as only having a single domain, suggest that a second still unknown domain might be present and investigations of these targets could also lead to new findings.

While three halotolerant proteins were satisfactorily expressed and characterised, the effects of these enzymes on recalcitrant biomass remains unknown. To investigate this, an experiment incubating biomass with each of these enzymes (individually and combined), followed by analysis of total sugars released (comparing with the results obtained by cellulase cocktails), would be of interest. In addition, in order to better understand the role of these enzymes in lignocellulose degradation some additional experiments are also needed. The AFase GH51 proved to be active on arabinoxylan (AX) and arabinan, but it is still unclear which bonds are being attacked (1,2, 1,3 or 1,5). The βglu GH3

was only characterised using pNP substrates. A deeper investigation of its action should be undertaken by incubation of this enzyme with different polysaccharides and oligos of different linkages, such as β 1,4; β 1,3 and α 1,4 glucans, which would allow us to have a better understanding of its substrate specificity. Also, the effect of different concentrations of glucose on the enzymatic activity of this enzyme would be useful to investigate if this enzyme could be tolerant or even stimulated by glucose. The FAE CE1 demonstrated the ability to release ferulic acid from methyl ferulate, however its incubation with different methylated substrates (methyl p-coumarate, methyl caffeate and methyl sinapate) would help us to understand the substrate affinity and preference exhibited by this enzyme. In addition, incubation of FAE CE1 with biomass and feruloylated AX, could provide evidence of its ability to release ferulate from these substrates. Finally, experiments to determine the kinetic parameters (Vmax and Km) of these three enzymes would allow us to compare our results with the ones already published for equivalent enzymes, and elucidation of the structure of these proteins by x-ray crystallography would be of great importance for comparison with enzymes already published and potentially to help us to understand the mechanisms developed by these enzymes regarding their salt and cold tolerance.

Overall this work shows the potential for novelty that can be obtained by the approaches we have used. First, saltmarsh sediments proved to be a reliable source for halotolerant lignocellulose-degrading enzymes and the use of these enzymes during saccharification of biomass, can potentially make it possible the replacement of fresh water by seawater into biorefineries. In addition, saltmarsh sediments can be a good source for cold tolerant enzymes. Finally, the approach of using biomass that has already been extensively degraded, combined with meta-omics approaches, has demonstrated an efficient way to mine for unusual biomass degrading enzymes.

Appendices

Clone 8GH51 – AFase GH51

Nucleotide sequence (predicted signal peptide underlined):

ATGAAAAGACTGATATCCGCATTTGCGCTTTCAATCGCATGTTTTGGTATGGCGAGCGCACAGAACGCCGTCACTCTGGAT
AAAGACGCTTCGCTGGGAACAATCCAGCCCGAAGTTTACGGACAGTTCCTTGAGCATTTAGGCACACAAATTTATGACGG
AATGTGGGTCGGCGAAGACAGCTCCAGACCGAATGTTGGCGGAATTCGGAAAGATGTTTTTGACGCGCTTGATGCGCTG
GATATTCCTGTCATCCGTTGGCCGGGCGGCTGTTTTGCCGATATCTATCACTGGCGTGATGGTGTGGGATCCAGAGATGA
AAGAACCCCACGCGTGAATGTCAGTTGGGATTCAACGCCAGAATCCAATCAATTTGGTACGCACGAATTTTTTAATCTGGC
CGAAGCCCTTGGTGCGAAAACCTATTTGAATTTCAATCTCGGTACCGGAACGCCGGAAGAAGCGACAGATTGGATGGAAT
ATATCACAGCTGATCATGATTCAGCGCTGGCTCAGGAAAGACGCGCAAATGGCCGCGCAGAGCCTTGGAAAGTCGATTA
CATTTCCATCGGTAACGAAACATGGGGATGCGGCGGCAATATGCGGCCCGATTATTATGCTGACCTCTACGTGCAGTGGT
CGACCTTCATCAGATCCCACAGCGGCGACCAGCCCAAGCGTATAATCTCCGGATCTCACAATGGGAATATAGATTACTCCG
ATACGATTTTGGACCACTGGGCGATGAGAAACCTGTCTGACGGCATTGCGTTGCACTACTACACACTGCCAACGGCGGAT
TGGGGCGACAAAGGCGAAGGTGTTGATTTCCCGGAAGAGCAGTGGGCAAGCACGATTGCAAATACGATAGAGATGGAC
GCTTTCATTTCCGAGCAATTGGCGATGTTCGAAAAACATAAGTACCTGAAGGATGATTTTGGTCTCTATGTCGACGAATGG
GGTGTTTGGACAAATACGCCAGAGCGCATGCCAGCCTTGTGGAACCACAGCACAATTCGTGAGGCGGTTGTTGCCGGCCT
GAACTTCAACATTTTCCACAAATACGCGGAAGATGTGCCCATGACCAACATTGCTCAGATGTTGAATGTGCTGCAGTCCAT
GATCCTGCTGGAGGGCGACGATATGGTCCTCACGCCAACTTATCACGTGTTCGAAATGTACAAGCCATTTCAAGGCGCCG
AGTCTGTGAGTGTGTCTATTGAAACGCCAACTTTGACGAATGGGGAAAATAGCTTTCCTGCGCTTTCTGTTTCTGCTGCAA
AAACGGCTGACGGCAAATTGGTTGTTGGGTTAGTGAATGCGGATTCGAACAGCGCTCATGAAGTGTCATTCCCGCGTCAA
AACGGTCAAACGGTTGCCGGGCGTGTCCTGACAGCGGACGAGAATGACGCGCATAACAGCTTCGAGAATCCCGAGCTTA
TCAAGCCGATGCCGGCGAGTGTTTCGTCGACATCAGACGCCTTTACAGCAACATTACCTGCACGGTCGGTTTCCGTTTGGG
TAATTGAATAG

Protein sequence (predicted signal peptide underlined):

MKRLISAFALSIACFGMASAQNAVTLDKDASLGTIQPEVYGQFLEHLGTQIYDGMWVGEDSSRPNVGGIRKDVFDALDALDIP
VIRWPGGCFADIYHWRDGVGSRDERTPRVNVSWDSTPESNQFGTHEFFNLAEALGAKTYLNFNLGTGTPEEATDWMEYITA
DHDSALAQERRANGRAEPWKVDYISIGNETWGCGGNMRPDYYADLYVQWSTFIRSHSGDQPKRIISGSHNGNIDYSDTILDH
WAMRNLSDGIALHYYTLPTADWGDKGEGVDFPEEQWASTIANTIEMDAFISEQLAMFEKHKYLKDDFGLYVDEWGVWTNT
PERMPALWNHSTIREAVVAGLNFNIFHKYAEDVPMTNIAQMLNVLQSMILLEGDDMVLTPTYHVFEMYKPFQGAESVSVSIE
TPTLTNGENSFPALSVSAAKTADGKLVVGLVNADSNSAHEVSFPRQNGQTVAGRVLTADENDAHNSFENPELIKPMPASVSST
SDAFTATLPARSVSVWVIE-

Clone 14GH3 – βglu GH3

Nucleotide sequence (predicted signal peptide underlined):

ATGATGCGATCTTTTATAGCGGCACTCTGCCTGAGCGCAGCCCTTGGCGCCTGTTCAAATCCTGCGACGGAGGCGCCTGC
CCCAAAAACAGATGCCGACGCGGCATCCAATACATTGACGGTATGGCCGGATCTTGATGGGAGCTTCATGATTGACCCCG
CTATCGAAGCGCAAATTACCGATATCATGTCACGTATGACATTAGAGCAAAAAGTCGGACAGGTCATTCAAGGCGATAGC
ACTACTGTCACGCCGGAAGACGTTAAAACATACCGTTTAGGCTCTGTTCTAAGCGGCGGAAATTCAGCTCCGGGTGAGCA
TCCTTATGCCTCTATTGAGGAGTGGGTCGAGGCGGCGGATGCTTATTATCTGGCCTCTATTGATGACAGTGATGTTGAAGT
TGCGATCCCTGTCATATGGGGGATCGATGCCGTACATGGTCATGGCAATGTGATCGGCGCAACCGTCTTTCCGCATAATAT
CGGCCTCGGGGCAATGCGTAATCCGGCTTTAATCGGTGATATTGCCGCCGTGACAGCCCGCGAATTGCGCGCCACCGGAC
ATGATTGGACTTTCGCGCCCACTGTCGCAGTTCCTCAGGATGACCGTTGGGGTCGGACCTATGAAGGATTTAGTGAAAAC

CCCGAAGTCGTCGCCTCGTATTCCGGTGAGATTGTCAAAGGCATTCAAGGCGATCTTACCCAAACCAAGACAATCGATTCC
GACCATGTCATCTCAACCGCCAAACATTTCCTGGCGGATGGCGGTACGGATATGGGTAAGGATCAAGGCGATGCGCTCGC
CAATGTCGAAGACTTGGTCCGTATTCATAATGCTGGCTATCCGCCGGCGCTCGATGCAGGCGCCCTCTCAGTCATGGCCTC
TTTTTCAAGCTGGCAAGGCAATAAAGTTCATGGCTCTAAATATCTTTTAACGGACGCCTTAAAAGACCGGATGGACTTTAA
AGGTTTTGTCGTCGGCGACTGGAACGCCCATGGCCAGATTCCTGGCTGTACAAATGAAGATTGCCCAGCCGCACTTGAGG
CCGGACTCGATATGTATATGGCGCCGGATAGCTGGAAAGGGCTGTATAATAGTTTGCTGGCGCAAGCGAAGTCCGGAGA
GCTGTCCATGACCCGGCTCGATGATGCTGTGCGCCGTATATTGCGCGCGAAAATTCGCTATGGCCTGTTCGATATGGGCA
AACCCTCCGACCGTCCTTTGGCCGGAGACCGCTCTGTTCTCGGTGCGCCGGATCATAAGGCCGTAGCCCGTCAAGCCGTTC
GGGAATCTCTCGTGTTGCTTAAAAATGAGGGTCAAATTTTACCTCTATCGCCAAACCAAAATATATTAGTCGCGGGCGGAG
GCGCAGATGACATTTCCAAACAGGCAGGCGGCTGGACACTGACATGGCAGGGCGGTGGATTGGGCAATGATCTTTTCCC
GTCTGGCGAGTCTATTTTTAGCGGCATTCAAAAGGCAGCCCTTGCAGGCGGCGGCACCGTTCAGCTCTCCGAAGATGGTA
CTTTTACGCAAACACCGGATGTCGCCATTGTCGTCTTCGGCGAAGACCCTTATGCCGAATTTCAAGGCGATAGACCGCATG
TCGGTTATGACCCTTTTAGTCAAAAGGAAGTCCGGCTCCTGCGCGAGTTTCAAAACCAAGGCATACCGACTGTCTCGGTTT
TCCTTTCAGGGCGCCCACTTTGGGTTAATCCCGAACTGAACGCTTCCGATGCTTTTGTCGCAGCGTGGCTGCCGGGTACAG
AAGGGGCCGGAATAGCGGATGTTTTATTCCGCGATGAAAGCGGCAATATCGGGTATGATTTTACCGGAAAGCTCTCTTAC
TCTTGGCCAAAATCGGCGGGGCAAACCCCGCTTAATTATAGCGATAGTAATTATGATCCGCTCTTTGCCTATGGCTTCGGT
TTGACCTATGCCGATGATGTTTCCCTCCCCGTTCTGGACGAAGCGCCAGAAATTGATCTCTCAAAAGCCGGATTAAATCTG
ACCCTTTTCAAGGACGGGCAGGTTCAAGCGCCTTGGGCTCTCACGCTAAGCGGGGATGCCAGCACAATGGCGGTCGATC
ACCAAGCGCAAGAAGACGCGCTAAAATTTGAGTTTAACGGCCCCGGAACCGCCGTTATCGGCGTGACTGATTCCGTCGAC
CTGTCCCGCGAGACAACAGGCGCACTTGAACTCGCCTTCAACATAAAACGGAACAGCACACGCGAAGGCGGGATGACCT
TATCTGCAAAATGCCCGAATGATAGCTGCGCCGGGCCTTTGGATTTGTCCAAATCGGTGGACAACCTTGGCGATGACTGG
ACACCGGTGCGAATTGCGCTGTCCTGTTTCCGGGATTCGGGTGCGGATATGTCCAACATTCAAACGCCTTTCCGGCTCGTC
ACCAGCGGCCCGGTTTCGATCTCCATATCCGACTTGCATATCGCCGAAGATGACAATGGCGAGGCAAGCTGCACATTCTA
A

Protein sequence (predicted signal peptide underlined):

MMRSFIAALCLSAALGACSNPATEAPAPKTDADAASNTLTVWPDLDGSFMIDPAIEAQITDIMSRMTLEQKVGQVIQGDSTTV
TPEDVKTYRLGSVLSGGNSAPGEHPYASIEEWVEAADAYYLASIDDSDVEVAIPVIWGIDAVHGHGNVIGATVFPHNIGLGAM
RNPALIGDIAAVTARELRATGHDWTFAPTVAVPQDDRWGRTYEGFSENPEVVASYSGEIVKGIQGDLTQTKTIDSDHVISTAKH
FLADGGTDMGKDQGDALANVEDLVRIHNAGYPPALDAGALSVMASFSSWQGNKVHGSKYLLTDALKDRMDFKGFVVGDW
NAHGQIPGCTNEDCPAALEAGLDMYMAPDSWKGLYNSLLAQAKSGELSMTRLDDAVRRILRAKIRYGLFDMGKPSDRPLAG
DRSVLGAPDHKAVARQAVRESLVLLKNEGQILPLSPNQNILVAGGGADDISKQAGGWTLTWQGGGLGNDLFPSGESIFSGIQK
AALAGGGTVQLSEDGTFTQTPDVAIVVFGEDPYAEFQGDRPHVGYDPFSQKEVRLLREFQNQGIPTVSVFLSGRPLWVNPELN
ASDAFVAAWLPGTEGAGIADVLFRDESGNIGYDFTGKLSYSWPKSAGQTPLNYSDSNYDPLFAYGFGLTYADDVSLPVLDEAPE
IDLSKAGLNLTLFKDGQVQAPWALTLSGDASTMAVDHQAQEDALKFEFNGPGTAVIGVTDSVDLSRETTGALELAFNIKRNST
REGGMTLSAKCPNDSCAGPLDLSKSVDNLGDDWTPVRIALSCFRDSGADMSNIQTPFRLVTSGPVSISISDLHIAEDDNGEASC
TF-

Clone 34CE1 – FAE CE1

Nucleotide sequence (predicted signal peptide underlined):

ATGGCGACCACTCTGTCCTTAGCAAGCACAGCAATCAACGCGGCTGCACAGGACGAAACTTCTGTCACAACGCAGAAAGT
GACGATTCACAGCGATGCCGTTGAAGGCAATCTCGAGGGGAATTCAGCCGAACGGGATTTGTTGATTTATTTGCCGCCGT
CATACGACACAGACACAGATAAAAGATATCCTGTGATCTATGGACTGCATGGGTACAGTATCGACAATGACCAGTGGTCG
AAAGAAATACAGACCCCGACAACTATCGATGCCGCGTTTACGGACGGCGTTTCTGAAATGATCGTGGTGTTGCCAGATTC
GAAAACGCTGCATAACGGCTCCATGTATTCGAGCTCCGTCACCACGGGTGATTTCGAGACATTTATCGCAGAAGACGTTG
TCAATTATATCGATGCGAATTACCGCACGATCCCAAAAAGGGAATCACGCGGATTGGCGGGTCACTCAATGGGCGGCTAC
GGCACACTCAGAATTGCGATGAAGCGTCCGGATGTGTTTAGCAGCTTTTATTCCATGAGCCCTTGTTGTCTTTCTGCGCGC

GGTGCGCCGCCGGATGAGATGATGGAGACCCTAAGAAATATTGAAAGTACCGAGGCCGCTGCCGAGTTTGGGTTTATGG
GCCGCGCGACCTTGGCGGTCGCATCTGCCTGGTCACCCAATCCGAATAAGCCGCCGCTTTTCATCGACCTACCGGGCGAT
GAAGAGGTTGATGGCGACGTCATAGCCCGATGGGCGGCAAACGCACCGCTGTCTATGGTCGGTCAATACGTACCAGCCA
TGAAGACATATAAGGCCGGAGCCATCGATGTCGGTGATCAGGACGGCCTGAAAACAGATGCAGAAATGATGCATAAATT
GCTTGGAAAATACGGGGTCGATACGACTTTCGAGATTTACGAAGGTGATCATGTCAACCGGGTTCACATCCGGTTCGAAG
ATTACGTTTTGCCCTTTTTTGCGGCCAATTTGGAATTTGAATAG

Protein sequence (predicted signal peptide underlined):

MATTLSLASTAIN<u>AAAQD</u>ETSVTTQKVTIHSDAVEGNLEGNSAERDLLIYLPPSYDTDTDKRYPVIYGLHGYSIDNDQWSKEIQT
PTTIDAAFTDGVSEMIVVLPDSKTLHNGSMYSSSVTTGDFETFIAEDVVNYIDANYRTIPKRESRGLAGHSMGGYGTLRIAMKRP
DVFSSFYSMSPCCLSARGAPPDEMMETLRNIESTEAAAEFGFMGRATLAVASAWSPNPNKPPLFIDLPGDEEVDGDVIARWA
ANAPLSMVGQYVPAMKTYKAGAIDVGDQDGLKTDAEMMHKLLGKYGVDTTFEIYEGDHVNRVHIRFEDYVLPFFAANLEFE

List of abbreviations

AA	auxiliary activities
ABSL	acetyl bromide soluble lignin
AEBSF	4-benzenesulfonyl fluoride hydrochloride
AFases	Arabinofuranosidases
AI	auto-induction medium
AX	Arabinoxylans
AXE	acetyl xylan esterases
BLAST	basic local alignment search tool
bp	base pairs
BSA	Bovine Serum Albumin
CAZy	carbohydrate active enzyme database
CAZymes	Carbohydrate active enzymes
CBM	carbohydrate binding module
CE	carbohydrate Esterases
CTAB	cetyl trimethylammonium bromide
dbCAN	database for automated carbohydrate-active enzyme annotation
DEPC	diethyl pyrocarbonate
dH20	deionised water
DMF	Dimethylformamide
DNase	DeoxyriboNuclease
dNTP	deoxyriboNucleotide TriPhosphate
DTE	Dithiothreitol
DTT	Dithiothreitol
emPAI	exponentially modified protein abundance
Expasy	Expert Protein Analysis SYstem
FA	Ferulic acid
FAE	Feruloyl esterase
G	Guaiacyl
G II	Rhamnogalacturonan II
GalA	galacturonic acid
GAX	Glucuronoarabinoxylans
gDNA	genomic DNA
GH	glycoside Hydrolases
GlcA	glucuronic acid
GT	glycosyltransferases
H	p-hydroxyphenyls
HG	Homogalacturonan
HPAEC	High-Performance Anion-Exchange Chromatography
HPLC	high performance liquid chromatography
HRP	HorseRadish Peroxidase
IPTG	isopropyl ß-D-1-thiogalactopyanoside
kb	kilobase

kDa	kiloDalton
Km	The Michaelis constant
LB	Lysogeny Broth
LCC	lignin-carbohydrate complex
LC-MS/MS	liquid chromatography tandem mass spectrometry
LPMO	lytic polysaccharide monooxygenase
M	Molar
MFA	methyl ferulate
MLG	Mixed linkage glucans
mRNA	messenger RNA
NCBI	national center for biotechnology information
NCBInr	nCBI non-redundant protein database
OD	Optical Density
oNP-βXyl	ortho-nitrophenyl β-D-xylopyranoside
ORF	open reading frame
OTU	operational taxonomic unit
PBS	phosphate buffered saline
pCA	coumaric acid
PCR	polymerase chain reaction
PL	polysaccharide lyases
pNP	4-nitrophenyl
pNP-Ace	4-nitrophenyl Acetate
pNP-Ara	4-nitrophenyl-α-L-arabinofuranoside
pNP-Glc	4-nitrophenyl β-D-glucopyranoside
pNP-Xyl	4-nitrophenyl β-D-xylopyranoside
PSL	polyserine linker
PUL	polysaccharide utilisation loci
RG I	Rhamnogalacturonan I
RIN	RNA Integrity Number
RNase	RiboNuclease
rRNA	ribosomal RNA
S	Syringyls
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SOC	Super Optimal broth with Catabolite repression
SOD	superoxide dismutases
TBE	Tris-borate-EDTA buffer
TFA	triflouracetic acid
Tm	melting temperatures
v/v	volume to volume ratio
Vmax	maximum rate of reaction
w/v	weight to volume ratio
w/w	Weight to weight ratio
WB	Western Blot

XGA	xylogalacturonan
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
βglu	β-glucosidase
µM	microMolar

References

1. Hahn-Hagerdal, B., et al., *Bio-ethanol--the fuel of tomorrow from the residues of today.* Trends Biotechnol, 2006. **24**(12): p. 549-56.
2. Saini, J.K., R. Saini, and L. Tewari, *Lignocellulosic agriculture wastes as biomass feedstocks for second-generation bioethanol production: concepts and recent developments.* 3 Biotech, 2015. **5**(4): p. 337-353.
3. Li, Q., et al., *Plant biotechnology for lignocellulosic biofuel production.* Plant Biotechnol J, 2014. **12**(9): p. 1174-92.
4. Isikgor, F.H. and C.R. Becer, *Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers.* Polymer Chemistry, 2015. **6**(25): p. 4497-4559.
5. Sun, Y. and J. Cheng, *Hydrolysis of lignocellulosic materials for ethanol production: a review.* Bioresour Technol, 2002. **83**(1): p. 1-11.
6. Marriott, P.E., L.D. Gomez, and S.J. McQueen-Mason, *Unlocking the potential of lignocellulosic biomass through plant science.* New Phytol, 2016. **209**(4): p. 1366-81.
7. Gomez, L.D., C.G. Steele-King, and S.J. McQueen-Mason, *Sustainable liquid biofuels from biomass: the writing's on the walls.* New Phytol, 2008. **178**(3): p. 473-85.
8. Vogel, J., *Unique aspects of the grass cell wall.* Curr Opin Plant Biol, 2008. **11**(3): p. 301-7.
9. Klemm, D., et al., *Cellulose: fascinating biopolymer and sustainable raw material.* Angew Chem Int Ed Engl, 2005. **44**(22): p. 3358-93.
10. Gusakov, A.V., et al., *Design of highly efficient cellulase mixtures for enzymatic hydrolysis of cellulose.* Biotechnol Bioeng, 2007. **97**(5): p. 1028-38.
11. Reddy, N. and Y. Yang, *Biofibers from agricultural byproducts for industrial applications.* Trends Biotechnol, 2005. **23**(1): p. 22-7.
12. Kumar, R., S. Singh, and O.V. Singh, *Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives.* J Ind Microbiol Biotechnol, 2008. **35**(5): p. 377-391.
13. Zhao, X., L. Zhang, and D. Liu, *Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose.* 2012. **6**(4): p. 465-482.
14. Harris, D. and S. DeBolt, *Synthesis, regulation and utilization of lignocellulosic biomass.* Plant Biotechnol J, 2010. **8**(3): p. 244-62.
15. Tayeb, A.H., et al., *Cellulose Nanomaterials-Binding Properties and Applications: A Review.* Molecules, 2018. **23**(10).
16. Chundawat, S.P., et al., *Deconstruction of lignocellulosic biomass to fuels and chemicals.* Annu Rev Chem Biomol Eng, 2011. **2**: p. 121-45.
17. Scheller, H.V. and P. Ulvskov, *Hemicelluloses.* 2010. **61**(1): p. 263-289.
18. Park, Y.B. and D.J. Cosgrove, *Xyloglucan and its interactions with other components of the growing cell wall.* Plant Cell Physiol, 2015. **56**(2): p. 180-94.
19. Hatfield, R.D., D.M. Rancour, and J.M. Marita, *Grass Cell Walls: A Story of Cross-Linking.* Front Plant Sci, 2016. **7**: p. 2056.
20. Ebringerova, A. and T. Heinze, *Xylan and xylan derivatives - biopolymers with valuable properties, 1 - Naturally occurring xylans structures, procedures and properties.* Macromolecular Rapid Communications, 2000. **21**(9): p. 542-556.
21. Ralph, J., et al., *Identification and Synthesis of New Ferulic Acid Dehydrodimers Present in Grass Cell-Walls.* Journal of the Chemical Society-Perkin Transactions 1, 1994(23): p. 3485-3498.
22. Buanafina, M.M de O., *Feruloylation in grasses: current and future perspectives.* Mol Plant, 2009. **2**(5): p. 861-72.

23. de Oliveira, D.M., et al., *Ferulic acid: a key component in grass lignocellulose recalcitrance to hydrolysis.* Plant Biotechnol J, 2015. **13**(9): p. 1224-32.
24. Lagaert, S., et al., *beta-xylosidases and alpha-L-arabinofuranosidases: accessory enzymes for arabinoxylan degradation.* Biotechnol Adv, 2014. **32**(2): p. 316-32.
25. Busse-Wicher, M., et al., *The pattern of xylan acetylation suggests xylan may interact with cellulose microfibrils as a twofold helical screw in the secondary plant cell wall of Arabidopsis thaliana.* Plant J, 2014. **79**(3): p. 492-506.
26. de Gonzalo, G., et al., *Bacterial enzymes involved in lignin degradation.* J Biotechnol, 2016. **236**: p. 110-9.
27. Zhu, D., et al., *Biodegradation of alkaline lignin by Bacillus ligniniphilus L1.* Biotechnol Biofuels, 2017. **10**: p. 44.
28. Boerjan, W., J. Ralph, and M. Baucher, *Lignin biosynthesis.* Annu Rev Plant Biol, 2003. **54**: p. 519-46.
29. Sattler, S.E. and D.L. Funnell-Harris, *Modifying lignin to improve bioenergy feedstocks: strengthening the barrier against pathogens?* Front Plant Sci, 2013. **4**: p. 70.
30. Harholt, J., A. Suttangkakul, and H. Vibe Scheller, *Biosynthesis of pectin.* Plant Physiol, 2010. **153**(2): p. 384-95.
31. Xiao, C. and C.T. Anderson, *Roles of pectin in biomass yield and processing for biofuels.* Front Plant Sci, 2013. **4**: p. 67.
32. Mohnen, D., *Pectin structure and biosynthesis.* Curr Opin Plant Biol, 2008. **11**(3): p. 266-77.
33. De Souza, A.P., et al., *How cell wall complexity influences saccharification efficiency in Miscanthus sinensis.* J Exp Bot, 2015. **66**(14): p. 4351-65.
34. Latarullo, M.B., et al., *Pectins, Endopolygalacturonases, and Bioenergy.* Front Plant Sci, 2016. **7**: p. 1401.
35. Marcus, S.E., et al., *Pectic homogalacturonan masks abundant sets of xyloglucan epitopes in plant cell walls.* BMC Plant Biol, 2008. **8**: p. 60.
36. Marcus, S.E., et al., *Restricted access of proteins to mannan polysaccharides in intact plant cell walls.* Plant J, 2010. **64**(2): p. 191-203.
37. Jorgensen, H., J.B. Kristensen, and C. Felby, *Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities.* Biofuels Bioproducts & Biorefining-Biofpr, 2007. **1**(2): p. 119-134.
38. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* Nucleic Acids Res, 2014. **42**(Database issue): p. D490-5.
39. Henrissat, B., *A classification of glycosyl hydrolases based on amino acid sequence similarities.* Biochem J, 1991. **280 (Pt 2)**: p. 309-16.
40. Lombard, V., et al., *A hierarchical classification of polysaccharide lyases for glycogenomics.* Biochem J, 2010. **432**(3): p. 437-44.
41. Arnling Baath, J., et al., *Biochemical and structural features of diverse bacterial glucuronoyl esterases facilitating recalcitrant biomass conversion.* Biotechnol Biofuels, 2018. **11**: p. 213.
42. Nakamura, A.M., A.S. Nascimento, and I. Polikarpov, *Structural diversity of carbohydrate esterases.* Biotechnology Research and Innovation, 2017. **1**(1): p. 35-51.
43. Sista Kameshwar, A.K. and W. Qin, *Understanding the structural and functional properties of carbohydrate esterases with a special focus on hemicellulose deacetylating acetyl xylan esterases.* Mycology, 2018. **9**(4): p. 273-295.
44. Levasseur, A., et al., *Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes.* Biotechnol Biofuels, 2013. **6**(1): p. 41.
45. Boraston, A.B., et al., *Carbohydrate-binding modules: fine-tuning polysaccharide recognition.* Biochem J, 2004. **382**(Pt 3): p. 769-81.
46. Karboune, S., P.A. Geraert, and S. Kermasha, *Characterization of selected cellulolytic activities of multi-enzymatic complex system from Penicillium funiculosum.* J Agric Food Chem, 2008. **56**(3): p. 903-9.

47. Gupta, V.K., et al., *Fungal Enzymes for Bio-Products from Sustainable and Waste Biomass.* Trends Biochem Sci, 2016. **41**(7): p. 633-645.
48. Singh, G., A.K. Verma, and V. Kumar, *Catalytic properties, functional attributes and industrial applications of beta-glucosidases.* 3 Biotech, 2016. **6**(1): p. 3.
49. Dimarogona, M., E. Topakas, and P. Christakopoulos, *Cellulose degradation by oxidative enzymes.* Comput Struct Biotechnol J, 2012. **2**: p. e201209015.
50. Frandsen, K.E., et al., *The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases.* Nat Chem Biol, 2016. **12**(4): p. 298-303.
51. Andlar, M., et al., *Lignocellulose degradation: An overview of fungi and fungal enzymes involved in lignocellulose degradation.* Engineering in Life Sciences, 2018. **18**.
52. Xu, H., et al., *Characterization of a glucose-, xylose-, sucrose-, and D-galactose-stimulated beta-glucosidase from the alkalophilic bacterium Bacillus halodurans C-125.* Curr Microbiol, 2011. **62**(3): p. 833-9.
53. Meleiro, L.P., et al., *A novel beta-glucosidase from Humicola insolens with high potential for untreated waste paper conversion to sugars.* Appl Biochem Biotechnol, 2014. **173**(2): p. 391-408.
54. Cameron, R.G., et al., *Purification and characterization of a beta-glucosidase from Citrus sinensis var. Valencia fruit tissue.* J Agric Food Chem, 2001. **49**(9): p. 4457-62.
55. Pontoh, J. and N.H. Low, *Purification and characterization of beta-glucosidase from honey bees (Apis mellifera).* Insect Biochem Mol Biol, 2002. **32**(6): p. 679-90.
56. Harvey, A.J., et al., *Comparative modeling of the three-dimensional structures of family 3 glycoside hydrolases.* Proteins, 2000. **41**(2): p. 257-69.
57. Zang, X., et al., *The structural and functional contributions of beta-glucosidase-producing microbial communities to cellulose degradation in composting.* Biotechnol Biofuels, 2018. **11**: p. 51.
58. Gruno, M., et al., *Inhibition of the Trichoderma reesei cellulases by cellobiose is strongly dependent on the nature of the substrate.* Biotechnol Bioeng, 2004. **86**(5): p. 503-11.
59. Chamoli, S., et al., *Secretory expression, characterization and docking study of glucose-tolerant beta-glucosidase from B. subtilis.* Int J Biol Macromol, 2016. **85**: p. 425-33.
60. Salgado, J.C.S., et al., *Glucose tolerant and glucose stimulated beta-glucosidases - A review.* Bioresour Technol, 2018. **267**: p. 704-713.
61. Henrique Moreira Souza, F., et al., *Glucose and xylose stimulation of a β-glucosidase from the thermophilic fungus Humicola insolens: A kinetic and biophysical study.* Journal of Molecular Catalysis B: Enzymatic, 2013. **94**: p. 119-128.
62. de Giuseppe, P.O., et al., *Structural basis for glucose tolerance in GH1 beta-glucosidases.* Acta Crystallogr D Biol Crystallogr, 2014. **70**(Pt 6): p. 1631-9.
63. Guo, B., Y. Amano, and K. Nozaki, *Improvements in Glucose Sensitivity and Stability of Trichoderma reesei beta-Glucosidase Using Site-Directed Mutagenesis.* PLoS One, 2016. **11**(1): p. e0147301.
64. Zimbardi, A.L., et al., *Optimization of beta-glucosidase, beta-xylosidase and xylanase production by Colletotrichum graminicola under solid-state fermentation and application in raw sugarcane trash saccharification.* Int J Mol Sci, 2013. **14**(2): p. 2875-902.
65. Biely, P., *Microbial carbohydrate esterases deacetylating plant polysaccharides.* Biotechnol Adv, 2012. **30**(6): p. 1575-88.
66. Moreira, L.R. and E.X. Filho, *Insights into the mechanism of enzymatic hydrolysis of xylan.* Appl Microbiol Biotechnol, 2016. **100**(12): p. 5205-14.
67. Gündüz Ergün, B. and P. Çalık, *Lignocellulose degrading extremozymes produced by Pichia pastoris: current status and future prospects.* Bioprocess and Biosystems Engineering, 2016. **39**(1): p. 1-36.

68. Kumar, V., J. Marin-Navarro, and P. Shukla, *Thermostable microbial xylanases for pulp and paper industries: trends, applications and further perspectives.* World J Microbiol Biotechnol, 2016. **32**(2): p. 34.
69. Uday, U.S., et al., *Classification, mode of action and production strategy of xylanase and its application for biofuel production from water hyacinth.* Int J Biol Macromol, 2016. **82**: p. 1041-54.
70. Zhang, J., et al., *The role of acetyl xylan esterase in the solubilization of xylan and enzymatic hydrolysis of wheat straw and giant reed.* Biotechnology for Biofuels, 2011. **4**(1): p. 60.
71. Yang, X., et al., *Two xylose-tolerant GH43 bifunctional beta-xylosidase/alpha-arabinosidases and one GH11 xylanase from Humicola insolens and their synergy in the degradation of xylan.* Food Chem, 2014. **148**: p. 381-7.
72. Dos Santos, C.R., et al., *The mechanism by which a distinguishing arabinofuranosidase can cope with internal di-substitutions in arabinoxylans.* Biotechnol Biofuels, 2018. **11**: p. 223.
73. Phuengmaung, P., et al., *Identification and characterization of GH62 bacterial alpha-l-arabinofuranosidase from thermotolerant Streptomyces sp. SWU10 that preferentially degrades branched l-arabinofuranoses in wheat arabinoxylan.* Enzyme Microb Technol, 2018. **112**: p. 22-28.
74. Chavez Montes, R.A., et al., *Cell wall modifications in Arabidopsis plants with altered alpha-L-arabinofuranosidase activity.* Plant Physiol, 2008. **147**(1): p. 63-77.
75. Bouraoui, H., et al., *The GH51 alpha-l-arabinofuranosidase from Paenibacillus sp. THS1 is multifunctional, hydrolyzing main-chain and side-chain glycosidic bonds in heteroxylans.* Biotechnol Biofuels, 2016. **9**: p. 140.
76. Lagaert, S., et al., *Substrate specificity of three recombinant alpha-L-arabinofuranosidases from Bifidobacterium adolescentis and their divergent action on arabinoxylan and arabinoxylan oligosaccharides.* Biochem Biophys Res Commun, 2010. **402**(4): p. 644-50.
77. van den Broek, L.A., et al., *Cloning and characterization of arabinoxylan arabinofuranohydrolase-D3 (AXHd3) from Bifidobacterium adolescentis DSM20083.* Appl Microbiol Biotechnol, 2005. **67**(5): p. 641-7.
78. Sorensen, H.R., et al., *A novel GH43 alpha-L-arabinofuranosidase from Humicola insolens: mode of action and synergy with GH51 alpha-L-arabinofuranosidases on wheat arabinoxylan.* Appl Microbiol Biotechnol, 2006. **73**(4): p. 850-61.
79. Wilkens, C., et al., *GH62 arabinofuranosidases: Structure, function and applications.* Biotechnol Adv, 2017. **35**(6): p. 792-804.
80. Beylot, M.H., et al., *The Pseudomonas cellulosa glycoside hydrolase family 51 arabinofuranosidase exhibits wide substrate specificity.* Biochem J, 2001. **358**(Pt 3): p. 607-14.
81. Alvira, P., M.J. Negro, and M. Ballesteros, *Effect of endoxylanase and alpha-L-arabinofuranosidase supplementation on the enzymatic hydrolysis of steam exploded wheat straw.* Bioresour Technol, 2011. **102**(6): p. 4552-8.
82. Selig, M.J., et al., *Debranching of soluble wheat arabinoxylan dramatically enhances recalcitrant binding to cellulose.* Biotechnol Lett, 2015. **37**(3): p. 633-41.
83. Dilokpimol, A., et al., *Diversity of fungal feruloyl esterases: updated phylogenetic classification, properties, and industrial applications.* Biotechnology for Biofuels, 2016. **9**(1): p. 231.
84. Xu, Z., et al., *Characterization of Feruloyl Esterases Produced by the Four Lactobacillus Species: L. amylovorus, L. acidophilus, L. farciminis and L. fermentum, Isolated from Ensiled Corn Stover.* Front Microbiol, 2017. **8**: p. 941.
85. Selig, M.J., et al., *Synergistic enhancement of cellobiohydrolase performance on pretreated corn stover by addition of xylanase and esterase activities.* Bioresour Technol, 2008. **99**(11): p. 4997-5005.

86. Tabka, M.G., et al., *Enzymatic saccharification of wheat straw for bioethanol production by a combined cellulase xylanase and feruloyl esterase treatment.* Enzyme and Microbial Technology, 2006. **39**(4): p. 897-902.
87. Gottschalk, L.M.F., R.A. Oliveira, and E.P.d.S. Bon, *Cellulases, xylanases, β-glucosidase and ferulic acid esterase produced by Trichoderma and Aspergillus act synergistically in the hydrolysis of sugarcane bagasse.* Biochemical Engineering Journal, 2010. **51**(1): p. 72-78.
88. Benoit, I., et al., *Biotechnological applications and potential of fungal feruloyl esterases based on prevalence, classification and biochemical diversity.* 2008. **30**(3): p. 387-396.
89. Fang, C., et al., *Seawater as Alternative to Freshwater in Pretreatment of Date Palm Residues for Bioethanol Production in Coastal and/or Arid Areas.* ChemSusChem, 2015. **8**(22): p. 3823-31.
90. Vörösmarty, C.J., et al., *Global Water Resources: Vulnerability from Climate Change and Population Growth.* Science, 2000. **289**(5477): p. 284.
91. Domínguez de María, P., *On the Use of Seawater as Reaction Media for Large-Scale Applications in Biorefineries.* 2013. **5**(7): p. 1643-1648.
92. Yu, Q., et al., *The effect of metal salts on the decomposition of sweet sorghum bagasse in flow-through liquid hot water.* Bioresour Technol, 2011. **102**(3): p. 3445-50.
93. Liu, L., et al., *Corn stover pretreatment by inorganic salts and its effects on hemicellulose and cellulose degradation.* Bioresour Technol, 2009. **100**(23): p. 5865-71.
94. Gaur, R., et al., *Isolation, Production, and Characterization of Thermotolerant Xylanase from Solvent Tolerant Bacillus vallismortis RSPP-15 %J International Journal of Polymer Science.* 2015. **2015**: p. 10.
95. Zhuo, R., et al., *Induction of laccase by metal ions and aromatic compounds in Pleurotus ostreatus HAUCC 162 and decolorization of different synthetic dyes by the extracellular laccase.* Biochemical Engineering Journal, 2017. **117**: p. 62-72.
96. Karan, R., M.D. Capes, and S. Dassarma, *Function and biotechnology of extremophilic enzymes in low water activity.* Aquat Biosyst, 2012. **8**(1): p. 4.
97. Zaccai, G., *The effect of water on protein dynamics.* Philos Trans R Soc Lond B Biol Sci, 2004. **359**(1448): p. 1269-75; discussion 1275, 1323-8.
98. Baldwin, R.L., *How Hofmeister ion interactions affect protein stability.* Biophys J, 1996. **71**(4): p. 2056-63.
99. Pennings, S.C. and M. Bertness, *Salt marsh communities*. 2001. p. 289-316.
100. Boorman, L., *Saltmarsh Review: An overview of coastal saltmarshes, their dynamic and sensitivity characteristics for conservation and management.* JNCC Report, 2003. **334**: p. 132.
101. Dini-Andreote, F., et al., *Dynamics of bacterial community succession in a salt marsh chronosequence: evidences for temporal niche partitioning.* ISME J, 2014. **8**(10): p. 1989-2001.
102. Bowen, J.L., et al., *Salt marsh sediment diversity: a test of the variability of the rare biosphere among environmental replicates.* The Isme Journal, 2012. **6**: p. 2014.
103. Rietl, A.J., et al., *Microbial Community Composition and Extracellular Enzyme Activities Associated with Juncus roemerianus and Spartina alterniflora Vegetated Sediments in Louisiana Saltmarshes.* Microb Ecol, 2016. **71**(2): p. 290-303.
104. Bowen, J.L., et al., *Salt marsh sediment bacteria: their distribution and response to external nutrient inputs.* ISME J, 2009. **3**(8): p. 924-34.
105. Fukushima, R.S. and R.D. Hatfield, *Extraction and isolation of lignin for utilization as a standard to determine lignin concentration using the acetyl bromide spectrophotometric method.* J Agric Food Chem, 2001. **49**(7): p. 3133-9.
106. Foster, C.E., T.M. Martin, and M. Pauly, *Comprehensive compositional analysis of plant cell walls (lignocellulosic biomass) part II: carbohydrates.* J Vis Exp, 2010(37).

107. Leyva, A., et al., *Rapid and sensitive anthrone-sulfuric acid assay in microplate format to quantify carbohydrate in biopharmaceutical products: method development and validation.* Biologicals, 2008. **36**(2): p. 134-41.
108. Parada, A.E., D.M. Needham, and J.A. Fuhrman, *Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples.* 2016. **18**(5): p. 1403-1414.
109. Nelson, K.A., N.S. Moin, and A.E. Bernhard, *Archaeal diversity and the prevalence of Crenarchaeota in salt marsh sediments.* Appl Environ Microbiol, 2009. **75**(12): p. 4211-5.
110. Seyler, L.M., L.M. McGuinness, and L.J. Kerkhof, *Crenarchaeal heterotrophy in salt marsh sediments.* ISME J, 2014. **8**(7): p. 1534-43.
111. Apprill, A., et al., *Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton.* Aquatic Microbial Ecology, 2015. **75**(2): p. 129-137.
112. Lundberg, D.S., et al., *Practical innovations for high-throughput amplicon sequencing.* Nat Methods, 2013. **10**(10): p. 999-1002.
113. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4**: p. e2584.
114. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. pp. 10-12.
115. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-2461.
116. Edgar, R.C., *UPARSE: highly accurate OTU sequences from microbial amplicon reads.* Nature Methods, 2013. **10**(10): p. 996-+.
117. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
118. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
119. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
120. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
121. Alessi, A.M., et al., *Revealing the insoluble metasecretome of lignocellulose-degrading microbial communities.* Scientific Reports, 2017. **7**(1): p. 2356.
122. Dowle, A.A., J. Wilson, and J.R. Thomas, *Comparing the Diagnostic Classification Accuracy of iTRAQ, Peak-Area, Spectral-Counting, and emPAI Methods for Relative Quantification in Expression Proteomics.* J Proteome Res, 2016. **15**(10): p. 3550-3562.
123. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol Cell Proteomics, 2005. **4**(9): p. 1265-72.
124. Yin, Y., et al., *dbCAN: a web resource for automated carbohydrate-active enzyme annotation.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W445-51.
125. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Anal Biochem, 1976. **72**: p. 248-54.
126. McIlvaine, T.C., *A buffer solution for colorimetric comparison.* Journal of Biological Chemistry, 1921. **49**(1): p. 183-186.
127. Zhao, X.B., L.H. Zhang, and D.H. Liu, *Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose.* Biofuels Bioproducts & Biorefining-Biofpr, 2012. **6**(4): p. 465-482.
128. de Gonzalo, G., et al., *Bacterial enzymes involved in lignin degradation.* Journal of Biotechnology, 2016. **236**: p. 110-119.
129. Zhu, D., et al., *Biodegradation of alkaline lignin by Bacillus ligniniphilus L1.* Biotechnology for Biofuels, 2017. **10**(1): p. 44.

130. Jørgensen, H., J.B. Kristensen, and C. Felby, *Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities.* Biofuels, Bioproducts and Biorefining, 2007. **1**(2): p. 119-134.
131. Cragg, S.M., et al., *Lignocellulose degradation mechanisms across the Tree of Life.* Curr Opin Chem Biol, 2015. **29**: p. 108-19.
132. Dutta, S. and K.C.W. Wu, *Enzymatic breakdown of biomass: enzyme active sites, immobilization, and biofuel production.* Green Chemistry, 2014. **16**(11): p. 4615-4626.
133. Daniel, R., *The metagenomics of soil.* Nat Rev Micro, 2005. **3**(6): p. 470-478.
134. Warnecke, F. and M. Hess, *A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts.* Journal of Biotechnology, 2009. **142**(1): p. 91-95.
135. Hettich, R.L., et al., *Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities.* Anal Chem, 2013. **85**(9): p. 4203-14.
136. van Vliet, A.H., *Next generation sequencing of microbial transcriptomes: challenges and opportunities.* FEMS Microbiol Lett, 2010. **302**(1): p. 1-7.
137. Rashid, G.M., et al., *Identification of manganese superoxide dismutase from Sphingobacterium sp. T2 as a novel bacterial enzyme for lignin oxidation.* ACS Chem Biol, 2015. **10**.
138. Simmons, C.W., et al., *Metatranscriptomic analysis of lignocellulolytic microbial communities involved in high-solids decomposition of rice straw.* Biotechnol Biofuels, 2014. **7**(1): p. 495.
139. Strachan, C.R., et al., *Metagenomic scaffolds enable combinatorial lignin transformation.* Proc Natl Acad Sci U S A, 2014. **111**(28): p. 10143-8.
140. Fang, Z.M., et al., *A new marine bacterial laccase with chloride-enhancing, alkaline-dependent activity and dye decolorization ability.* Bioresour Technol, 2012. **111**: p. 36-41.
141. Noinaj, N., et al., *TonB-dependent transporters: regulation, structure, and function.* Annu Rev Microbiol, 2010. **64**: p. 43-60.
142. Zhu, B. and H. Yin, *Alginate lyase: Review of major sources and classification, properties, structure-function analysis and applications.* Bioengineered, 2015. **6**(3): p. 125-31.
143. Akileswaran, L., et al., *1,4-benzoquinone reductase from Phanerochaete chrysosporium: cDNA cloning and regulation of expression.* Appl Environ Microbiol, 1999. **65**(2): p. 415-21.
144. Rashid, G.M.M., et al., *Identification of Manganese Superoxide Dismutase from Sphingobacterium sp. T2 as a Novel Bacterial Enzyme for Lignin Oxidation.* ACS Chemical Biology, 2015. **10**(10): p. 2286-2294.
145. Rashid, G.M.M., et al., *Sphingobacterium sp. T2 Manganese Superoxide Dismutase Catalyzes the Oxidative Demethylation of Polymeric Lignin via Generation of Hydroxyl Radical.* ACS Chemical Biology, 2018. **13**(10): p. 2920-2929.
146. Alalouf, O., et al., *A new family of carbohydrate esterases is represented by a GDSL hydrolase/acetylxylan esterase from Geobacillus stearothermophilus.* J Biol Chem, 2011. **286**(49): p. 41993-2001.
147. Lapébie, P., et al., *Bacteroidetes use thousands of enzyme combinations to break down glycans.* Nature Communications, 2019. **10**(1): p. 2043.
148. Andersson, S.G., et al., *The genome sequence of Rickettsia prowazekii and the origin of mitochondria.* Nature, 1998. **396**(6707): p. 133-40.
149. Newton, R.J., et al., *A guide to the natural history of freshwater lake bacteria.* Microbiol Mol Biol Rev, 2011. **75**(1): p. 14-49.
150. Morris, R.M., et al., *SAR11 clade dominates ocean surface bacterioplankton communities.* Nature, 2002. **420**(6917): p. 806-10.
151. Biers, E.J., S. Sun, and E.C. Howard, *Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome.* Applied and Environmental Microbiology, 2009. **75**(7): p. 2221.

152. Biely, P., *Microbial Glucuronoyl Esterases: 10 Years after Discovery.* Applied and Environmental Microbiology, 2016. **82**(24): p. 7014.
153. Bork, P. and A. Bairoch, *Go hunting in sequence databases but watch out for the traps.* Trends Genet, 1996. **12**(10): p. 425-7.
154. Terpe, K., *Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems.* Applied Microbiology and Biotechnology, 2006. **72**(2): p. 211.
155. Hannig, G. and S.C. Makrides, *Strategies for optimizing heterologous protein expression in Escherichia coli.* Trends in Biotechnology, 1998. **16**(2): p. 54-60.
156. Kaur, J., A. Kumar, and J. Kaur, *Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements.* Int J Biol Macromol, 2018. **106**: p. 803-822.
157. Shen, H., et al., *Deletion of the linker connecting the catalytic and cellulose-binding domains of endoglucanase A (CenA) of Cellulomonas fimi alters its conformation and catalytic activity.* J Biol Chem, 1991. **266**(17): p. 11335-40.
158. Rixon, J.E., et al., *Do the non-catalytic polysaccharide-binding domains and linker regions enhance the biobleaching properties of modular xylanases?* Appl Microbiol Biotechnol, 1996. **46**(5-6): p. 514-20.
159. Howard, M.B., et al., *Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium Microbulbifer degradans.* Protein science : a publication of the Protein Society, 2004. **13**(5): p. 1422-1425.
160. Schallus, T., et al., *Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation.* Mol Biol Cell, 2008. **19**(8): p. 3404-14.
161. Nothaft, H. and C.M. Szymanski, *Protein glycosylation in bacteria: sweeter than ever.* Nature Reviews Microbiology, 2010. **8**: p. 765.
162. Mortimer, J.C., et al., *Absence of branches from xylan in Arabidopsis gux mutants reveals potential for simplification of lignocellulosic biomass.* Proc Natl Acad Sci U S A, 2010. **107**(40): p. 17409-14.
163. Várnai, A., et al., *Effects of enzymatic removal of plant cell wall acylation (acetylation, p-coumaroylation, and feruloylation) on accessibility of cellulose and xylan in natural (non-pretreated) sugar cane fractions.* Biotechnology for Biofuels, 2014. **7**(1): p. 153.
164. Wagschal, K., et al., *Genetic and biochemical characterization of an α-l-arabinofuranosidase isolated from a compost starter mixture.* Enzyme and Microbial Technology, 2007. **40**: p. 747-753.
165. Carvalho, D.R.d., et al., *A halotolerant bifunctional β-xylosidase/α-l-arabinofuranosidase from Colletotrichum graminicola: Purification and biochemical characterization.* International Journal of Biological Macromolecules, 2018. **114**: p. 741-750.
166. Xu, B., et al., *Characterization of a novel salt-, xylose- and alkali-tolerant GH43 bifunctional beta-xylosidase/alpha-l-arabinofuranosidase from the gut bacterial genome.* J Biosci Bioeng, 2019.
167. Oliveira, D.M., et al., *Feruloyl esterases: Biocatalysts to overcome biomass recalcitrance and for the production of bioactive compounds.* Bioresour Technol, 2019. **278**: p. 408-423.
168. Hunt, C.J., A. Tanksale, and V.S. Haritos, *Biochemical characterization of a halotolerant feruloyl esterase from Actinomyces spp.: refolding and activity following thermal deactivation.* Appl Microbiol Biotechnol, 2016. **100**(4): p. 1777-87.
169. Nieter, A., et al., *A halotolerant type A feruloyl esterase from Pleurotus eryngii.* Fungal Biol, 2014. **118**(3): p. 348-57.
170. Sigoillot, J.-C., et al., *Fungal Strategies for Lignin Degradation.* Advances in Botanical Research, 2012. **61**: p. 263-308.
171. Dashtban, M., et al., *Fungal biodegradation and enzymatic modification of lignin.* Int J Biochem Mol Biol, 2010. **1**(1): p. 36-50.

172. Bugg, T.D., et al., *Pathways for degradation of lignin in bacteria and fungi.* Nat Prod Rep, 2011. **28**(12): p. 1883-96.
173. Mohorcic, M., et al., *Expression of soluble versatile peroxidase of Bjerkandera adusta in Escherichia coli.* Bioresour Technol, 2009. **100**(2): p. 851-8.
174. Fernandez-Fueyo, E., et al., *Comparative genomics of Ceriporiopsis subvermispora and Phanerochaete chrysosporium provide insight into selective ligninolysis.* Proc Natl Acad Sci U S A, 2012. **109**(14): p. 5458-63.
175. Biswal, A.K., et al., *Sugar release and growth of biofuel crops are improved by downregulation of pectin biosynthesis.* Nature Biotechnology, 2018. **36**: p. 249.
176. Lionetti, V., et al., *Engineering the cell wall by reducing de-methyl-esterified homogalacturonan improves saccharification of plant tissues for bioconversion.* Proceedings of the National Academy of Sciences, 2010. **107**(2): p. 616.
177. Singh, S.A., H. Plattner, and H. Diekmann, *Exopolygalacturonate lyase from a thermophilic Bacillus sp.* Enzyme and Microbial Technology, 1999. **25**(3): p. 420-425.
178. Marín-Rodríguez, M.C., J. Orchard, and G.B. Seymour, *Pectate lyases, cell wall degradation and fruit softening.* Journal of Experimental Botany, 2002. **53**(377): p. 2115-2119.
179. Kester, H.C., et al., *Performance of selected microbial pectinases on synthetic monomethyl-esterified di- and trigalacturonates.* J Biol Chem, 1999. **274**(52): p. 37053-9.
180. Micheli, F., *Pectin methylesterases: cell wall enzymes with important roles in plant physiology.* Trends Plant Sci, 2001. **6**(9): p. 414-9.
181. Lyczakowski, J.J., et al., *Removal of glucuronic acid from xylan is a strategy to improve the conversion of plant biomass to sugars for bioenergy.* Biotechnology for Biofuels, 2017. **10**(1): p. 224.
182. Marriott, P.E., et al., *Range of cell-wall alterations enhance saccharification in Brachypodium distachyon mutants.* Proc Natl Acad Sci U S A, 2014. **111**(40): p. 14601-6.
183. Chiniquy, D., et al., *XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan.* Proc Natl Acad Sci U S A, 2012. **109**(42): p. 17117-22.
184. Arnling Baath, J., et al., *A glucuronoyl esterase from Acremonium alcalophilum cleaves native lignin-carbohydrate ester bonds.* FEBS Lett, 2016. **590**(16): p. 2611-8.
185. Ryabova, O., et al., *A novel family of hemicellulolytic alpha-glucuronidase.* FEBS Lett, 2009. **583**(9): p. 1457-62.
186. Tadeo, X., et al., *Structural basis for the aminoacid composition of proteins from halophilic archea.* PLoS Biol, 2009. **7**(12): p. e1000257.
187. Paul, S., et al., *Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes.* Genome Biol, 2008. **9**(4): p. R70.
188. Siddiqui, K.S. and R. Cavicchioli, *Cold-adapted enzymes.* Annu Rev Biochem, 2006. **75**: p. 403-33.
189. Cavicchioli, R., et al., *Biotechnological uses of enzymes from psychrophiles.* Microb Biotechnol, 2011. **4**(4): p. 449-60.