Physiology and enzymology of lignocellulose digestion in the shipworm *Lyrodus pedicellatus*

Giovanna Pesante

PhD

University of York

Biology

September 2018

Abstract

Shipworms are marine bivalve molluscs, known for their wood boring abilities. They use modified shells to bore into and grind wood, which is then digested. The shipworm's ability to feed on lignocellulose is dependent on the presence of endosymbiotic bacteria that live in the animal's gills inside specialized eukaryotic cells called bacteriocytes. These bacteria provide the animal with hydrolytic enzymes for wood digestion, which are translocated from the gills to the caecum, the main site of wood digestion. Unlike other lignocellulose degrading organisms, which harbour symbiotic microbes in their digestive tract, the shipworm caecum hosts only few bacteria but contains a large amount of carbohydrate active enzymes (CAZymes) of both endogenous and bacterial origin.

This study investigates the anatomical, physiological and molecular basis of wood digestion in the shipworm *Lyrodus pedicellatus*. A combination of meta-transcriptomics, meta-proteomics and microscopic studies of the shipworm digestive organs was used, coupled with recombinant production and characterisation of some of the most expressed lignocellulolytic enzymes.

This multidisciplinary analysis revealed how two structures, the food groove (a mucus stream utilised by filter feeding molluscs to transport food particles from the gills to the digestive system) and the crystalline style (a rotating structure hosted in the stomach involved in extra-cellular digestion) have been co-opted in shipworms to translocate bacteria and their enzymes from the gills to the caecum, to facilitate wood digestion. The transcriptomic and proteomic results indicate that bacterial lignocellulolytic enzymes are expressed in the gills, while the endogenous enzymes are mainly produced by the digestive glands, with complementary CAZy classes being expressed by the bacteria and the shipworms, indicating a subdivision of roles. Five bacterial CAZymes were recombinantly expressed and characterised, showing activity on cellulose, galacto-glucomannan and xylan. This study provides new insights into the mechanisms of wood digestion in shipworms, which may have biotechnological relevance.

List of contents

List of tables

List of figures

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisors, Prof. Simon McQueen-Mason and Prof. Neil Bruce, for giving me the opportunity to join the "wood borers project", of which this PhD thesis is part. They took a leap of faith in giving this project to me, a Marine biologist who became a Paramedic and then turned into a Lab Scientist. I really appreciate their perhaps unconventional choice, and I hope I have not made them regret it!

I am very grateful to my sponsor, the UK Biotechnology and Biological Sciences Research Council (BBSRC), for providing the funding that allowed me to carry out this work, and the White Rose BBSRC Doctoral Training Partnership for allowing me to join their training scheme.

Many thanks to my TAP supervisors, Robert White and Michael Plevin, for their assistance and encouragement throughout these four years, and for the invaluable advices, comments and revisions they provided for many TAP reports.

In CNAP (Centre for Novel Agricultural Project), the research centre where the work was carried out, I have too many people to acknowledge. I have never worked in such a friendly, supportive and relaxed environment, where high standard research is carried out nevertheless this laid-back approach. I am extremely grateful to Federico Sabbadin, who passed me some of his extensive knowledge of molecular biology, biochemistry and bioinformatics and taught me all I now know. Luisa Elias has been a great friend and colleague, with whom I shared many laughs around the lab benches, and has been mine (and the whole lab's) reference point throughout my PhD. Katrin Besser was another invaluable font of knowledge and a model for perseverance and dedication. Other colleagues to mention, in no particular order, are: Claire-Steele-King (without whom no microscopy work would have been possible), Yi Li, Rachael Hallam, Leonardo Gomez, Laura Faas, Carla Botehlo Machado, Julia Crawford, Caragh Whitehead, Juliana Sanchez Alponti, Linda Sainty, Joe Bennett, Alexandra Lanot, Aritha Dornau, Daniel Leadbeater (thank you for your help with the heat maps), David Neale, Liz Rylott, Thierry Tonon, Margaret Cafferky, Nicola Oates, Veronica Ongaro, Anna Alessi and Susannah Bird.

Many thanks go to the Institute of Marine Science staff at Portsmouth University (in particular Simon Cragg (I have learnt so much from you about shipworm biology and taxonomy), Elizabeth Clutton, Amaia Etxabe, Lourdes Cruz Garcia and Graham Maylon) who took care of the shipworm rearing, retrieval and dissections. The staff of the Istituto di Science Marine in Venice (ISMAR-CNR) was extremely helpful (mainly Davide Tagliapietra, Marco Sigovini and Irene Guarnieri) and even if the data collected in Venice have not been included in this dissertation, their knowledge, hospitality and helpfulness won't be forgotten.

The Technology Facility staff at York University has been extremely helpful with many aspects of the data analysis process. In particular, I would like to mention Jared Cartwright, Rebecca Preece and Mick Miller for their friendly support with protein production; Adam Dowle, Rachel Bates and Swen Langer for their help with the proteomic analysis; Peter O'Toole, Meg Stark, Graeme Park, Karen Hogg, Jo Marrison and Karen Hodgkinson from the imaging and cytometry lab; Peter Ashton and John Davey for support with genomics and bioinformatics analysis.

Staff at the Biorenewable Development Centre in Dunnington, York, made my PIPs project stimulating and enjoyable, and I am grateful to them (Deborah Rathbone, Alexandra Jukes, Rosie Nolan, Susan Heywood, David Vaughan and Joe Bennett).

I am thankful to the Oxford Protein Production Facility UK for allowing me to use their specialised high throughput equipment for protein expression and purification. Without the help of Ray Owens, Luise Bird, Valtteri Jarvinen and Joanne Nettleship, *Lps*GH11 would have never been expressed.

The Next Generation Sequencing Facility staff at the University of Leeds, namely Ian Carr, Sally Harrison, Carolina Lascelles and Ummey Hany, was extremely helpful and competent in their assistance with the RNA sequencing.

Many thanks also to Paul Dupree, Marta Busse-Wicher and Rita Delgado Silva Marques from the Department of Biochemistry in Cambridge, who carried out the PACE analysis.

Acknowledgements are also owed to all the support staff at York University. There are too many names to mention, but the people that have been particularly friendly and helpful are Lorna Warnock, Amanda Barnes, Andrew Smith, Monica Bandeira, Fiona Scaife, Paul Genever, Emma Rand, Betsy Pownall and David Nelmes.

Last, but not least, I would like to thank my family (my mother Bianca, my father Vitaliano and my sister Giulia) for always supporting *ciccia biologa* in her maybe unusual and frustrating choices, and for always being there when needed. Sadly, my father did not make it to see me as Dr Pesante, but I know he would have been very proud. My partner Robert Foxon was also supportive and sympathetic throughout my PhD life, thank you!

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.

The following publication has arisen from this work:

Sabbadin, F., Pesante, G., Elias, L., Besser, K., Li, Y., Steele-King, C., Stark, M., Rathbone, D.A., Dowle, A.A. and Bates, R. 2018. Uncovering the molecular mechanisms of lignocellulose digestion in shipworms. *Biotechnology for biofuels*, 11, 59.

Please find a reprint of this publication at the end of the thesis.

Signature:

Chapter 1: General introduction

1.1 Biofuels and the plant cell wall

1.1.1 The need for biofuels

Greenhouse gases (GHGs) causing global warming are a topical issue. It has been estimated that the average global surface temperature of Earth has increased of 0.8 °C since 1880, with most of the warming occurring since 1975 (Hansen *et al.*, 2010), an increase that is directly linked to human activities rather than natural fluctuations due to weather variability. This global warming, if not kept under control, is likely to have devastating effects on parts of the Planet. The Agreement signed in Paris in 2015 by the United Nations Framework Convention on Climate Change aimed at reducing GHG emissions to net zero by 2300 (U.N.F.C.C.C., 2015). However, it is foreseen that the gases already accumulated in the atmosphere will cause a global temperature increase despite the GHGs reduction, with a consequent rise of the ocean levels up to 1.5 m, due to the melting of the polar ice sheets (Mengel *et al.*, 2018).

Among the GHGs, carbon dioxide (CO_2) is produced in highest quantities, and the amounts recorded in the atmosphere are directly proportional - and the main causal driver - to the temperature increase registered since 1850 (Stips *et al.*, 2016). CO_2 production is directly linked to the burning of fossil fuels, which are estimated to be responsible for 80% of the anthropogenic GHG emission (Quadrelli and Peterson, 2007), a figure that is predicted to rise rapidly due to the increase of the human population and the consequent global energy demand (U.N., 2013). Fossil fuels are, therefore, one of the most significant contributors to climate change. The environmental problems caused by the burning of fossil fuels, coupled with their limited availability - oil, gas and coal are foreseen to be exhausted in less than 100 years (Shafiee and Topal, 2009) - have led to a growing effort being put into the discovery and production of fuels that not only are more sustainable and less polluting, but also renewable and available in large quantities. One of the best alternatives to fossil fuels are biofuels such as bioethanol, biodiesel or biogas obtained by fermenting the sugars contained in plant material. Indeed, biomass is available in large quantities, it is inexpensive and renewable, and biofuels have the potential to be carbon neutral, meaning that the CO_2 released during their use equals the amount that is re-absorbed from the air by growing plants (Gomez *et al.*, 2008). In light of these advantages, the European Union in 2007

agreed that at least 10% of the transport fuels in Europe must come from renewable resources by 2020 (E.U., 2007; Klessmann *et al.*, 2011), and the United States in 2008 increased of almost five times the set production target of renewable transport fuels (I.E.A., 2008).

1.1.2 First- and second-generation biofuels

The first non-fossil based fuels to be utilised at a commercial scale have been produced by fermentation of sugars that are easily accessible in sugar cane, sugar beet, corn or starch (Gomez *et al.*, 2008). These first generation biofuels, and particularly bioethanol, have now become familiar commodities worldwide. Global biofuel production showed almost a fivefold increase from 2002 to 2012, with the USA and Brazil being the major producers, followed by Europe and Asia (UNCTAD, 2016). However, the fact that the crops used for the production of these biofuels are the same as those used by the food industry has generated a heated "food versus fuel" debate, particularly after a food price increase in the USA (Odling-Smee, 2007). Furthermore, recent research has highlighted the high amount of fertilisers and pesticides needed for the production of these sugar and starch crops, as well as the fact that these crops are not really carbon neutral but have a net CO_2 emission, even if lower than burning fossil fuels (Hill *et al.*, 2006; Waldrop, 2007; DeCicco *et al.*, 2016). As a result of these economic and environmental issues, and also as a consequence of a recent drop of oil prices, the biofuel production growth rates have decreased worldwide, and the annual production in 2015 has indeed plummeted to values similar to 2010 (UNCTAD, 2016).

A more environmentally sustainable alternative are the second-generation biofuels, which are produced from biomass material not used by the food industry. These alternative feedstocks range from agricultural by-products (unused plant stems, straw and wood shavings, sugar cane bagasse) to municipal solid waste and dedicated crops grown on marginal or degraded land (Carroll and Somerville, 2009; Li *et al.*, 2014). Second-generation biofuels started to be marketed on a commercial level between 2013 and 2015, and it is estimated that by 2020 there will be an increase of 50% of their market (UNCTAD, 2016). However, second-generation biofuel production is not at present cost-effective, due the recalcitrant nature of the lignocellulose that comprises the majority of woody biomass (see next section, 1.1.3) (Chen and Dixon, 2007). The practice of processing the polysaccharide constituents of lignocellulose into simple sugars (called saccharification) on an industrial

scale has been found to be challenging and expensive (Yang and Wyman, 2008). Indeed, plants' lignocellulosic tissues are strong, tough and resistant to chemical and biological attack, which renders their sugar contents difficult to access. For this reason, harsh chemical and physical pretreatments are needed to render the plant material more accessible (Mosier *et al.*, 2005; Galbe and Zacchi, 2007), and then a complex mixture of different cellulolytic enzymes are required to hydrolyse the carbohydrates, producing monosaccharides or short oligosaccharides (Jørgensen *et al.*, 2007). A cost-effective combination of pretreatments and enzymes has still not been found, mainly due to high costs and low enzyme efficiency, despite significant new progress being continuously made (Sims *et al.*, 2010). Considerable current research is focused on the discovery and characterization of novel cellulases and associated enzymes from bacteria, fungi and invertebrates in order to develop more efficient approaches to lignocellulose saccharification and therefore second-generation biofuels production.

1.1.3 Lignocellulosic biomass

Second generation biofuels are obtained from lignocellulosic biomass, which comprises of plant secondary cell walls, the thick layer that the cells lay down inside the primary cell walls to strengthen plant structure. Plants produce the primary cell wall during growth, while the secondary wall is formed once expansion has ceased, and its function is to make the plant mechanically strong and resistant to external and internal stress (Zhong *et al.*, 2010). The secondary cell wall principally comprises cellulose, hemicellulose and lignin (Figure 1.1), with pectins and structural proteins also being an important part of its structure, despite being less abundant than the other components (Vogel, 2008). The composition and structure of the secondary cell wall makes it resilient to chemical, physical and biological attack, therefore it is challenging to extract from it the monosaccharides that can be fermented into bioethanol, hence the challenge for the second-generation biofuels (Wingren *et al.*, 2003; Harris and DeBolt, 2010).

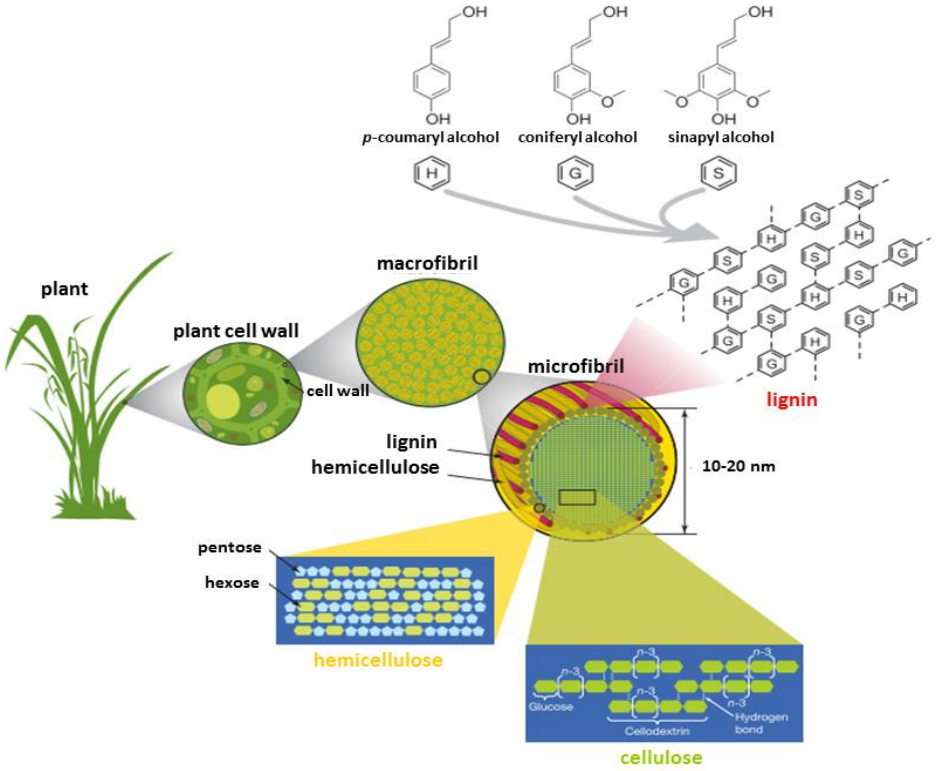


Figure 1.1. Plant cell wall localisation and structure. Figure adapted from Rubin (2008).

1.1.4 Cell wall composition

1.1.4.1 Cellulose

Cellulose is the main structural component of the plant cell wall and can make up to 35-50% of its dry weight (Pauly and Keegstra, 2008), making it the most abundant organic polymer on Earth (Klemm *et al.*, 2005). It is composed of unbranched parallel chains of β-1,4 linked glucans, which are polymers of pure glucose. Consecutive glucose residues in cellulose are rotated 180° around the glycosidic bond, forming repetitive units of cellobiose (Figure 1.2A). The glucan chains form crystalline microfibrils, held together by interchain hydrogen bonds and Van derWaals forces (Nishiyama *et al.*, 2002), which render them insoluble and therefore enzyme resistant (Dimarogona *et al.*, 2012). Recent research suggests that the microfibrils are made up of 24 glucan chains; groups of three chains are organised in eight sheets, which are arranged into a rectangular shape (Fernandes *et al.*, 2011; Thomas *et al.*, 2013). The previous theory from 1983 suggested subunits of six proteins to make a complex of 36 chains (Herts, 1983). The cellulose microfibrils are produced by large protein complexes called rosettes located in the plasma membrane (Gomez *et al.*, 2008).

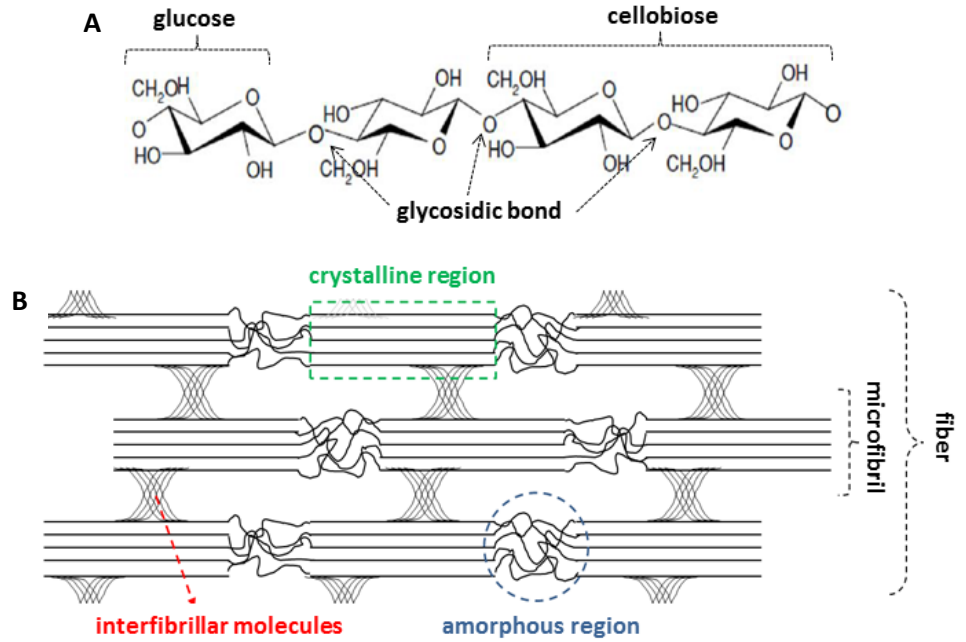


Figure 1.2. The structure of cellulose. **A.** The β-1,4 glucan chain, showing glucose residues rotated 180^{o} around the glycosidic bond and forming repetitive units of cellobiose. Figure adapted from Marriott (2014). **B**. The crystalline and amorphous regions of the cellulose microfibrils. Figure adapted from http://granalgae.eu/granalgae/blog/1-The-organic-graphene.html.

Being abundant in nature, and containing large amount of glucose, cellulose is the ideal biological material to be used for ethanol production, since the glucose can be fermented into ethanol by yeasts or bacteria (Carroll and Somerville, 2009). However, the main hindrance to the process is the crystallinity of cellulose, which makes it inaccessible to the cellulolytic enzymes and therefore more difficult to digest (Gomez *et al.*, 2008). However, not all the cellulose is organised in a crystalline structure, since some areas of the glucan chain are non-covalently linked to other polysaccharides such as mannans and xylans, and are therefore less organised and amorphous (Figure 1.2B) (Heredia *et al.*, 1995).

1.1.4.2 Hemicellulose

Hemicellulose is a collective term for a group of polysaccharides that cover the cellulose microfibrils, to which they associate by hydrogen bonding (McCann *et al.*, 1990; Gomez *et al.*, 2008; Simmons *et al.*, 2016). Their main role is to give the wall integrity by holding the cellulose chains together, but at the same time they ensure flexibility by keeping the microfibrils spatially separated (Scheller and Ulvskov, 2010). Hemicelluloses are a highly heterogeneous assembly of carbohydrates, and the name of the group was given when the organisation and biosynthesis of hemicelluloses were not well understood. According to

the review written by Scheller and Ulvskov in 2010, hemicelluloses should be defined those plant wall carbohydrates that are not cellulose or pectins and are composed by a β-1,4 linked backbone with an equatorial conformation in C1 and C4; the backbone is highly branched with variegated side chains.

Hemicelluloses can be comprised of pentose sugars (β-D-xylose, α-L-arabinose), hexose sugars (β-D-mannose, β-D-glucose, α-D-galactose), uronic acids (α-D-glucuronic, α-D-4-O-methylgalacturonic and α-D-galacturonic) and sometimes sugars like α-L-rhamnose and α-L-fucose (Gírio *et al.*, 2010). The composition and abundance of these sugars and acids varies highly between different species of plant and tissue type (Scheller and Ulvskov, 2010), therefore it is difficult to give an overall description for each group. Xyloglucans are the most abundant components of hemicelluloses in the primary cell wall of dicotyledons, but are less abundant and with a different structure in grasses and conifers and are only found in traces in secondary cell walls (Heredia *et al.*, 1995; Vogel, 2008). Xylans are dominant in secondary walls of angiosperms, including hardwoods, cereals and perennial grasses, which are the most important source of biomass for biofuel production, and they represent around 20-30% of the dry weight of the walls (Dodd and Cann, 2009). Mannans and glucomannans constitute a smaller fraction of the cell wall (Vogel, 2008), but are important as a source of energy and hydration in seeds, and are the major hemicellulose in gymnosperm species such as softwoods (Buckeridge, 2010). Beta 1,3- and 1,4-glucans are not very widely distributed, being mainly restricted to few plant groups including grasses, where they are a minor component of secondary walls (Ochoa-Villarreal *et al.*, 2012).

Hemicelluloses typically comprise between 15-35% of plant biomass, therefore constituting an important source of sugars for biofuel production. However, the heterogeneity of the hemicelluloses and the presence of pentose sugars that are not easily fermented by yeast and bacteria that are normally used for ethanol production, results in hemicellulose being largely unused and therefore discarded in most of the ethanol plants around the world (Gomez *et al.*, 2008; Gírio *et al.*, 2010).

1.1.4.3 Pectins

Pectin is the name given to the third group of the polysaccharides found in the plant cell wall. Their main roles are to contribute to the integrity and rigidity of primary cell walls, and they are also involved in ion transport, water retention and defense mechanisms (Caffall and Mohnen, 2009). Pectins are quite diverse but are characterised by having a high

percentage of galacturonic acid and by being easily extracted using water, acids or chelating agents (Heredia *et al.*, 1995; Scheller and Ulvskov, 2010). They are an abundant component of primary cell walls of monocotyledons (20-35%) but are significantly less important in secondary walls of dicotyledons and grasses (Vogel, 2008), therefore they are not considered important in biofuels production and will not be discussed further in this work.

1.1.4.4 Lignin

Lignin makes up about 30% of most secondary plant cell walls (Scheller and Ulvskov, 2010). It is the main barrier for successful extraction of polysaccharides from plants, as it forms a hydrophobic coating that is highly resistant to degradation (Sattler and Funnell-Harris, 2013). Lignin is deposited to reinforce cell walls after cell expansion has ceased. It is a polyphenol comprised of hydroxycinnamyl alcohol monolignols, *p*-courmaryl alcohol, coniferyl alcohol and sinapyl alcohol, which are secreted into the cell wall and then polymerised by free radical coupling (Figure 1.3) (Boudet *et al.*, 1995; Boerjan *et al.*, 2003). Once the monomers are incorporated into lignin, they constitute units that are categorised as p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S). These H, G and S units are used by different plants in different proportion, being mostly G and S in angiosperms, predominantly only G (with traces of H and S) in gymnosperms and a mixture of G and S in monocotyledons (Boerjan *et al.*, 2003). The resulting polymer is highly branched and lacks notable repeat structures, the reason why so far no enzymes have been discovered that can directly depolymerise it.

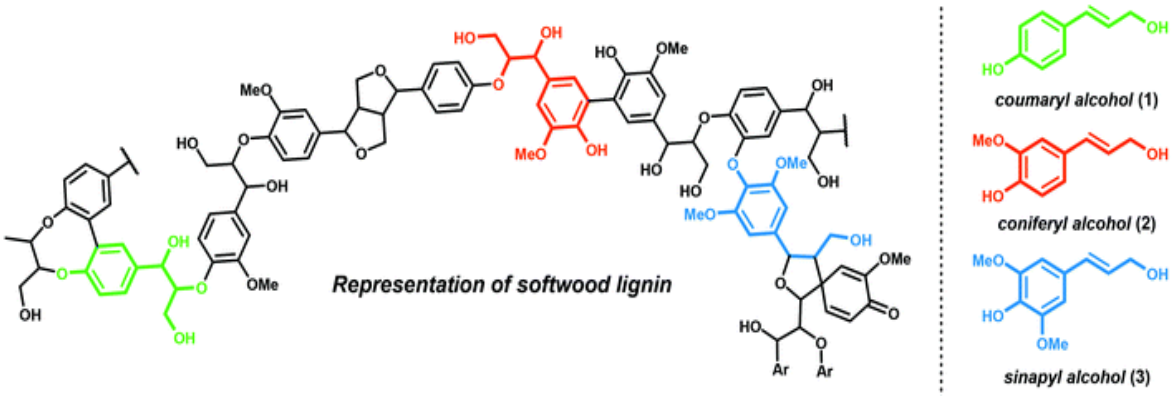


Figure 1.3. The structure of lignin. The three building blocks of lignin, the monolignol monomers, are represented in colours: p-coumaryl (green), coniferyl (red) and sinapyl (green). Figure reproduced from Kärkäs et al. (2016)

1.1.4.5 Structural proteins

Structural proteins are another element of the plant cell wall, but they are mostly found only in the primary cell walls (around 10%) of both monocotyledons and dicotyledons (Vogel, 2008). These proteins are classified in five main classes (extensins, glycine-rich proteins, solanaceous lectins, proline-rich proteins and arabinogalactan proteins) and their functions range from structural roles to morphogenesis and cell-cell interactions (Showalter, 1993; Cassab, 1998). They will not be discussed further as they are not considered important in the context of biofuel production.

1.1.5 Lignocellulolytic enzymes

The enzymes needed for cellulose digestion include four main categories, which are collectively called cellulases and act synergistically by creating new sites for the enzyme that follows. These include: 1. Lytic polysaccharide monooxygenases (LPMOs) oxidatively introduce breaks in glucan chains in crystalline cellulose to allow glycoside hydrolases to function; 2. Endo-glucanases randomly hydrolyse the internal glyosidic bonds in the cellulose microfibrils, creating free chain ends; 3. Exo-glucanases (also called cellobiohydrolases) cleave glyosidic bonds from the reducing and non-reducing ends of the cellulose chain, generating glucose or cellobiose moving along the cellulose chain; 4. Beta-glucosidases hydrolyse cellobiose into glucose and also act on cellulooligosaccharides by cleaving glucose units off (Teeri, 1997; Watanabe, 1998; Eriksson *et al.*, 2002; Väljamäe *et al.*, 2003; Jørgensen *et al.*, 2007).

A wide variety of enzymes, which make up the hemicellulases, is needed for the digestion of the hemicelluloses, given the complexity of these polymers. For example, the complex structure of grass biomass requires a range of endo-β-1,4-xylanases and β-1,4-xylosidases to attack internal and external xylan chains, endo-1,4-β-mannanases and 1,4-β-mannosidases to cleave internal bonds in mannans and release mannose from the non-reducing ends, α-L-arabinofuranosidases, α-L-arabinanases, α-glucuronidases, p-courmaric acid and acetylxylan esterases, feruloyl esterases and other enzymes (Prade, 1996; Beg *et al.*, 2001; Shallom and Shoham, 2003; Moreira and Filho, 2008). LPMOs that act on hemicellulose such as xylan, xyloglucan and glucomannan have also recently been described (Agger *et al.*, 2014; Hemsworth *et al.*, 2015; Couturier *et al.*, 2018).

In order to access the carbohydrates, lignin also must be degraded or at least modified. However, as previously described (section 1.1.4.4), lignin is difficult to attack enzymatically

because it lacks a repetitive structure, and its degradation or modification is achieved using free radical attack mediated by lignin and manganese peroxidases and laccases (Higuchi, 2004). In this type of reactions, C-C and C-O-C bonds from polyphenols and lignin are cleaved oxidatively to generate free radicals (Ertan *et al.*, 2012; Siddiqui *et al.*, 2014). The precise mechanism of lignin depolymerisation and the role and importance of Fenton chemistry over that of enzymes is still a matter of debate; it is probable that the two systems work in synergy (Scharf *et al.*, 2011).

The Carbohydrate-Active enZymes database (CAZy, www.cazy.org) is a comprehensive collection of enzymes active on carbohydrates and their associated modules (Cantarel *et al.*, 2008). The enzymes are classified according to similarities in their amino acid sequence (Henrissat, 1991). The database includes both the enzymes needed for the assembly of carbohydrates, named glycosyltransferases (GTs), and those necessary for their breakdown. The latter are grouped in five different classes: glycosyl hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate binding modules (CBMs) and the recently added auxiliary activities family (AA), which includes lignolytic enzymes and LPMOs (Lombard *et al.*, 2013).

1.2 Degradation of lignocellulose in nature

As discussed above, the industrial deconstruction of lignocellulose is a cost-inefficient process, which involves pretreatments at harsh conditions (high temperature, pressure and pH, use of acids or alkali) in order to disrupt the recalcitrant structure of the cell wall and expose the carbohydrates for depolymerisation into sugars for fermentation (Gomez *et al.*, 2008; Carroll and Somerville, 2009). However, in nature there are organisms that are able to degrade lignocellulose under ambient conditions, having evolved to utilise the plant carbohydrates for their metabolism. These organisms are known under the collective name of Natural Biomass Utilisation Systems (NBUS) and have become models for the development of lignocellulose conversion systems (Xie *et al.*, 2014). They include bacteria, protists, microbial consortia in vertebrate digestive systems, fungi, terrestrial and marine invertebrates (Guerriero *et al.*, 2015). They can degrade hemicellulose and cellulose with different degrees of efficiency, and some can also digest or modify the lignin (Figure 1.4). Different NBUS will be discussed in the following sections, with details on the lignocellulolytic enzymes present in each group.

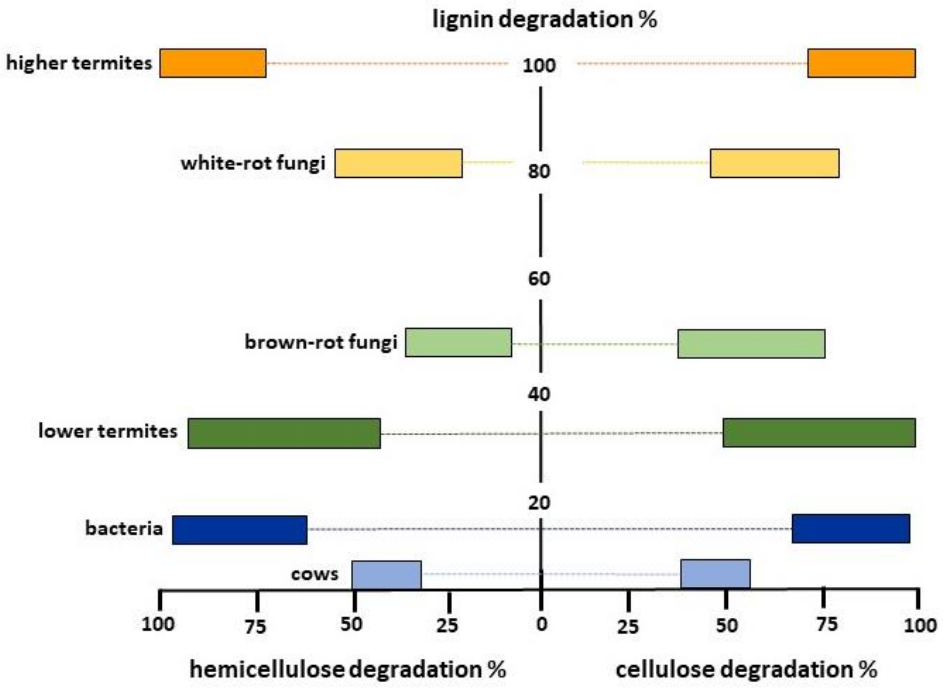


Figure 1.4. Conversion efficiency of lignin, hemicellulose and cellulose by different classes of NBUS. White rot and brown rot refers to the two types of fungi. Figure reproduced from Sun et al (2013).

1.2.1 Prokaryotes and archaea

Bacteria are well known for their ability to digest lignocellulosic material (Peralta-Yahya *et al.*, 2012), since it was once thought that hydrolysis of cellulose was almost exclusively performed by these microorganisms. Bacteria are indeed the main agents, together with fungi, responsible for the recycling of carbon in the soil (Lynd *et al.*, 2002). The aerobic bacteria *Actinomycetales* (phylum Actinobacteria) and the anaerobic *Clostridiales* (phylum Firmicutes) have the most prominent cellulolytic ability (Bélaïch *et al.*, 1997). There is a notable difference between anaerobic and aerobic bacteria in the way they perform biomass digestion. Anaerobic bacteria produce an assembly of enzymes called a cellulosome, which is generally located on the surface of the bacterial cell (Schwarz, 2001). Therefore cell adhesion to cellulose is necessary in these bacteria (Lynd *et al.*, 2002). Although this arrangement enhances the performance of the enzymes (Doi and Kosugi, 2004; Bayer *et al.*, 2008), the conversion of the substrate is relatively inefficient, giving low yields of products (Lynd *et al.*, 2002). Aerobic bacteria, on the contrary, produce a wide variety of individual enzymes, which are secreted and can be recovered in the culture supernatant (Rapp and Beerman, 1991); this mixture has an enhanced efficiency compared to single cellulases and hemicellulases (Wilson, 2011). Prokaryotes able to degrade lignocellulose produce a wide range of cellulases and hemicellulases belonging to

numerous GH families, while bacterial LPMOs have only been identified in the family AA10. Until recently, bacteria were thought to have no ability to degrade lignin, but lignolytic enzymes have now been discovered, which are mainly laccases but also include lignin, manganese and DyP-type peroxidases (McLeod *et al.*, 2006; Jing, 2010; Ausec *et al.*, 2011; Jing and Wang, 2012; Shi *et al.*, 2013b). However, the efficiency is lower than that described for equivalent enzymes found in fungi (Bugg *et al.*, 2011).

The involvement of Archaea in lignocellulose degrading communities has only recently been discovered, but they are now known to have a role in termite guts and in compost (Boucias *et al.*, 2013; de Gannes *et al.*, 2013). The mechanisms by which they deconstruct lignocellulose are not well known so far, but they are thought to rely on GHs and laccases (Kataoka and Ishikawa, 2014; Tian *et al.*, 2014).

1.2.2 Fungi

Fungi have long been known for their ability to degrade lignocellulose. Indeed, army material (e.g. tents, clothing) damaged by fungi during World War II prompted studies on these organisms, and allowed the discovery of the first cellulose and hemicellulose degrading enzymes (Makela *et al.*, 2014). Nowadays, some fungal species such as *Aspergillus niger* and *Trichoderma reesei* are still widely studied as model organisms for lignocellulose digestion and are used for the production of commercial cellulolytic enzymes (Xie *et al.*, 2014). The phylum Basidiomycetes includes the classes of white- and brown-rot fungi, which can degrade all the components of the plant cell wall. However, these two groups of fungi are quite different in the mechanisms and end products that they use in their attack of wood.

White-rot fungi are the only microorganisms known to be able to completely depolymerise and mineralise lignin, which is transformed into water and carbon dioxide, leaving behind a white residue that is mostly composed of cellulose (Hatakka and Hammel, 2011; Xie *et al.*, 2014). The majority (more than 90%) of the wood-rotting basidiomycetes indeed belong to the white-rot group (Gilbertson, 1980). This lignin degrading ability is due to the production of a particular group of extracellular enzymes, the class II peroxidases, which are lignin, manganese and versatile peroxidases, and laccases (Makela *et al.*, 2014; Riley *et al.*, 2014; Pollegioni *et al.*, 2015), which can oxidise lignin. However, a single species does not express all these types of enzymes, but only a subset (Ahmad *et al.*, 2010). The mechanism they use to digest the non-lignin fraction of the wood is mainly through

enzymatic attack, and they indeed produce many cellulases. Some white-rot fungi have also reported to express a wide range of LPMO enzymes (Busk and Lange, 2015). A further class of accessory enzymes, such as cellobiose and glucose dehydrogenases, quinone reductases, ary-alcohol dehydrogenases and oxidase, and glyoxal oxidase is involved in lignin degradation in white-rot fungi, playing an accessory role to the lignin peroxidases by serving as electron donors for the formation of free radicals that attack the lignin (ten Have and Teunissen, 2001; Xie *et al.*, 2014). These enzymes are important because the large size of the primary enzymes prevents them from penetrating the plant cell wall structure (Cragg *et al.*, 2015).

Brown-rot fungi, on the contrary, are only able to modify lignin (producing a brown residue) and cannot degrade it completely, in contrast to white-rot fungi (Floudas *et al.*, 2012). They produce only few carbohydrate active enzymes (CAZymes) and they use Fenton chemistry to cleave internal glyosidic bonds (Jensen *et al.*, 2001; Xie *et al.*, 2014). This radical chemistry system is unique in nature and is very efficient, allowing brown-rot fungi to exploit certain ecological niches such as coniferous forests (Cragg *et al.*, 2015) and man-made wooden structure (Eriksson *et al.*, 1990). Brown-rot fungi attack the hemicellulose fraction first, then the cellulose, and what it left at the end is a polymer containing aromatic rings, which derives from the original lignin (Hatakka and Hammel, 2011).

Soft-rot fungi are not effective wood degraders such as white-rot, but they can colonise environments difficult for other types of fungi, such as hot, cold or wet ecosystems (Nilsson and Daniel, 1989). They can also attack woods containing high levels of compounds difficult to digest, such as tannin and suberin (Vane *et al.*, 2006). Their method of attack is by growing their hyphae in the lumen of the cells and then forming holes in the secondary cell wall (Hatakka and Hammel, 2011). They have a negligible effect on lignin, which remains largely unmodified (the knowledge on the enzymes they use is limited), but they produce a wide range of cellulolytic enzymes which allows them to make use of the wood carbohydrates (Cragg *et al.*, 2015).

1.2.3 Protista

A few unicellular eukaryotic organisms are known to produce enzymes that can degrade cellulose and hemicellulose. The genome of the oomycete *Phytophthora parasitica* was analysed with bio-informatic tools to discover genes encoding for both cellulases and hemicellulases belonging to the GH families 1, 5, 6, 7 and 10 and to the LPMO families AA9

and AA10 (Blackman *et al.*, 2014). These genes allow the pathogen to penetrate into the plant cell wall during the infection process and are not related to lignocellulose digestion for nutritional purposes. Similarly, the dinoflagellate *Alexandrium catenella* produces a cellulase that allows it to break down cellulose, but the role is related to cell division (Toulza *et al.*, 2010), while the photoheterotrophic chlorophyte *Chlamydomonas reinhardtii* uses an endogenous endoglucanase to digest cellulose in the absence of other carbon sources (Blifernez-Klassen *et al.*, 2012). Recently, multiple species of protists in termite guts have been identified as able to phagocytose and digest wood particles. The enzymes responsible for this task are stored into their cell vacuoles and belong to different classes of cellulases and hemicellulases (Brune, 2014).

1.2.4 Terrestrial invertebrates

Many insects are able to utilise plant material to complement their diets, but some can actually live on lignocellulose as their only food (Sun and Scharf, 2010). These xylophagous insects have established a symbiosis with microbes (bacteria, archaea, fungi and protista) that live in their digestive system in order to digest the wood they feed on (Shi *et al.*, 2013a). Members of the orders Dictyoptera, Isoptera, Orthoptera, Coleoptera, Lepidoptera and others are known to be able to degrade lignocellulose, for a total of more than 20 families; these include beetles, roaches, wood wasps, silverfish, leaf-cutting ants and others, but the group that is the most well described and most efficient is that of the termites (Figure 1.5) (Sun and Scharf, 2010). Indeed, termites can metabolise up to 99% of cellulose and 87% of hemicellulose, and are therefore key organisms in the carbon cycle (Xie *et al.*, 2013).

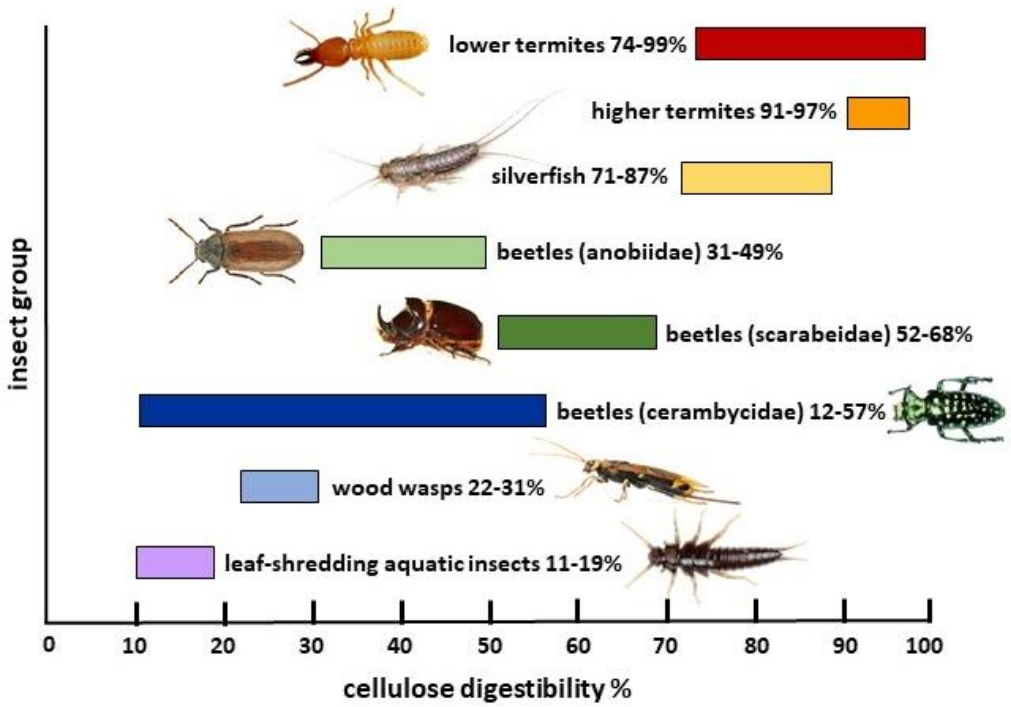


Figure 1.5. Cellulose digestion capability in some insects. The table shows the percentage of cellulose that is digested in different groups of insects. Data were obtained by comparing cellulose contents in food and faeces. Figure adapted from Xie at al. (2013).

The microbiota found in the termite digestive system produces cellulases and hemicellulases belonging to at least 45 different CAZy families, of which GH5 are the most abundant (Warnecke *et al.*, 2007). It was thought that the digestion of lignin depended mainly on enzymatic reactions, but recent studies suggest the involvement of oxidative chemistry (Taprab *et al.*, 2005; Geib *et al.*, 2008). Some species of termites are classed as "higher" termites – to distinguish them from the "lower" ones - because they have established a symbiosis with fungi. These *Termitomyces* fungi enhance the efficiency of lignin degradation and therefore enable the termites to access and metabolise more carbohydrates (Warnecke *et al.*, 2007). It was originally thought that the digestion of lignocellulose in termites was only performed by the symbionts. However, in 1998 Watanabe and colleagues showed that the termites themselves are able to produce some cellulases, mainly belonging to the CAZy families GH9 and GH1 (Watanabe, 1998).

Endogenous cellulases, mainly endo-glucanases and β-glucosidases, are produced also by other terrestrial invertebrates that lack large digestive gut chambers to harbor symbionts, such is the case for some nematodes, gastropods, crustacea and anellidae (Yokoe and Yasumasu, 1964; Nozaki *et al.*, 2009; Milatovič and Štrus, 2010; Bui and Lee, 2015; Cragg *et al.*, 2015).

1.2.5 Marine invertebrates

Marine bivalves are filter feeders, and were once believed to leave uneaten the cellulose fraction of suspended matter. However, it has recently been shown that quite the opposite is true, as bacteria in their digestive system have a very active role in the digestion of biomass and enable these molluscs to utilise the abundant lignocellulose material that is brought in the sea by river estuaries (Sakamoto *et al.*, 2007; King *et al.*, 2010). Further research has highlighted that some bivalves like the blue mussels *Mytilus edulis* (Xu *et al.*, 2001), the abalone *Haliotis discus hannai* (Suzuki *et al.*, 2003), and some gastropods such as *Ampullaria crossean* (Wang *et al.*, 2003) and *Aplisia curodai* (Tsuji *et al.*, 2013) produce endogenous cellulases and, therefore, do not rely entirely on symbiotic bacteria for cellulose degradation.

A group of marine bivalves that is very well known for its wood-digesting ability is that of shipworms, of which *Lyrodus pedicellatus* – subject of this work – is a member. The caecum of shipworms appears to harbour very few microbes (Betcher *et al.*, 2012), while endosymbiotic bacteria able to produce CAZymes have been found in bacteriocytes located in the shipworm's gills (Distel *et al.*, 2002a), although the bacteria reside quite at a distance from the caecum, the main site of lignocellulose digestion. The role that these bacteria play in wood digestion is not clear, and neither is the extent to which the shipworm itself produces enzymes for the degradation of the wood it feeds on. Shipworms are discussed in detail in section 1.3.

Other marine animals are well known for their ability to bore and ingest wood and, together with the shipworms, are colloquially called "wood borers". Evidence of their feeding habits date back to the Jurassic (Vahldiek and Schweigert, 2007) and they are considered both pests, as they destroy man-made wood structure like boats, pier and dykes (Si, 2000; Distel *et al.*, 2011), but also beneficial, because of their highly efficient wood-turnover ability (Distel, 2003). Among the wood borers, the crustaceans belonging to the genus *Limnoria* (order Isopoda) are fascinating, since their digestive tract is completely sterile and are able to feed on wood utilising their own endogenous enzymes, without depending on bacteria (King *et al.*, 2010). These crustaceans probably use free radical chemistry to modify the lignin and render the cellulose accessible to glycosyl hydrolases, which have been identified as belonging to many families, mainly 5, 7 and 9 (King *et al.*, 2010; Kern *et al.*, 2013). Hemocyanins are proteins found in abundance in their

hepatopancreas, suggesting that they are the generators of the free radicals used for the oxidative attack of lignocellulose (King *et al.*, 2010).

1.2.6 Vertebrates

Herbivores are animals adapted to eat plant material, the group including numerous vertebrates. Ruminants are the classic example, since they have evolved to specifically feed on grass. However, they are not themselves able to produce the enzymes needed for lignocellulose digestion, but like other NBUS they rely on a symbiosis with microbes that live in their digestive tract, which includes bacteria, archaea, fungi and protozoa (Xie *et al.*, 2014). Hundreds to thousands of species are present in the rumen at one time, each one of them producing various types of lignocellulolytic enzymes, making the system very difficult to study and reproduce (Weimer, 2013). The main hindrance to studying these enzymes is caused by the fact that only a small percentage of the microbes can be cultivated (Stewart, 2012; Pham and Kim, 2012). Furthermore, even with successful cultivation, the individual enzymes are not always possible to be separated from the mixture or expressed for characterisation (Taupp *et al.*, 2011). These enzymes can be found both on the microbes' cell surface or are secreted extracellularly (Dodd *et al.*, 2010) and are characterised by containing many carbohydrate binding modules (Cantarel *et al.*, 2008). Certain terrestrial herbivorous vertebrates, such as beavers and porcupines, can actually eat wood (not just plant material). However, it is not clear the extent of their xylophagy and whether they have to complement their diet with other types of food (Vispo and Hume, 1995). In the aquatic environment the catfish *Panaque nigrolineatus* is known to ingest wood as part of its herbivorous diet (up to 70% of the gut content) and it has been discovered that the wood is actually digested with the help of a microbial community found in its gastrointestinal tract. This microbiome is composed of various species of bacteria and fungi (Watts *et al.*, 2013; Marden *et al.*, 2017) and it is thought to complement the wood diet by fixing atmospheric nitrogen (McDonald *et al.*, 2012), similarly to what happens in shipworms (see section 1.3.3).

1.3 The wood boring bivalves

The superfamily Pholadacea includes some of the most fascinating and specialised of all bivalve molluscs, those that can live in and eat wood and are therefore colloquially called wood borers. This group of bivalves is divided in two separate but closely related families, the Pholadidae (or piddocks) and the Teredinidae (or shipworms) (Distel, 2003). Members of the Pholadidae family have close resemblance to typical bivalves, but can burrow inside a variety of hard substrates, which include wood but also peat, stone, mud, clay, other molluscs' shells, coral, sand and even plastic (Jenner *et al.*, 2003; Distel *et al.*, 2011; Paalvast and van der Velde, 2013). Grinding into the substrate is performed by specialised shells containing external abrasive spikes (Purchon, 1955), but nutriment is not obtained from the substrate but from organic material suspended in the water (Turner, 1955). The Teredinids (members of the Teredinidae family), on the contrary, are obligate wood borers (with the exception of the genders *Kuphus* and *Zachsia*), which have evolved the remarkable ability to utilise the sugars in the wood for their nutrition (Dore and Miller, 1923). Some species are able to live on a diet of only wood (Becker, 1959; Gallager *et al.*, 1981), though the majority of the species alternate filter feeding to wood boring (Morton, 1970; Paalvast and van der Velde, 2013). They only resemble classical bivalves while in the larval stage, but once they find suitable wood they grow into adults by undergoing a major metamorphosis, which sees the body elongating while the shells not growing at the same rate, resulting in an animal with a vermiform shape, hence the name of "shipworm" (Turner, 1966). Adult specimens can grow from few millimetres to almost two meters in lengths, and several centimetres in diameter, depending on the species (Distel *et al.*, 2002a).

The rest of this chapter will focus on the Teredinidae family only, since it is the one that includes the shipworm *Lyrodus pedicellatus*, species under study for this PhD thesis.

1.3.1 Distribution, ecological role and economic impact of the Teredinidae

Shipworms can be found worldwide, from tropical to temperate areas, excluding only the Polar Regions (Turner, 1966; Saraswathy and Nair, 1971; Borges *et al.*, 2012). They are distributed across the seas, but they are mainly located around the coastline, rivers mouths and mangrove forests (Bienhold *et al.*, 2013). They mainly inhabit shallow waters, but can live up to 200-700 meters deep (Bartsch, 1927; Roch, 1940; Distel *et al.*, 2011); some

species have been retrieved alive from 7,488 meters of depth, but colonisation had probably occurred prior to the wood sinking (Turner, 1966). Wood boring bivalves living at depths of up to 5,000 meters belong to the Pholadidae family only, which represents the deep-sea version of the Teredinids (Voight, 2007). Shipworms can colonise floating, fixed, submerged and living wood like that of mangrove trees (Distel *et al.*, 2011), and the environmental factors that control their distribution are mainly temperature, salinity, oxygen concentration and wood presence (Turner, 1966; Borges *et al.*, 2014). However, many species can tolerate a broad fluctuation of these parameters, and a few species are adapted to brackish waters and even freshwater, like *Nausitoria hedlei* and *Psiloteredo healdi* (Turner, 1966).

The Teredinidae are important organisms in both marine and estuarine ecosystems, because of their role in the carbon cycle, which involves unlocking the organic carbon contained in wood and redistributing it in the environment (Lopes *et al.*, 2000). Indeed, it has been established that in mangrove systems they account for 70% of the wood turnover (Distel, 2003) and that they can cause the loss of 90% of the weight from mangrove trunks, triggering their eventual collapse (Nair and Saraswathy, 1971; Cragg, 2008). Their ability to digest wood is linked to their evolution, the first shipworm fossil ever recorded belonging to the lower Cretaceous (Hatai, 1951), which corresponds to the evolution of woody plants and in particular of salt-tolerant ones. Wood was initially exploited for protection, and it likely became a source of nutrition only in second instance (Turner, 1966; Turner, 1971a).

Shipworms are not selective when it comes to the timber to attack, indeed they colonise man-made wooden structures such as pier, boats, navigation poles, dock, buoys, floats and even coconuts shells (Turner, 1966; Nair and Saraswathy, 1971; Distel *et al.*, 2011), hence the name of "termites of the sea" (Turner, 1966). The first accounts of shipworm damage date back to Pliny, Ovid, Aristophanes and Homer, and the Dutch naturalist Sellius in the early 18th Century wrote some of earliest scholarly writings on shipworms, which is not surprising since the Netherlands coast at that time was shielded by a large number of wooden dykes, which would have been riddled by shipworms (Nair and Saraswathy, 1971; Distel, 2003). Many famous exploration voyages were plagued by the presence of shipworms in the wooden ships, as reported in accounts by Cook, Drake and Columbus (Turner, 1959). The latter reported infestation so extensive during its fourth trip to the Americas in 1502, that the crew was eventually left stranded in Jamaica (Scofield, 1975).

There is no estimation of the economic damage caused by shipworms in European waters; however, they are considered to be the cause of billions of dollars damage worldwide each year (Distel *et al.*, 2011). The Ballast Water Management Convention in 2004 estimated that Teredinids destruction accounts for millions of Euros loss each year in Germany alone, for example. In the Venice lagoon (Northern Italy), wood is still utilised for boats and many maritime structures, including more than 22,000 navigation poles called *briccole*, thousands of which have to be substituted each year due to damage cause by shipworms (and other crustacean wood borers), generating great economical loss (Schröder, 2016). Despite the use in present days of metal and fibreglass instead of wood in most boats (Cobb, 2002) and of anti-fouling chemical treatments (Distel, 2003), the problem created by Teredinids' damage is still extensive, particularly in developing countries and also considering that some chemicals used for wood protections are highly polluting and therefore have been regulated in Europe, North America and Asia (Distel, 2003; Borges *et al.*, 2008).

1.3.2 Xylotrophy versus filter feeding

The ability of the Teredinids to use wood as a food source (xylotrophy) has been speculated upon since 1733, when Sellius first described the species *Teredo navalis* and wondered whether its wood boring skills were coupled with the ability to digest lignocellulose. For long time the main opinion remained that the wood passed out of the body without being digested, while only plankton was used for nutrition, though some authors at the beginning of the 20th Century started suspecting that the wood might be a complement to the filter feeding diet (Sigerfoos, 1908; Moll, 1914; Calman, 1919). The general consensus started changing in 1921, when Harington highlighted the cellulose degrading properties of the digestive glands of *Teredo norvegica*. Soon after Potts (1923) observed wood fragments in the digestive glands of *T. navalis* and also debated their role in the shipworm nutrition. Dore and Miller (1923) studied the composition of wood before and after ingestion by *T. navalis* and noted that most of the cellulose was consumed by the shipworm, together with part of the hemicellulose, concluding that those carbohydrates are probably utilised as food. Since then, numerous authors have studied the ability of the Teredinids' digestive tract to digest various polysaccharides (Nair and Saraswathy, 1971), and the fact that shipworms can digest wood is now a well-known and accepted phenomenon.

Shipworms are known to alternate wood boring and ingestion with filter feeding, which is the ability to extract phytoplankton from the water (Morton, 1970; Paalvast and van der Velde, 2013). Filter feeding is a characteristic of the majority of bivalves, the class of molluscs the shipworms belong to. In the Teredinids the water is pumped into the body cavity by the inhalant siphon, and is drawn over the gills by a current created by the ciliated epithelium. While passing over the ctenidia (synonym of gills), the suspended food is captured by a layer of mucus, and is then transported to the mouth and from there to the rest of the digestive system. The water then leaves the body using the exhalant siphon, taking with it the digested food expelled from the anus.

Wood boring and filter feeding are mutually exclusive: the shipworms cannot bore into wood when they are eating suspended matter, and vice-versa (Morton, 1970). Therefore, when both food sources are available, they alternate the two processes, retracting the siphons and sealing their burrow with the pallets when feeding on wood (Shipway, 2013). According to some authors, shipworms can survive by eating phytoplankton only, once the wood they have been feeding on is completely exhausted (Bartsch, 1922; Johnson *et al.*, 1936). The length of the ctenidia has been related to efficiency of filter feeding and the extent to which they are dependent on this type of nutrition, with species having the largest gills being described as completely reliant on phytoplankton. However, both these concepts have been challenged by many authors (Potts, 1923; Roch, 1932; Lane, 1955; Becker, 1959), and at present other combined anatomical characteristics are used to demonstrate adaptation toward xylotrophy or filter feeding, such as caecum size, intestine length and digestive gland specialisation (Shipway, 2013).

While it has been known that shipworms digest the lignocellulose and use its sugars for nutrition, the debate is still open on whether they can live on a diet of wood only for prolonged periods when suspended food is not available. It was originally thought that this was not possible, since the wood is nitrogen-poor (Hungate, 1940; Cowling and Merrill, 1966) and lacks in certain amino acids such as phenylalanine, tyrosine, proline and valine (Nair and Saraswathy, 1971), and their only source could be found in protein-rich phytoplankton. Indeed, species such *T. furcifera* (Karande *et al.*, 1968) have never been successfully reared on a diet of just wood, and the same was initially true for *T. navalis* (Becker, 1959). However, some authors reported on the ability of some shipworms to live on wood only, as did Becker (1959), who was able to nurture *L. pedicellatus* for four generations in artificial sea water without any phytoplankton supplement. More recently,

Gallager and colleagues (1981) grew specimens of the same species for four months on either phytoplankton-free or phytoplankton-enriched water, noticing no significant difference in the animals' growth and metabolism. A similar result was obtained with *T. navalis*, which was grown on fir panels for 232 days in filtered water in the presence or absence of the planktonic alga *Isochrysis galbana* (Mann and Gallager, 1985). Gallager et al. (1981) concluded that *L. pedicellatus* must have a mechanism that allows it to conserve nitrogen, as very little appeared to be lost during the experiment. They also stated that the wood itself is not the likely the only origin of the shipworm nitrogen, as its levels are too low to justify the amount found in the animals; hence they must have the ability to complement their diet with other nutritional sources. This was also recently confirmed by stable isotopes analysis of three species of shipworms (including *L. pedicellatus*) in the Venice lagoon, which highlighted that while the animals rely on both xylotrophy and filter feeding, the latter is most likely the main source of nitrogen, with the endosymbiotic bacteria (see next section) having a minor input (Schröder, 2016). This is in contrast with a previous finding obtained by the stable isotopes method, which observed that the nitrogen source for shipworms found in sunken wood in Japan is not likely to be neither the wood nor the suspended particles, while they suggest that the endosymbiotic bacteria provide the major nitrogen supply by nitrogen fixation (Nishimoto *et al.*, 2009). This though can be explained by the different habitat where the studied shipworms came from, as deep-sea wood borers might have different feeding behaviours to those of the intertidal zone. More recently a study of *T. navalis* showed that this species relies more on filter feeding rather than wood consumption (Paalvast and van der Velde, 2013) and work performed on *B. carinata* showed that most of the carbon is derived from wood carbohydrates, while up to 82% of nitrogen from N_2 fixation (Charles *et al.*, 2018). The debate is, therefore, still open, and it might be possible that the overall feeding strategy is not uniform throughout the different shipworm species and the different habitats they live in, with species such as *L. pedicellatus* having a preference for wood and other for phytoplankton, as is the case of *T. navalis*.

Despite being poor in nitrogen and amino acids, wood is a very good food source as it is extremely rich in carbohydrates. The shipworm's ability to digest the wood's carbohydrates, coupled with the ability to convert its sugars into fat and to store high quantities of glycogen (Potts, 1923), allows them to use this rich resource to sustain the

energy-sapping wood boring activity and the extreme rapid growth, a feature of the Teredinids (Dore and Miller, 1923).

1.3.3 The symbiosis between shipworms and bacteria

In 1848 Gérard Paul Deshayes, a French geologist and conchologist, noted the presence in the Teredinids gills of some glandular-like structures, which he hypothesised to be involved in the secretion of substances used for the nutrition of the embryos developing on the gills. These structures (Figure 1.6) were later called "Glands of Deshayes" by Sigerfoos (1908), in honour of the first observer, but he could not produce an educated guess on their function.

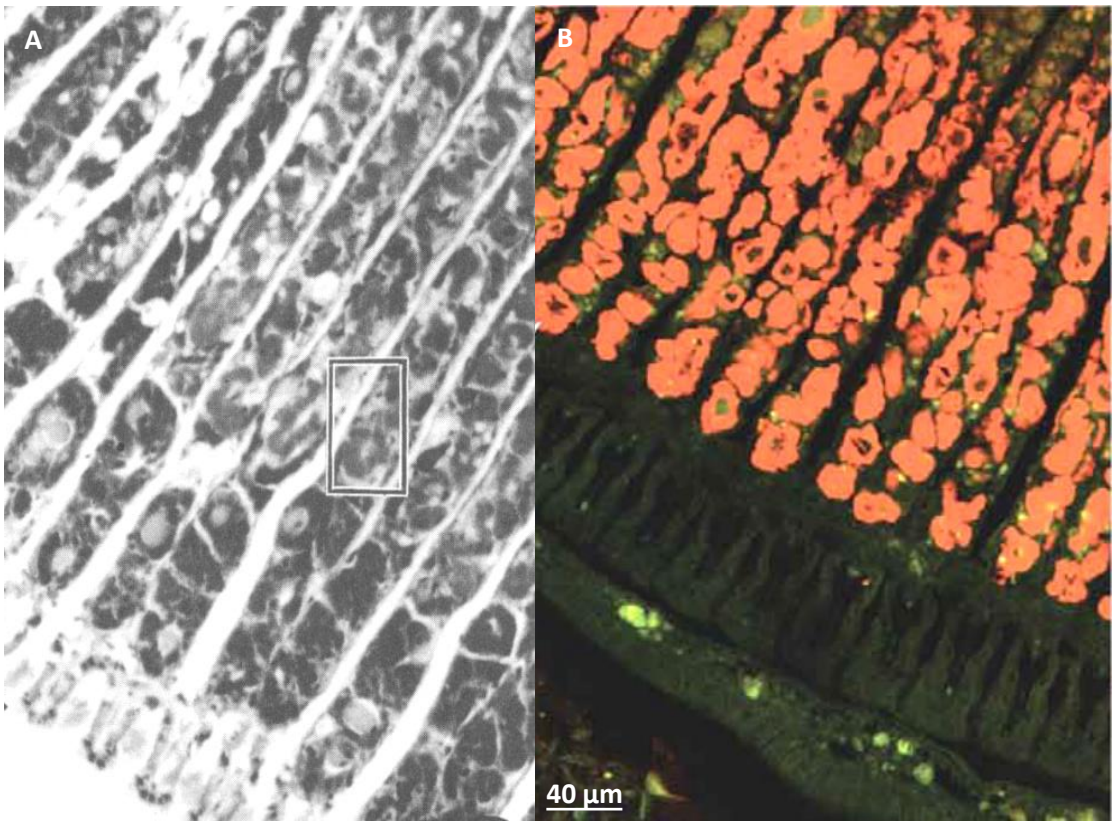


Figure 1.6. Location of the bacteriocytes found within the Gland of Deshayes in the gill's lamellae of *L. pedicellatus*. **A.** Light microscopy (Distel *et al.*, 1991) and **B.** confocal microscopy (Betcher *et al.*, 2012) pictures of gills sections showing multiple gill filaments containing bacteriocytes. The bacteriocytes are in dark coloration in A, while in B they appear in bright orange due to fluorescence obtained by hybridisation with bacteria 16S rRNA directed oligonucleotide probes.

We had to wait until 1973 for the discovery in *Bankia australis* that the alleged gland was instead a region of the gills containing vast numbers of gram-negative bacteria, which the authors showed by electron microscopy and appeared to live in symbiosis with the shipworms (Figure 1.7) (Popham and Dickson, 1973).

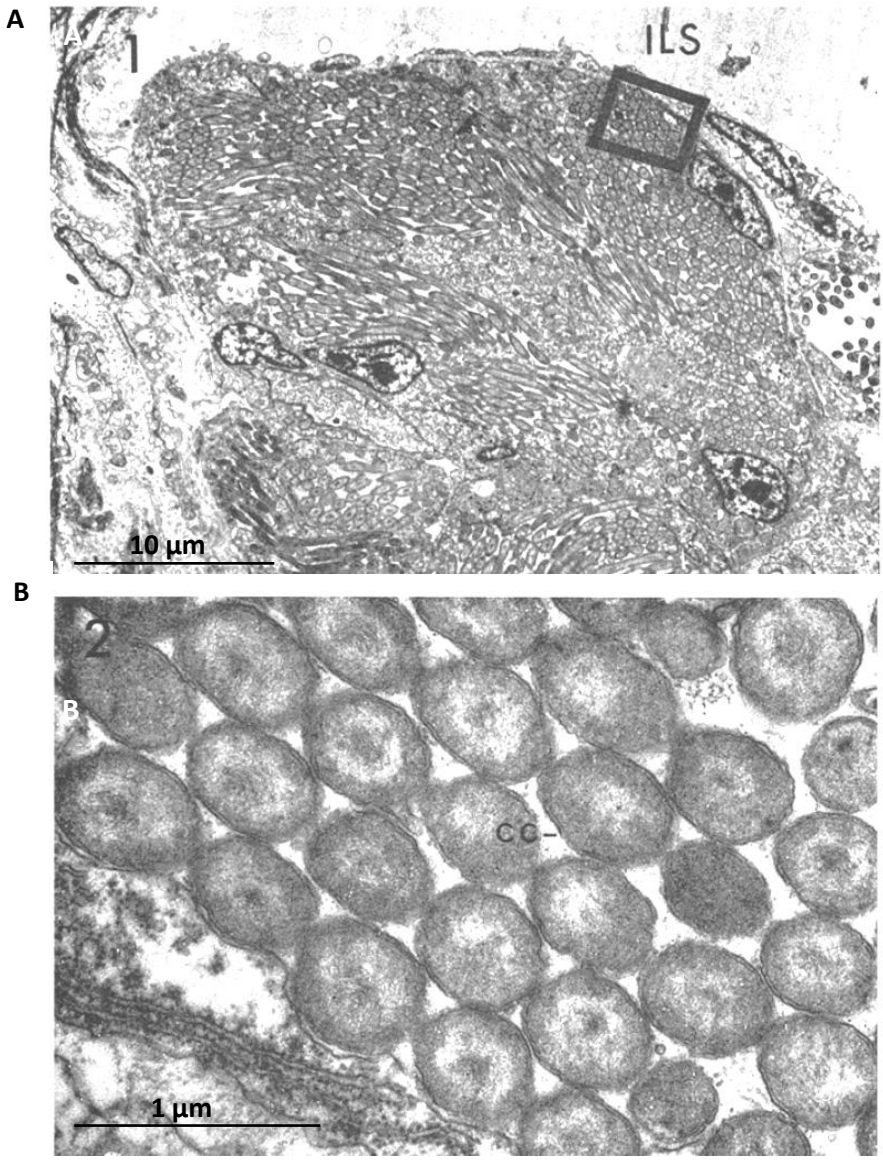


Figure 1.7. Bacteriocytes containing bacteria in *Bankia australis*. **A**. Electron microscopy image of a gill lamina section, showing one bacteriocyte (2,250x). **B.** Higher magnification of the square from picture A, showing tightly packed bacteria (33,000x). Picture taken from Popham and Dickson (1973).

Soon after this discovery, Carpenter and Culliney (1975) were able to demonstrate that nitrogen fixation (the conversion of atmospheric nitrogen into ammonia) is associated with four different shipworm species, including *L. pedicellatus,* and that the rate of fixation is inversely correlated to filter feeding capabilities, therefore it is higher when less nitrogen is available. The nitrogen is then transported into the shipworms tissues and used for their metabolism, to complement the nitrogen-poor wood diet (Waterbury *et al.*, 1983; Lechene *et al.*, 2007). Nitrogen fixation was thought to be associated with bacteria found in the caecum (Carpenter and Culliney, 1975), though we now know that the caecum is almost devoid of bacteria and that those that are able to fix nitrogen are instead found in the gills (Betcher *et al.*, 2012). In 1983, the link between the bacteria found in the gills and the ability

to fix nitrogen was finally demonstrated by isolating in pure culture bacteria extracted from the gills (Waterbury *et al.*, 1983). The bacteria were also shown to be able to digest cellulose and other carbohydrates and they were thought to allow the shipworm to utilise wood for its diet. They were described as proteobacteria, gram-negative, aerobic, chemoheterotrophic, obligate marine symbionts and they were found in five species of shipworms, suggesting that the relationship is widely spread among Teredinids. The bacteria are hosted in specialised shipworm cells called bacteriocytes (Distel, 2003), which can contain hundreds of symbionts, an arrangement that is found in other symbiosis (Baumann *et al.*, 2006).

The symbionts were thought to belong to a single, new species of bacterium, distinct from other known bacterial genera, a result that was confirmed by in situ hybridisation and qPCR in *L. pedicellatus* (Distel *et al.*, 1991). These bacteria, called *Teredinibacter turnerae* (Figure 1.8), were isolated and described in 24 different shipworm species in 2002 (Distel *et al.*, 2002b) and their activity was shown on different carbohydrate substrates. Further studies revealed that there is actually not only one species of bacteria in the shipworm's gills, but that endosymbionts of many species cohabit the same host, all closely related but distinct gamma proteobacteria (Distel *et al.*, 2002a), giving the picture of a symbiosis more complex than previously thought. In recent years it has been shown that this symbiosis is indeed more multifaceted, since various papers have presented evidence that these endosymbiotic bacteria do not just provide the shipworm with nitrogen and lignocellulolytic enzymes, but they are also involved in the production of antibiotics (Elshahawi *et al.*, 2013; Han *et al.*, 2013) and of secondary metabolites that are beneficial to the host, such as polyketides, peptides, lipopeptides and other bioactive compounds (Haygood *et al.*, 2015; Brito *et al.*, 2018).

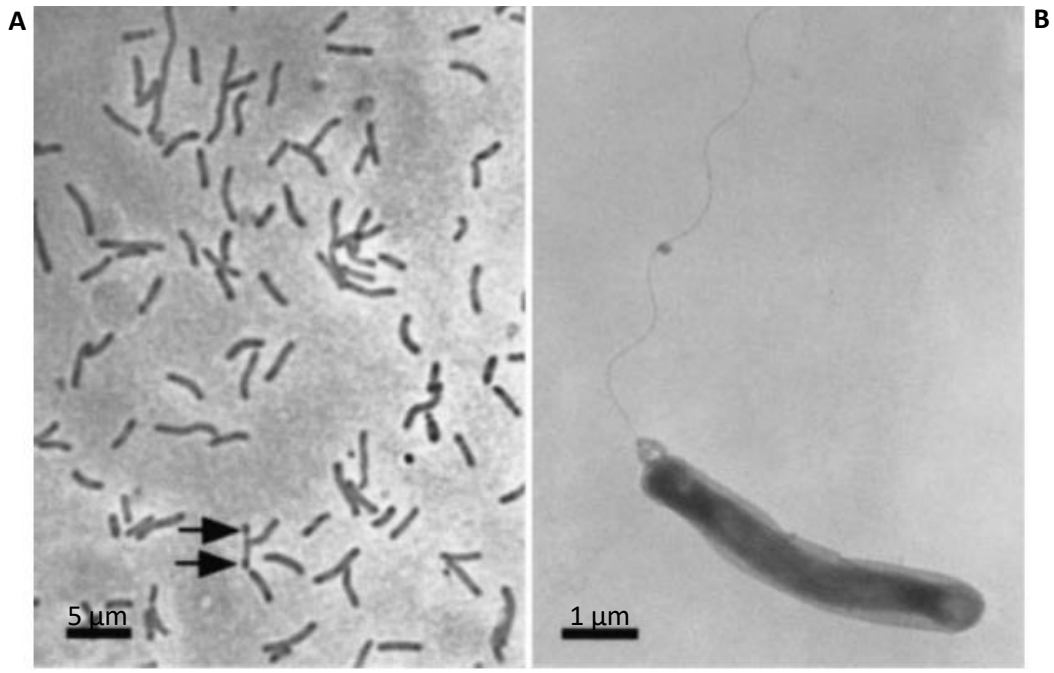


Figure 1.8. The bacterium *Teredinibacter turnerae*, strain T7902T, isolated from the gills of *L. pedicellatus*. **A.** Phase-contrast light picture of the bacteria. **B.** Negatively stained transmission electron microscopy picture of one bacterium, showing the single polar flagellum. Picture taken from (Distel *et al.*, 2002b).

1.3.4 The transport of the bacterial enzymes

The symbiosis that allows complex animals to make use of abundant carbohydrate sources such as wood or grass through the association with microorganisms is widespread in nature and found in both vertebrates and invertebrates, as outlined in section 1.2. However, the particular symbiosis between shipworms and bacteria is unique, since the symbionts are not found within the digestive tract, but in an organ that is part of the respiratory system, the gills (Waterbury *et al.*, 1983). The endosymbiotic bacteria found in the gills produce a range of lignocellulolytic enzymes that are translocated to the caecum, where they are used to digest the wood that has been pulverised by the grinding action of the shells. Only 11% of the bacterial gills proteome is made up of lignocellulolytic enzymes, while they represent almost the entirety of the caecum microbial proteome (98%). This indicates that the wood-degrading enzymes are selected from the other bacterial proteins before entering the caecum (O'Connor *et al.*, 2014). This mechanism is thought to have evolved to keep the bacteria, which can feed on sugars, physically separated from the shipworm's site of digestion and absorption, to keep competition between host and symbionts to a minimum (O'Connor *et al.*, 2014). Indeed, it was demonstrated that the caecum is almost devoid of bacteria, which therefore are not feeding on the same food source as the shipworm (Betcher *et al.*, 2012).

The mechanism behind this selective transport of bacterial CAZymes from the gills to the caecum, as well as the physical path that the enzymes follow, is at present unknown. The bacterial CAZymes that are found in the caecum present a signal peptide for secretion outside the bacterial cells (O'Connor *et al.*, 2014), which explains how they are excreted from the bacterial cell. However, it is unclear how these enzymes are secreted from the bacteriocytes, separated from other bacterial proteins and transported outside the gills tissue up to the caecum, an organ that is not physically adjacent. Some authors suggested that the transfer happens *via* a duct that connects the gills to the oesophagus, from which they could then easily reach the caecum. However, this duct, named Duct of Deshayes, was only once described in 1971 in the shipworm species *Teredo furcifera*, and only depicted by hand drawing (Saraswathy and Nair, 1971). Further work aimed at identifying the Duct of Deshayes by x-ray micro-computed tomography failed to show its existence in *L. pedicellatus* (Shipway, 2013).

1.3.5 The shipworm *Lyrodus pedicellatus*

1.3.5.1 Classification and distribution

Shipworms belong to the phylum Mollusca, order Bivalvia, Superfamily Pholadacea and Family Teredinidae. They are divided into three subfamilies, six groups, 15 genera and 69 species (as detailed in Table 1.1), whose systematic was reviewed and updated by Ruth Turner in 1966 in the comprehensive catalogue “A Survey and Illustrated Catalogue of the Teredinidae". However, shipworms are not simple to identify from morphological features, since shell morphology (which is used for identification for other molluscs and is used for shipworms instead of the pallets by non-specialist) varies greatly among single species and is similar among different ones (Borges *et al.*, 2012; Voight, 2015). Furthermore the pallets are not always retrievable and can also be affected by different types of wood, environmental conditions and erosion (Cragg *et al.*, 2009). Molecular tools (e.g. DNA barcoding) are now available alongside morphology-based identifications to help taxonomists in species identification, but this generates some confusion on the number of species that are recognised by different authors. Indeed, Treneman and colleagues (2018) recently reported the number of species being more than 70 and a new species, *Lyrodus mersinensis,* has been described in July 2018 on the base of molecular diagnostic characters only (Borges and Merckelbach, 2018).

Table 1.1. Classification of the Teredinidae family, based on morphological features (pallets and shells), anatomical characteristics and breeding strategies. The table is taken from Jones at al. (1976), based on Turner (1966) and two more species have been added from Calloway and Turner (1988) and from Turner and Yakolev (1983). The species Zachsia *zenkewitschi* belongs to a genus that was not described by Turner in 1966, therefore it is not listed under any specific subfamily nor into a numbered group.

Subfamily	Group	Genus	Species
Kuphinae	1	*Kuphus*	*Kuphus polythalamia*
Teredininae	2	*Bactronophorus* *Neoteredo* *Dicyathifer* *Teredothyra*	*Bactronophorus thoracites* (Linnaeus) 1767 *Neoteredo reynei* (Bartsch) 1920 *Dicyathifer manni* (Wright) 1866 *Teredothyra dominicensis* (Bartsch) 1921 *Teredothyra excavate* (Jeffreys) 1860 *Teredothyra smithi* (Bartsch) 1927 *Teredothyra matocotana* (Bartsch) 1927
	3	*Teredora* *Uperotus* *Psiloteredo*	*Teredora malleolus* (Turton) 1822 *Teredora princesae* (Sivickis) 1928 *Uperotus clavus* (Gmelin) 1791 *Uperotus rehderi* (Nair) 1956 *Uperotus panamensis* (Bartsch) 1927 *Uperotus lieberkindi* (Roch) 1931 *Psiloteredo megotara* (Hanley) 1848 *Psiloteredo healdi* (Bartsch) 1931 *Psiloteredo senegalensis* (Blainville) 1828
	4	*Teredo* *Lyrodus*	*Teredo aegypos* (Moll) 1941 *Teredo bartschi* (Clapp) 1923 *Teredo clappi* (Bartsch) 1923 *Teredo fulleri* (Clapp) 1924 *Teredo furcifera* (von Martens) 1894 *Teredo johnsoni* (Clapp) 1924 *Teredo mindanensis* (Bartsch) 1923 *Teredo navalis* (Linnaeus) 1758 *Teredo poculifer* (Iredale) 1936 *Teredo portoricensis* (Clapp) 1924 *Teredo somersi* (Clapp) 1924 *Teredo triangularis* (Edmonson) 1942 ***Lyrodus pedicellatus* (Quatrefages) 1849** *Lyrodus affinis* (Deshayes) 1863 *Lyrodus floridans* (Bartshc) 1922 *Lyrodus takonoshimensis* (Roch) 1929 *Lyrodus medilobatus* (Edmonson) 1942 *Lyrodus bipartitus* (Jeffreys) 1860 *Lyrodus massa* (Lamy) 1923
Bankiinae	5	*Nototeredo*	*Nototeredo norvagica* (Spengler) 1792 *Nototeredo knoxi* (Bartsch) 1917 *Nototeredo edax* (Hedley) 1895
	6	*Spathoteredo* *Nausitora* *Bankia*	*Spathoteredo spatha* (Jeffreys) 1860 *Spathoteredo obtusa* (Sivickis) 1928 *Nausitora dunlopei* (Wright) 1864 *Nausitora hedleyi* (Schepman) 1919 *Nausitora dryan* (Dall) 1909 *Nausitora fusticula* (Jeffreys) 1860 *Bankia anechoensis* (Roch) 1929 *Bankia australis* (Calman) 1920 *Bankia bagidaensis* (Roch) 1929 *Bankia barthelowi* (Bartsch) 1927 *Bankia bipalmulata* (Lamarck) 1801 *Bankia bipennata* (Turton) 1819 *Bankia brevis* (Deshayes) 1863 *Bankia campanellata* (Moll and Roch) 1931 *Bankia carinata* (Gray) 1827 *Bankia cieba* (Clench and Turner) 1946 *Bankia destructa* (Clench and Turner) 1946 *Bankia fimbriatula* (Moll and Roch) 1931 *Bankia fosteri* (Clench and Turner) 1946 *Bankia gouldi* (Bartsch) 1908 *Bankia gracilis* (Moll) 1935 *Bankia johnsoni* (Bartsch) 1927 *Bankia martensi* (Stempell) 1899 *Bankia nordi* (Moll) 1935 *Bankia orcutti* (Bartsch) 1923 *Bankia philippinensis* (Bartsch) 1927 *Bankia rochi* (Moll) 1931 *Bankia setacea* (Tryon) 1863 *Bankia zeteki* (Bartsch) 1921
?	?	*Zachsia*	*Zachsia zenkewitschi* (Bulatoff and Rjabtschikoff) 1983

Lyrodus pedicellatus (Quatrefages, 1849), also known as the blacktip shipworm, belongs to the family Teredinidae, subfamily Teredininae, genus *Lyrodus*. It is found worldwide in tropical and warm-temperate waters. It was first described in Spain, but it is probably of Indo-Pacific origin, and it now colonises (after being introduced) areas such as South Africa, New Zealand, North America and Hawaii (Turner, 1966; Turner, 1971a). Recent genetic analysis suggests that the populations found in the Mediterranean Sea and in the Northeast Atlantic could actually represent two separate cryptic species (Borges *et al.*, 2012; Borges and Merckelbach, 2018). In England, this species was introduced via ships' hulls and thermal discharges (Coughlan, 1977), but it is now established in the English Channel, probably as a result of increased sea temperatures (Borges, 2007).

1.3.5.2 General overview and ecology

The overall body structure of *L. pedicellatus* is similar to that of other shipworms, with a long, naked, slender body presenting reduced shells protecting the head only (Figure 1.9). The body size can vary from a few centimetres to a maximum of 30 cm (Poutiers, 1998). The shells are covered on the dorsolateral anterior slope with abrasive denticles used to grind the wood, where this species excavates its burrow, which it lines with a calcareous tube for further protection. This tube is secreted by the mantle, a semi-transparent organ that covers the whole body. The posterior part of the body ends with two siphons utilised to pump water in and out of the body cavity, and two calcareous structures called pallets used to seal the burrow when the wood is not submerged in water and the siphons are retracted. This sealing technique allows the shipworms to survive for long times (up to a week) in adverse conditions. During these periods the gas exchanges are performed by diffusion through the mantle (Manwell, 1963) and nutriment is obtained from glycogen stores (Potts, 1923).

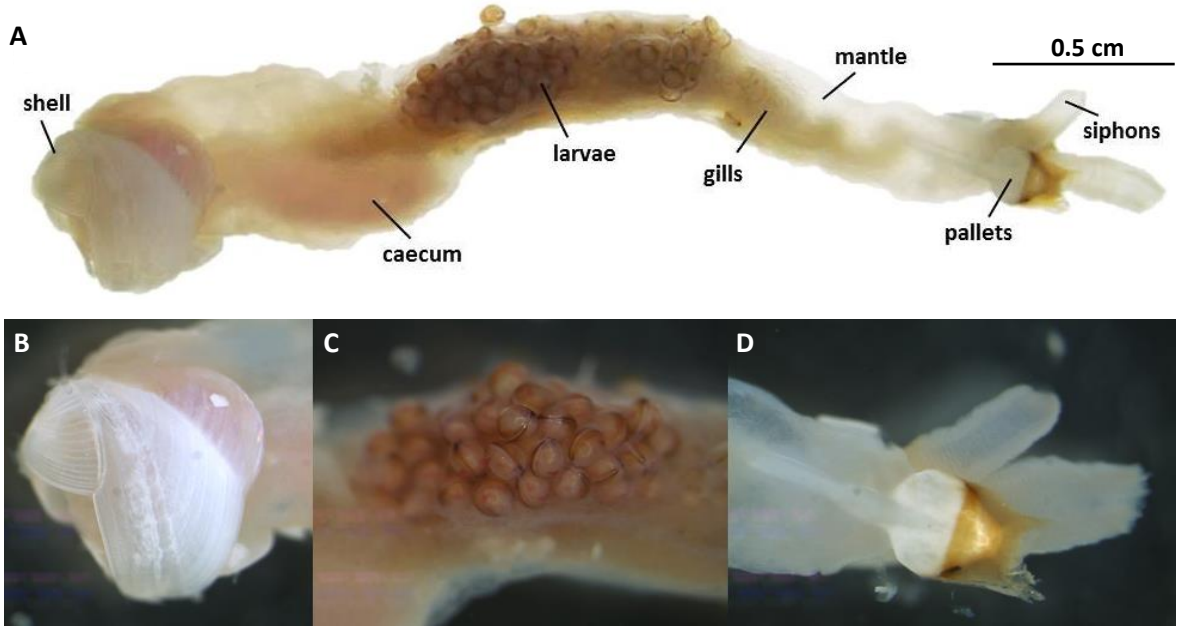


Figure 1.9. Overall body organisation of *Lyrodus pedicellatus*. **A.** Picture of the whole body of an adult specimen bearing larvae in the gills. The main organs visible before dissections are indicated. **B.** The rasping shells. **C.** The larvae contained in the gill's pouches. **D**. The inhalant (bottom) and exhalant (top) siphons and one of the pallets.

Colonisation of wood by this and other Teredinids species is not visible from the outside, and the presence of the animals can only be detected by the siphons protruding from the burrow's opening when the wood is submerged, or by x-ray photographs (Figure 1.10). The larvae, as in other shipworms, undergo a major metamorphosis when they find a piece of wood to colonise. They attack the wood with their shells and, when the body starts to grow and elongate, the head remains immersed in the wood, extending the length of the tunnel, with only the siphons remaining near the burrow's entrance.

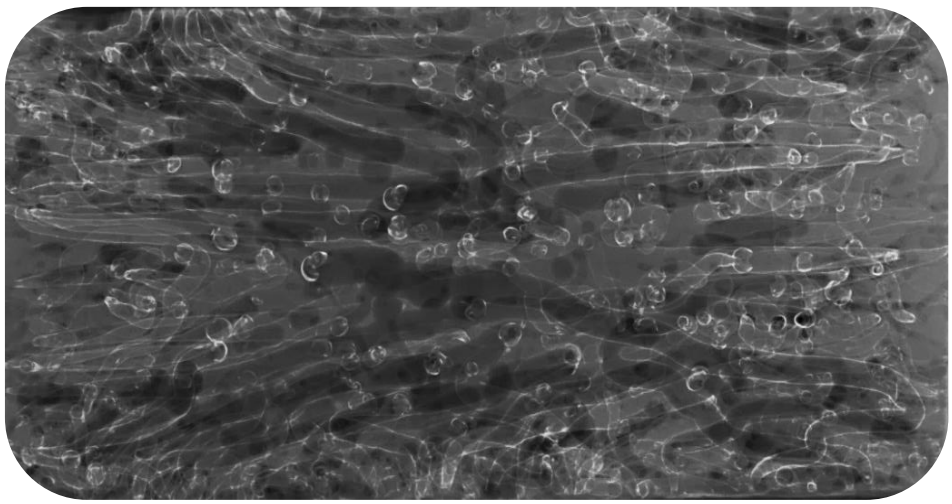


Figure 1.10. X-ray picture of a block of wood riddled with burrows containing *L. pedicellatus*. Picture curtesy of Elisabeth Clutton and Kirsten Farrell-Savage, Portsmouth University.

1.3.5.3 Reproduction

L. pedicellatus is a protandrous hermaphrodite (the male reproductive organs come to maturity before the females ones in the same individual) but is not able to self-fertilise (Eckelbarger and Reish, 1972). The sperm is injected by the males with the exhalant siphon into the females inhalant siphon to fertilise the eggs (internal fertilisation), which are then brooded in the gills in special pouches (Turner, 1966). In *L. pedicellatus*, a series of fertilisation events occur, therefore brooding of the larvae is asynchronous, meaning that there are different stages of development on the gills of the same individual (Cragg *et al.*, 2009). Once they have reached the pediveliger state (they are long-term brooders), the larvae (Figure 1.11) are released from the siphons and spend a maximum of 36 hours as plankton before finding a piece of wood, reason why dispersal can happen only by ships or driftwood (Lebour, 1946; Turner, 1971a; Calloway, 1982).

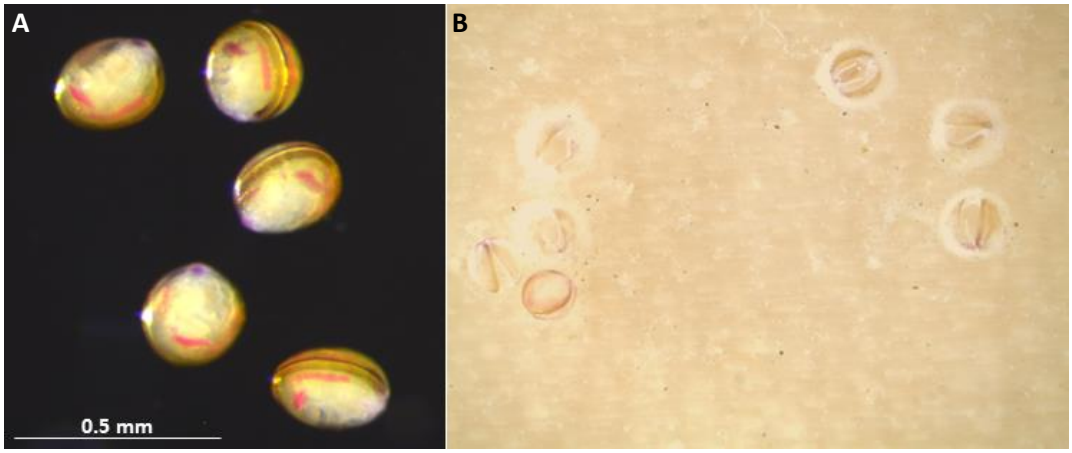


Figure 1.11. *Lyrodus pedicellatus* larvae. **A.** Larvae free in the water column. **B.** Larvae that have just colonised a log of wood.

Within two to three days, the larvae start to bore into the wood and their metamorphosis begins (Figure 1.12). Once the boring starts, the shells become more calcified and develop the denticles used for grinding, while the foot turns into a short organ only used for attachment to the wood. Sexual maturity is already reached after 10 weeks (Turner, 1971a), allowing this species to colonies new areas very rapidly, given the favourable environmental conditions.

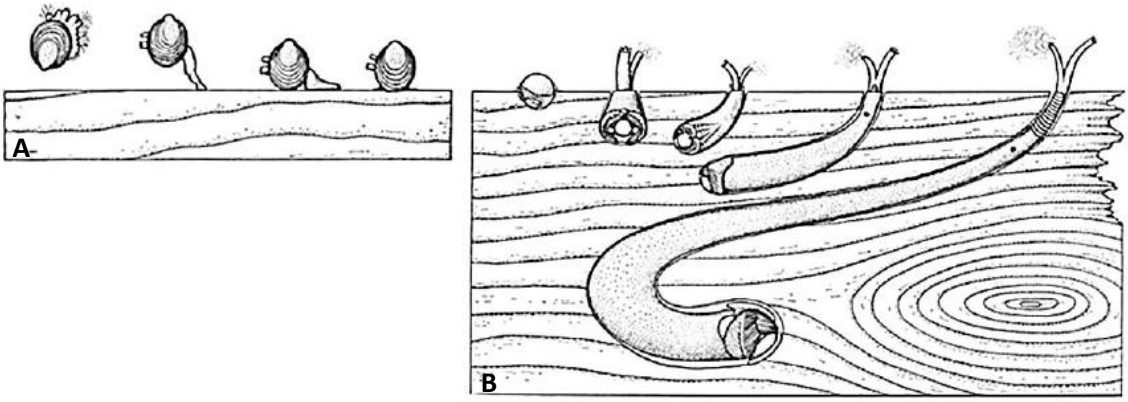


Figure 1.12. *Lyrodus pedicellatus* larvae settlement into the wood. **A.** The larvae are initially free-swimming, and once they find wood they start crawling and then the boring stage begins. **B.** Metamorphosis once boring has begun. Both sketches are taken from Nair and Saraswathy (Nair and Saraswathy, 1971)

1.3.5.4 Digestive system anatomy

Figure 1.13 presents a diagram of the principal organs of *L. pedicellatus*. In this section, we focus on describing the anatomy of the digestive system, because it is where wood is sorted, digested and absorbed, processes that are at the heart of this PhD thesis.

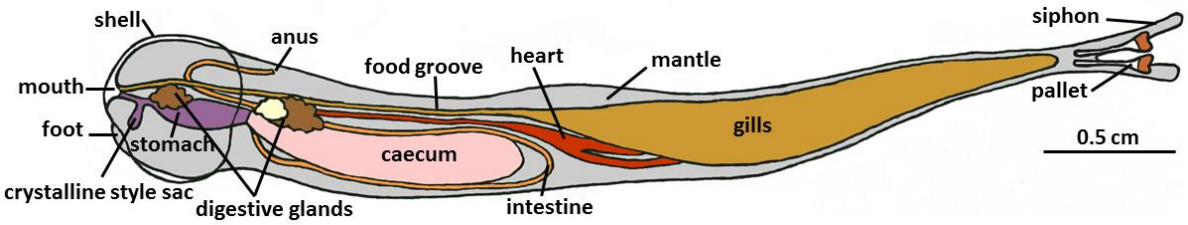


Figure 1.13. Diagram of the shipworm *L. pedicellatus*. The figure presents the principle organs, which are of importance to this study.

Gills

The gills perform three different functions, and this is likely the reason why they are the biggest shipworm organ and can cover up most of the body length in certain species (Distel *et al.*, 2011) (Figure 1.14). They carry out respiration and feeding, like in most of the bivalves, but in certain genera such as *Lyrodus* and *Teredo* they are also involved in brooding, as they host the larvae during their development (Turner, 1971b; Wurzinger-Mayer *et al.*, 2014). For this work, the respiratory and reproductive abilities will not be investigated further, and the gills are included in the discussion regarding the digestive system because of their involvement in both filter feeding and xylotrophy.

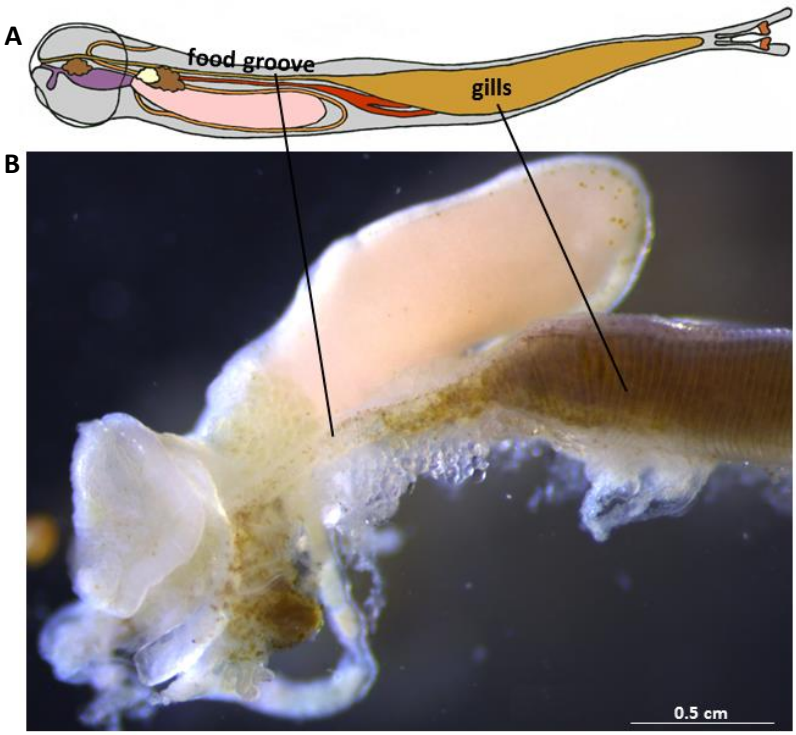


Figure 1.14. Light microscopy of the gills and food groove. **A**. *L. pedicellatus* diagram showing the anatomical position of the gills and food groove in the shipworm body. **B.** Image illustrating the gills (the brownish structures) on the right side and the food groove running along the lower portion of the caecum, which is pink sac on the middle top of the picture.

The teredinds gills – also called ctenidia - have the typical eulamellibranch arrangement, which consists of the repetition of multiple units, the lamellae or gills filaments, stacked together along the length of the gills and hanging into the mantle cavity (Figure 1.15). The lamellae are made up of two demibranchs or hemi-ctendia (Figure 1.16A), which are v-shaped and paired up at the ctenidal axis, to form the classical w-shaped structure. The demibranchs separate the mantle cavity into the infrabranchial and suprabranchial channels (Turner, 1966). The afferent and efferent branchial veins run along the ctenidial axis, near the branchial nerves (Gosling, 2015). On the peak of each v-shaped demibranch and along the length of the ctenidium is found a ciliated, mucous tunnel-like structure, the food groove (Figure 1.17) (Orton, 1912), which has a diameter of 50-60 µm, reducing toward to the anterior part of the gill (Shipway, 2013). There are two food grooves, one for each demibranch, which both extend after the end of the gills until the labial palps – running dorsally of the caecum - where they run alongside.

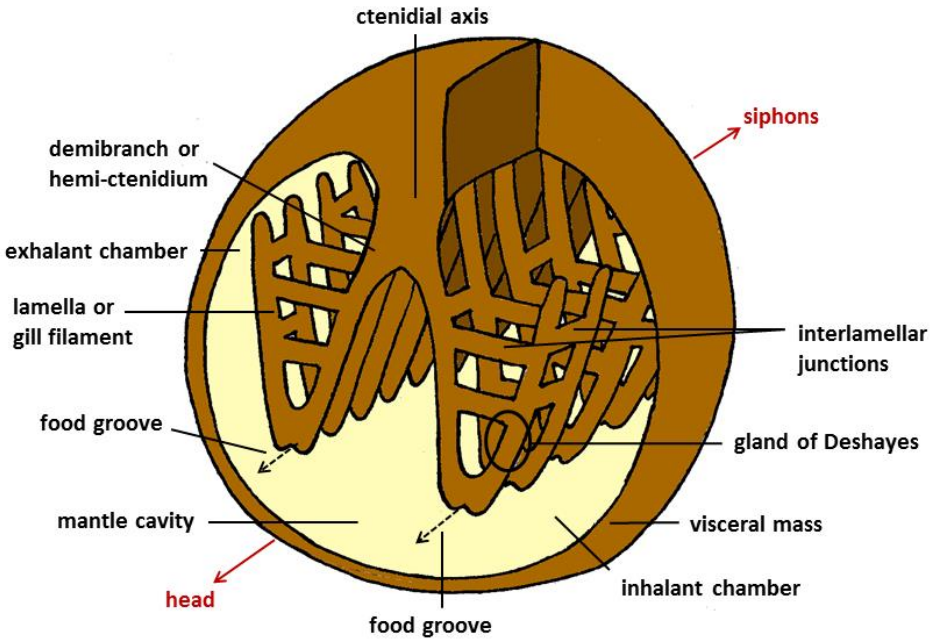


Figure 1.15. Shipworm gills. Diagram representing a transverse section of the shipworm body, showing the structure of the gills. The dotted lines represent the area where the food grove is found, and the arrows the direction of travel of the food particles. The red arrows indicate the location of the head and siphons, to give directionality to the drawing.

When water, drawn into the mantle cavity by the action of the inhalant siphon, passes along the ctenidia, the cilia (Figure 1.16B) and the mucus capture the suspended food and draw it toward the food grooves, which ultimately have the function of transporting it towards the mouth (Morton, 1970). The so-called gland of Deshayes, where the bacteriocytes are found, covers the area of the lamellae up to the afferent branchial vein (Distel *et al.*, 1991).

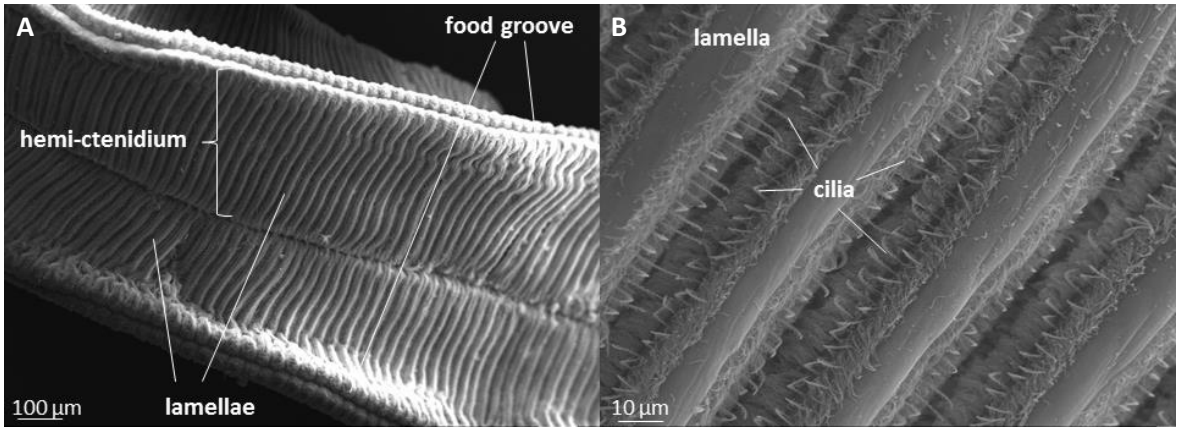


Figure 1.16. Scanning electron microscopy of *L. pedicellatus* gills. **A.** Picture showing the gills lamellae arranged in two v-shaped hemi-ctenidia, which pair up and form a w-shaped structure. Magnification 160x. **B.** Picture zooming close to the lamellae to show the numerous cilia that capture food and draw it to the food groove. Magnification 2,000x. Pictures by Clare Steele-King.

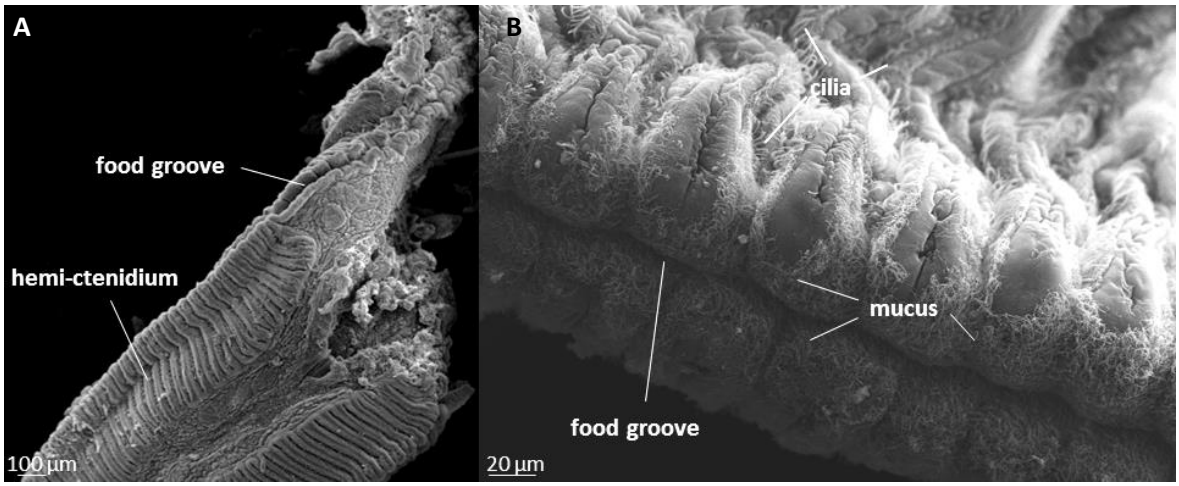


Figure 1.17. Scanning electron microscopy of *L. pedicellatus* food groove**. A.** Picture showing the food groove running at the apex of the hemi-ctenidium and then continuing towards the mouth once the lamellae are finished. Magnification 114x. **B.** Picture zooming close to the food groove to show the numerous cilia and some mucus (most of which is lost during the fixing procedure). Magnification 1,020x. Pictures by Clare Steele-King and Simon Cragg.

Caecum

The caecum (Figure 1.18) is a long, cylindrical, blind sac connected to the posterior side of stomach and positioned horizontally at the end of the digestive system (Sigerfoos, 1908), and is found only in wood boring bivalves (Purchon, 1941; Betcher *et al.*, 2012).

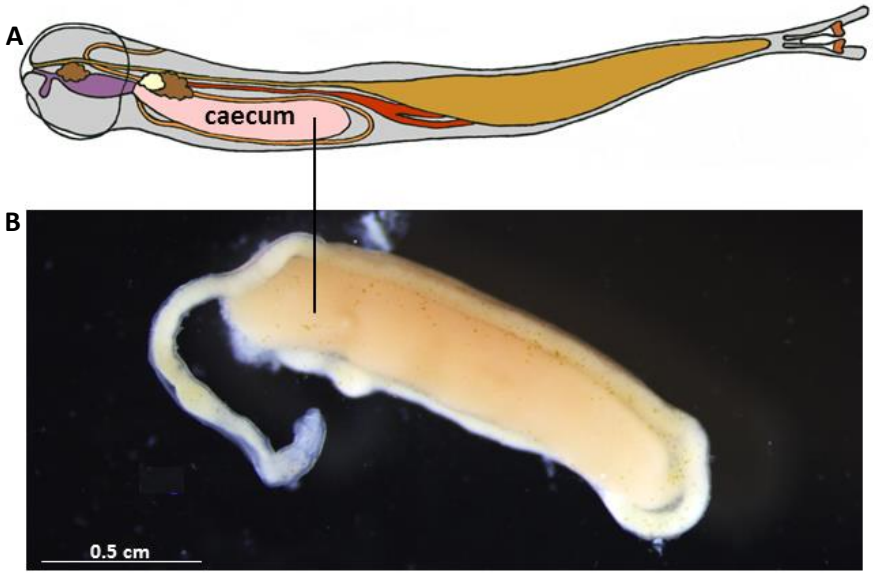


Figure 1.18. Light microscopy of the caecum. **A**. *L. pedicellatus* diagram showing the position of the caecum in the shipworm digestive system. **B.** Image illustrating the caecum (the pink sac) with the intestine (white tube) wrapped around it.

The caecum is filled with wood particles and its walls are lined throughout with a highly microvillar epithelium, with cilia also present close to the stomach opening (Bazylinski and

Rosenberg, 1983). The internal wall surface is folded inwards in the ventral region along the full length of the caecum to form a structure called typhlosole, which is also covered with microvilli and has the function of increasing the surface area, thus enhancing food digestion (Sigerfoos, 1908; Betcher *et al.*, 2012).

The caecum was initially thought to have the only function of storing wood, to be used when the timber was exhausted or when the animal was not boring, and it was believed it lacked the ability of digestion and absorption and that these functions were instead performed by the stomach and intestine (Purchon, 1941; Morton, 1970; Saraswathy and Nair, 1971). However, the presence of large amount of wood, the ciliated epithelium (Greenfield and Lane, 1953), the existence of the typhlosole and the extensive vascularisation of its tissues are all elements indicating that the caecum function is that of digestion and absorption of lignocellulosic material (Bazylinski and Rosenberg, 1983; Betcher *et al.*, 2012), as suggested by Sigerfoos in 1908.

Digestive glands

The adult shipworms have two separate digestive glands (Figure 1.19) projecting from the stomach and connected to it by ducts; the smaller one is found on the anterior part of the stomach, the bigger one is placed at the posterior end (Sigerfoos, 1908). They both have an almost spherical shape with a lobular appearance and are highly branched. Food particles are able to travel into the glands moved by the cilia found on the ducts epithelium (Potts, 1923).

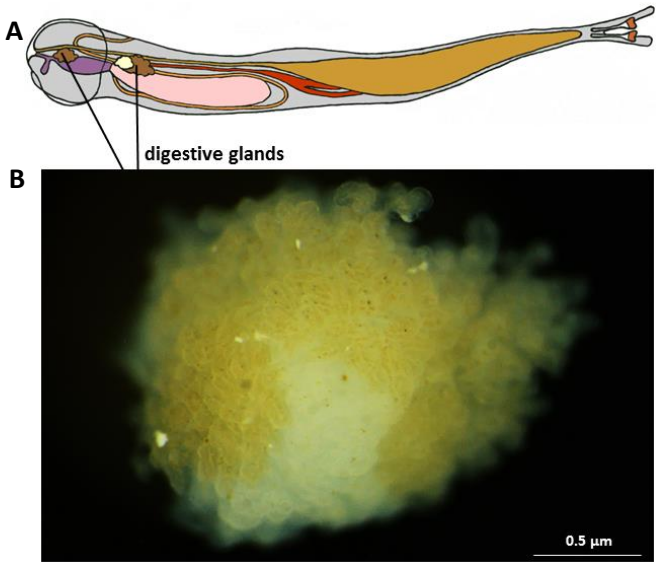


Figure 1.19. Light microscopy of the digestive glands. **A.** *L. pedicellatus* diagram showing the position of the digestive glands in the shipworm digestive system. **B.** Image illustrating the digestive gland, showing its "brown" and "white" portions, which are also named "excretory" and "digestive".

As their name suggests, the digestive glands have an important role in the digestion of both plankton and wood in the shipworms (Morton, 1970). The first author who recognised this function was Potts, describing in 1923 the details of the glands anatomy in the species *T. navalis*. She was also the first one to explain the presence of two distinct histological areas in the posterior gland, which were previously noted by Sigerfoos but left unexplained. One portion, which was called "excretory" by Potts, is characterised by the presence of a thick layer of columnar cells and a many long cilia and has the classical function of an exocrine gland, namely to synthesise and release substances, in this case digestive enzymes. The other portion, named "digestive", has quite a different histology, presenting short round tubules with thin walls and large amount of wood fragments in the lumen. This portion does not have a secretory nature, but instead is specialised in the intracellular digestion of wood, a characteristic peculiar to the Teredinids and probably associated with being able to digest wood (Potts, 1923). This unusual ability is performed by phagocytes, amoeboid cells found within the gland epithelium or free in the lumen (Figure 1.20). They are characterised by the presence of wood and the ability of projecting pseudopodia in the gland lumen in order to take up wood fragments into the cytoplasm (Potts, 1923). Morton suggests that not only the posterior gland, but also the anterior one presents the division into digestive and excretory portions (Morton, 1970), thought the latter was never observed during the dissections performed for this PhD thesis.

Figure 1.20. The phagocytes of *T. navalis*. Drawing extracted from Potts (1923) representing the phagocytes containing wood fragments found in the digestive portion of the digestive glands. According to the author, the wood can be found within vacuoles in some of these cells. Magnification 1,000x.

RNA extracted from the digestive glands of *L. pedicellatus* was sequenced (Shipway, 2013) and it revealed that around 5% of the glands transcriptome accounts for endogenously produced enzymes belonging to a range of glycosyl hydrolases classes, suggesting that it

has an important role in the production of enzymes for the digestion of wood, which are moved to the caecum where wood digestion takes place.

Crystalline style

The crystalline style is an organ found only in bivalves and in some gastropods. It is a gelatinous, acellular rod contained in and produced by an evagination of the stomach called crystalline style sac (Figure 1.21) (Mackenzie and Marshall, 2014). The anterior section of the style projects into the stomach and is in contact with the gastric shield, a chitinous structure at the back of the stomach (McQuiston, 1970). Cilia line the epithelium of the sac and make the style rotate and grind against the gastric shield, therefore consuming itself and realising the enzymes it contains (Judd, 1987). The role of the crystalline style is to help digestion by mechanically reducing the size of food particles, mixing digestive enzymes with them and redirecting the food to the right compartment (Nelson, 1917; Edmondson, 1920; Yonge, 1931; Lavine, 1946; Morton, 1952; Bailey and Worboys, 1960; Horiuchi and Lane, 1966; Kristensen, 1972; Judd, 1987; Alyakrinskaya, 2001).

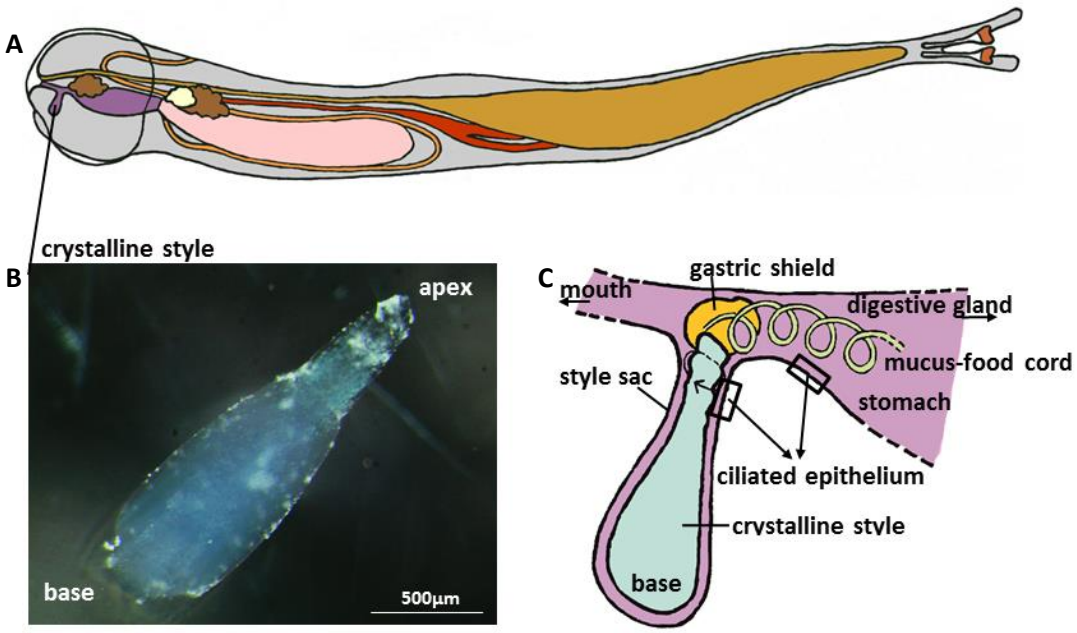


Figure 1.21. The crystalline style. **A.** *L. pedicellatus* diagram showing the position of the crystalline style in the shipworm digestive system. **B.** Light microscopy picture of the crystalline style once the sac has been removed. **C.** Diagram of the style showing its position in the anterior digestive system. The ciliated epithelium lining the style sac causes the style to rotate and scrape against the gastric shield, releasing the enzymes it contains and creating a cord of mucus and food. The cilia also line the stomach and help sorting food particles by size and redirecting them to the stomach, digestive glands or cecum.

The style of bivalves is mainly composed of water, with 10-20% being proteins and carbohydrates (Mackenzie and Marshall, 2014). The exact nature of the style proteins has not been studied in detail, but the presence of mucin-like proteins has been reported by some authors (Nelson, 1917; Morton, 1952; Bailey and Worboys, 1960), as well as that of globulin and albumen (Nelson, 1917; Mackintosh, 1925). Enzymes such as amylase, maltase, laminarinase, β-glucosidase, β-glucanase, cellulase, fucoidanase, xylanase, β-galactosidase, β-mannosidase and chitinase have been described in the style by various authors (Edmondson, 1920; Horiuchi and Lane, 1966; Coupin, 1900; Sova *et al.*, 1970; Wojtowicz, 1972; Alexander *et al.*, 1978; Smucker and Wright, 1984; Smucker and Wright, 1986; Hameed and Paulpandian, 1987; Sakamoto *et al.*, 2008; Xu *et al.*, 2002; Mackenzie and Marshall, 2014).

1.5 Aims of the project

As already discussed, many examples can be found in nature of organisms that can digest lignocellulose. The shipworm is one such organisms, and the extensive damage that it can cause to wooden structures is an indication of the wide array of enzymes that it is equipped with. This project sets out to explore the shipworms' mechanisms of wood digestion in detail, investigating the anatomical, physiological and molecular basis of wood digestion in the species *L. pedicellatus*. The final goal of this project is to gather in-depth knowledge of the processes that allow the shipworm, together with its gills endosymbionts, to break down wood and use it for nutritional purposes. This knowledge should address a series of questions regarding the nature, abundance, origin and movements of the lignocellulolytic enzymes utilised by this species for wood digestion. The information collected through this work could potentially be utilised both for the protection of man-made wooden structure from attack by shipworms and for the discovery of novel lignocellulosic enzymes to help the sustainable biofuel production industry.

Chapter Three aims to identify the main organs that are involved in wood digestion and the physiology of the process, using microscopic techniques, enzymatic assays and SDS gel electrophoresis. In particular, the chapter will address the open question regarding the route followed by bacterial enzymes as they relocate from their site of production in the gills to the caecum, where they are used for wood digestion. The hypothesis proposed is that there is no duct between the gills and the caecum, that the food groove (a mucous

stream utilised by filter feeding molluscs to transport food particles from the gills to the digestive system) is the only genuine connection, and that this structure has been co-opted by shipworms for the movement of bacteria and their enzymes. The digestive glands will also be examined, with the aim of identifying and imaging phagocytes, amoeboid cells involved in intracellular wood digestion that have so far only been described by hand drawings. Wood and frass will be analysed in their different components (cellulose, hemicellulose, lignin, ash and silica), to identify the main source of nutrition for the shipworm. The caecum will be tested for its ability to degrade different types of carbohydrate substrates and to confirm the presence of wood and of a microvillar typhlosole. Finally, the crystalline style will be investigated, since it has been overlooked in shipworms to date. Its involvement in extracellular wood digestion and its enzymatic properties will be explored, as well as the fine structure of the sac that supports the rotating mechanism of the style and that is involved in the secretion of proteins.

Chapter Four focuses on the transcriptomic and proteomic studies of the organs involved in wood digestion, with the aim of locating the sources of the endogenous and bacterial enzymes, as well as their final destination. The major genes and enzymes involved in wood digestion (as well as in immunity and sugar transport) will be investigated by the use of bioinformatics techniques, and the major CAZy classes will be identified. The premises for this study is the work of Shipway (2013). This work identified the digestive glands as the main organs where the endogenous wood-degrading enzymes are produced, and attributed only a limited role to the caecum. The hypothesis that the digestive glands are important for the production of proteins involved in immunity will be investigated, as the caecum has been reported to be almost bacteria-free (Betcher *et al.*, 2012). Finally, the chapter will look into the involvement of the caecum in the transcription of glucose transporters, which are essential in the uptake of sugars resulting from the degradation of wood.

Chapter Five will exploit the results from the transcriptomics and proteomics from chapter four to select some lignocellulosic enzymes for heterologous protein production, on the base of the enzyme abundance or transcript interest. The recombinant proteins produced in different types of host organisms will be purified and characterised to confirm the activity predicted with bioinformatics tools. Different carbohydrate substrates and environmental parameters such as pH, temperature and salinity will be tested to determine the enzymes preferences.

Chapter 2: Materials and methods

2.1 Animal cultures

2.1.1 Specimens acquisition and culturing

Shipworms of the species *Lyrodus pedicellatus* from the Atlantic lineage (see section 2.1.4 about species determination) were used for this work. Shipworm-infested wood composed of greenheart (*Chlorocardium* sp.) was initially collected from the pier of Portsmouth harbour, Hampshire, UK (50°47'47''N, 1°01'48''W). Larvae from the original wood were used to infest logs of Scots pine (*Pinus sylvestris*), which were kept in the laboratories of the Institute of Marine Science, University of Portsmouth, UK (Fig. 2.1). The tanks used to rear the animals measured 40x40x34 cm for a capacity of approximately 50 litres. The water was taken directly from the Langstone Harbour, Eastney, Hampshire, UK, using a flow-through system pumping approximately 10 litres per hour. The temperature was maintained constant at 15-18 °C using an Aquael Thermo-Precision electronic heater with a ATC-800+ microcomputer temperature controller, and it was kept aerated to prevent stagnation. Water salinity was kept constant at 34 PSU (Practical Salinity Unit) by using Perspex lids to cover the tanks and prevent evaporation.

L. pedicellatus infested wood logs were sent to the University of York, UK (Fig. 2.1) for the retrieval of individuals for analysis that needed to be performed with fresh animals or that required flash freezing in liquid nitrogen. The wood panels were extracted from the water, wrapped with a damp cloth and shipped in insulated containers. On arrival at the University of York (maximum one day after dispatch) the wood was immediately immersed in either seawater collected at Filey harbour, Yorkshire, UK (54°12'39''N, 0°16'60''W) or in artificial sea water prepared at least 24 hours before (to allow chlorine to evaporate) with marine salts (Seachem), for a salinity of 34 PSU. The logs were kept in plastic tanks of 35x27x27 cm for a capacity of 25 litres, constantly aerated with an air pump (Boyu S-500), and the water was left at room temperature (around 20 °C). The shipworms were allowed to acclimatise for a few days before performing the dissections (Foster *et al.*, 2010).

Figure 2.1 The tank set-up for the shipworm cultures at the Institute of Marine Sciences, University of Portsmouth (left) and at the University of York laboratories (right).

2.1.2 Wood preparation and compositional analysis

The Scots pine logs, measuring 15×10×2.5 cm, were prepared to host the shipworm by being impregnated in filtered seawater with a vacuum until waterlogged. They were then submerged in a flow-through tank for at least a month before being colonised by shipworms.

Compositional analysis was performed comparing the wood extracted from the panels and the frass (faeces) produced by the shipworms. The wood was finely grinded using a Cyclone Mill (Retsch). The frass was collected daily from the tanks with a 10 ml pipette and dried at 30 °C until all the moisture content was lost. Five different wood components were analysed: crystalline cellulose, hemicellulose (matrix polysaccharides), lignin, silica and ash content, following the protocols described below. The amount of the five components analysed were added together and the percentage of each fraction calculated from the total.

2.1.2.1 Hemicellulose content

The matrix polysaccharides (or hemicellulose content) were analysed using the trifluoroacetic acid (TFA) method (Foster *et al.*, 2010). In detail, 5 mg of wood or frass (in five biological replicates) were put into a 2 ml screw cap tube and 500 µl of 2 M TFA were added to perform hydrolysis. They were flushed with argon to displace the oxygen and incubated for four hours at 100 °C, vortexing every hour. The samples were cooled and the TFA was evaporated in a speed vacuum concentrator (SPD131DDA, Thermo Scientific) at

55 °C for two hours. The pellet was washed twice with 500 µl of 2-propanol, dried, and TFA-soluble sugars were removed in two extractions with 500 µl of dH_20. The supernatant was collected into a new tube, while the pellet was used for the crystalline cellulose content analysis (section 2.1.2.2). The supernatant (100 µl) was filtered with 0.45 µm filters and it was analysed by High-Performance Anion-Exchange Chromatography (HPAEC) using the monosaccharide program (see section 2.1.2.6). Standards comprising of a mixture of nine monosaccharides (arabinose, fucose, galactose, galacturonic acid, glucose, glucuronic acid, mannose, rhamnose and xylose) each at 100 µM were analyzed in the same way as the samples.

2.1.2.2 Crystalline cellulose content

The crystalline cellulose content was analysed using the anthrone-sulfuric acid method (Foster *et al.*, 2010). In detail, the pellets from section 2.1.2.1 were washed with 1.5 ml of dH_20 and then three times with 1.5 ml of acetone (C_3H_6O). The pellet was dried overnight and was then hydrolysed by adding 90 µl of 72% (w/v) sulfuric acid (H_2SO_4) and incubating at 25 °C for four hours, mixing every 15 minutes. Deionised water (1890) µl of were added to reduce the sulfuric acid concentration and they were then heated at 120 °C for four more hours. After cooling, the samples were centrifuged at 10,000 rpm for five minutes and the glucose content of the supernatant was quantified using the colorimetric anthrone assay (Leyva *et al.*, 2008) against a glucose standard curve. The samples (40 µl) were prepared by adding 360 µl of dH_20 and 800 µl of anthrone reagent. The samples and glucose standards were incubated at 80 °C for 30 minutes before transferring 200 µl to an optical plate and measuring the absorbance at 620 nm.

2.1.2.3 Lignin content

The lignin content was quantified using the acetyl bromide soluble lignin (ABSL) method (Fukushima and Hatfield, 2001). In detail, 400 mg of finely grinded wood or frass (in five biological replicates) were put into a 2 ml tube, and 250 µl of freshly prepared 25% acetyl bromide solution (25% v/v acetyl bromide in glacial acetic acid) were added to break the phenol bonds. The mixture was heated at 50 °C for three hours, vortexed every 15 minutes, and cooled at room temperature. It was then transferred into a 5 ml volumetric flask and 1 ml of 2 M sodium hydroxide (NaOH) was added, followed by 175 µl of 0.5 M hydroxylamine hydrochloride. After vortexing, the flask was filled up to 5 ml with glacial

acetic acid and mixed several times by inversion. Dilutions of 1 in 10 into glacial acetic acid were made, 1 ml of the solution was placed into a cuvette and the absorbance was measured at 280 nm against a blank of glacial acetic acid. The percentage of ABSL was measured using the following equation:

% ABSL = [absorbance/(coefficient x path length)] x [(total volume x 100 %)/biomass weight)] x dilution,where the coefficient used was 18.21 (poplar wood).

2.1.2.4 Silica content

Finely grinded wood or frass (1 g) was pelleted (in five biological replicates) using a 10 tons manual hydraulic press (Specac) for two seconds. The pellets were analysed using a portable x-ray fluorescence spectrometer (P-XRF, Niton XLt 900 GOLDD Analyzer, Thermo Scientific) to determine the amount of silicon present (Reidinger *et al.*, 2012). The pellets were scanned four times for each biological replicate, and a standard curve produced using certified reference materials was used to calculate the amount of silica.

2.1.2.5 Ash content

Wood or frass (1 g) was put in crucibles of known weight and placed in a convection oven at 600 °C for 24 hours. The samples were removed from the oven, cooled at room temperature and the weight was recorded. The ash content was determined with the following equation:

% ash = 100 x (weight of crucible and ash – weight of crucible)/(weight of crucible and sample – weight of crucible).

2.1.2.6 High-Performance Anion-Exchange Chromatography (HPAEC)

The monosaccharide content of wood and frass was analysed via HPAEC using an ICS-3000 PAD system with an electrochemical gold electrode, a CarboPac PA20 3x150 mm analytical column and a CarboPac PA20 3x30 mm guard column (Dionex). The protocol consisted in injecting 5 µl of sample and calibration standards and running them at a flow rate of 0.4-0.5 ml/minutes at a constant temperature. Initial equilibration with 100% dH_20 at 25 °C was followed by a linear gradient of 100% H_2O to 99%-1% of dH_2O-0.2 M NaOH in five minutes, then constant for 10 minutes, followed by a linear gradient to 47.5%-22.5-%-30% of dH_2O-0.2 M NaOH-0.5 M NaOH/0.1M NaOH in 7 minutes and then kept constant for 15 minutes. The column was re-equilibrated with 100% H_2O for 10 minutes before injection of the next

sample after a wash with 0.2 M NaOH for eight minutes. The identification of the carbohydrates was performed by comparison with the retention times of the standards and was quantified by comparing the integrated peak areas of the samples to those of the standards.

2.1.3 Specimen extraction and dissection

The wood logs were opened by splitting them with a hammer and screwdriver, and the animals were then extracted from the calcareous lining with tweezers, washed with seawater, placed in seawater containing EDTA-free protease inhibitors (1% v/v, Thermo Scientific) and kept on ice until dissection to anesthetize them. The dissections were performed using a stereomicroscope (Leica MZ6) after removing the mantle to expose the organs, which were dissected using tweezers and a scalpel and preserved with different methods depending on the downstream analysis to be performed.

2.1.4 Morphological and molecular species identification

Species identification was performed during dissections using pallets characteristics as described by Turner (1971b). To confirm the morphological identification, molecular identification was obtained by extracting DNA from either the mantle or the siphon tissues. The total genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Quiagen) following manufacture's protocol. The presence of DNA and its concentration and purity were determined by UV spectrophotometry using absorption at 260 nm. A 658 bp fragment from the 5' end of the cytochrome oxidase subunit I (COI-5P) and a 345 bp fragment of the 18S rRNA gene were amplified using primers and PCR set up as detailed in Borges et al. (2012). A longer fragment (1087 bp) of the 18S rRNA gene was also amplified to be able to distinguish *L. pedicellatus* from *Teredo navalis*, using the forward primer 5'-ACCGTCCCTTGGTGCTCTTG-3' and the reverse primer 5'-TACCTTGTTACGACTTTTACGACCTCTAATCC-3'. The DNA fragments were cloned by ligation cloning into the vector pSC-B-amp/kan using the Blunt PCR cloning Kit (StrataClone), followed by white-blue screening of the transformed cells and confirmation by colony PCR of the presence of the inserted DNA fragment. The DNA sequences were determined by Sanger sequencing and were blasted against the GenBank database to confirm species identity.

2.2 Microscopy

2.2.1 *In vivo* staining with x-cellobiose

Shipworms were extracted from the wood and the mantle was removed in order to expose the main digestive organs. The animals were immersed in filtered seawater containing EDTA free protease inhibitors (1% v/v, Thermo Scientific) and x-cellobiose (5-Bromo-4-chloro-3-indolyl β-D-cellobioside) in dimethylformamide (0.01% v/v). They were photographed using a stereomicroscope (Leica MZ6) at various time intervals for a total of four hours, to monitor for the appearance of blue coloration, to confirm the presence and distribution of active cellobiohydrolase activity.

2.2.2 Light microscopy

Shipworms were dissected and treated as for the *in vivo* staining (section above). Whole animals or single organs were photographed using a Leica M216F microscope mounted with a SPOT RT3 camera model 25.2 2Mp colour mosaic. Samples of the digestive glands were observed by placing a whole gland on a microscope slide and pressing the coverslip against it, to break the gland's lobule open and release the phagocytes to be photographed. Semi-thin sections (0.5 µm) of gills tissue were cut from resin blocks prepared for TEM (section 2.2.3) and adhered to glass slides (SuperFrost. Thermo Fisher Scientific) before staining with toluidine blue (0.6% toluidineblue in 0.3% sodium carbonate). Both gills and digestive glands sections were observed using a Nikon Eclipse E600 microscope mounted with the same camera as above.

2.2.3 Transmission electron microscopy

Shipworms were extracted from the wood, the mantle was removed and gills, food groove, caecum, digestive glands and crystalline style were dissected. The organs were fixed for 1–2 hours at room temperature in primary fixative (4% formaldehyde (w/v), 2.5% (w/v) glutaraldehyde in 100 mM sodium phosphate buffer pH 7.2) and washed in 100 mM sodium phosphate buffer pH 7.2 three times for 10 minutes. They were incubated in secondary fixative (1% osmium tetroxide in 100 mM sodium phosphate buffer pH 7.2) for one hour on ice. Samples were dehydrated through a graded ethanol series for 15 minutes each, followed by two washes of five minutes in epoxy propane. Samples were then infiltrated with a series of epoxy propane/Epon araldite (25%, 50%, 75% Epon Araldite with a

minimum of one hour at each stage, all at 30 °C), concluding with a minimum of two changes of Epon araldite resin over 24 hours at 30 °C, and polymerized at 60 °C for 48 hours in flat embedding moulds. Pale gold (70–90 nm) ultra-thin sections were cut with a Diatome diamond knife, using a Leica Ultracut UCT microtome, and mounted on hexagonal 200-mesh nickel grids. Sections were post-stained with 2% (w/v) aqueous uranyl acetate for 10 minutes, then with lead citrate for five minutes in a carbon dioxide-free chamber, and they were viewed using a FEI Tecnai 12 BioTWIN G2 TEM operating at 120 kV. Images were captured using AnalySIS software and a Megaview III CCD camera.

2.2.4 Immunogold labelling

2.2.4.1 Fixing, embedding and sectioning

Embedding for immunogold labelling proved difficult and several attempts were made to allow resin infiltration and at the same time preserve antigenicity. Freshly dissected shipworm tissues (food groove, gills and caecum) were fixed with 4% paraformaldehyde, 0.2% glutaraldehyde in sodium cacodylate buffer (0.2 M sodium cacodylate, 0.3 M sodium chloride, 2 mM calcium chloride, pH 7.4) on ice in a vacuum chamber for two hours, then on a rotator without vacuum for a further 12 hours at 4 °C. Samples were washed in 0.2 M sodium cacodylate buffer (three washes of twenty minutes each) and dehydrated through a graded ethanol series initially on ice (50%) and subsequently at -20 °C on a rotator (70%, 90%, 100%) with 20 minutes at each stage and two changes of 100% ethanol. Ethanol was gradually replaced with LR Gold resin (1:2, 1:1, 2:1 resin:ethanol) with one hour at each stage, followed by three changes of 100% LR Gold resin, 12 hours each, all at -20 °C on a rotator. Tissues were embedded in closed gelatine capsules and polymerised with UV light at -20 °C for 24 hours, followed by 24 hours at -10 °C. Pale gold (70–80 nm) ultra-thin sections were cut with a Diatome diamond knife, using a Leica Ultracut UCT microtome, and mounted on hexagonal 200-mesh nickel grids. All immunolabelling steps were achieved by floating grids on droplets of reagent. Sections were incubated in blocker (3% BSA in PBS pH 7.0) for 30 minutes at ambient temperature before incubation with primary antibodies against the bacterial CAZyme *Lps*GH5_8 (see section 2.2.4.2 for details on antibodies production). They were diluted 1:100 in 1% BSA in PBS at 30 °C for one hour and washed with PBS at ambient temperature (three brief washes followed by three washes of ten minutes each). Sections were incubated in secondary antibody (goat anti-rabbit IgG conjugated to 10 nm gold), diluted 1:100 in 1% BSA in PBS for one hour at 30°C, followed

by washes with PBS as before, and subsequently with ultrapure water. All immunolabelling procedures included negative controls of pre-immune serum (diluted 1:100 in 1% BSA in PBS) and buffer only (1% BSA in PBS). Sections were post-stained with 2% (w/v) aqueous uranyl acetate for 10 minutes, then lead citrate for five minutes in a carbon dioxide-free chamber and viewed using a FEI Tecnai 12 BioTWIN G2 operating at 120 kV. Images were captured using AnalySIS software and a Megaview III CCD camera.

2.2.4.2 Antibodies production and purification

The purified recombinant bacterial *Lps*GH5_8 (see chapter 5.4 for its cloning, expression and purification) was used (2 mg) to raise polyclonal antibodies in rabbits (ProteoGenix, France). To enrich the serum for antigen-specific antibodies, affinity columns were made using recombinant *Lps*GH5_8. Recombinant protein preparations were dialyzed against coupling buffer (0.1 M NaHCO3, 0.5 M NaCl, pH 8.3) and bound to CNBr-activated Sepharose™ 4 Fast Flow resin (GE Healthcare Life Sciences), followed by affinity purification of an aliquot of the crude antibody serum according to the resin manufacturer's instructions. The pre-immune serum was subject to the same purification procedure. Purified antibody and pre-immune serum fractions were characterized for their affinity by western blotting using both recombinant *Lps*GH5_8 (Fig. 2.2) and *L. pedicellatus* caecum fluids (not shown). Fractions showing the highest titre and no unspecific binding were selected for immunogold labelling.

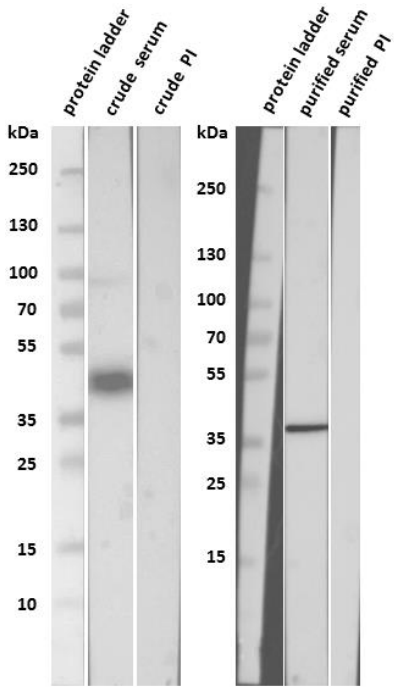


Figure 2.2. Western blot obtained by loading 6 µg of the recombinant purified bacterial *Lps*GH5_8 on a SDS-polyacrylamide gel and transferring into a nitrocellulose membrane. The left picture shows the blotting performed with either the crude serum or crude pre-immune serum (PI), while the right one shows the same procedure performed with purified serum and PI. The serum or PI concentration was 1:100, while the secondary antibody (stabilized goat anti-rabbit IgG (H+L) peroxidase-conjugated antibodies from Pierce) concentration was 1:500. Protein visualisation was performed using the SuperSignalT$_M$ West Pico PLUS chemiluminescent substrate (Thermo Scientific) and using the Syngene PXi gel documentation imaging system**.** A signal for the crude and purified serum is seen between 35 and 55 kDa, which corresponds to the predicted molecular weight for *Lps*GH5_8, which is around 42 kDa without the signal peptide. No signal is seen for the pre-immune serum.

2.2.5 Scanning electron microscopy

Samples for scanning electron microscopy were fixed in 4% (v/v) glutaraldehyde in a cacodylate buffer (0.2 M sodium cacodylate, 0.3 M sodium chloride, 2 mM calcium chloride) for two hours at room temperature and then rinsed once in buffer for thirty minutes. Samples were taken through an ethanol dehydration series (50-70-100% ethanol and twice in 100% acetone, each stage for 30 minutes), critical point dried and then mounted on aluminium stubs using adhesive carbon tabs. Sputter coating was carried out under an argon atmosphere using a gold and palladium target, at a voltage of 1.4 kV using a current of approximately 18 mA for three minutes. Specimens were examined using a Zeiss MA10 Scanning Electron Microscope with an accelerating voltage of 20 kV and the Zeiss Smart software.

2.3 Transcriptomic analysis

2.3.1 RNA extraction, clean up, depletion and quantification

Gills, digestive glands and caecum were dissected from two adults *L. pedicellatus*, as well as crystalline style sacs from 38 animals (which were then pooled together), flash frozen in liquid nitrogen and stored at -80 °C. Total RNA was extracted using TRIzol Reagent (Thermo Fisher Scientific) by placing the samples in a 2 ml tube, adding 1 ml of TRIzol Reagent and homogenising the tissues with a sterilised Teflon pestle. The tubes were vortexed for 10 seconds and incubated at room temperature for five minutes to allow dissociation of nucleoprotein complexes, and the lysate was flash-frozen in liquid nitrogen to disrupt the tissues further. After defrosting at room temperature, 200 µl of chloroform ($CHCl_3$) were added, vortexed for 15 seconds and centrifuged at 14,000 g for five minutes at room temperature. The chloroform extraction was repeated with 500 µl and then the top phase was transferred to a fresh 2 ml tube and 10 µg of glycogen were added, followed by 500 µl of isopropanol, mixing by inversion for several times and incubation at room temperature for 15 minutes. The tubes were centrifuged at 14,000 g for 10 minutes at 4 °C and the supernatant was removed, then 1 ml of cold 70% ethanol were added and the tubes were centrifuged at 14,000 g for five minutes at 4 °C. The supernatant was removed and air dried for 10 minutes and then re-dissolved in 50 µl of RNAse free water and heated at 56 °C for 10 minutes. To confirm the RNA integrity, 3 µl aliquots of each sample were loaded onto a 1 % w/v agarose gel, electrophoresed at 120 V for 15 minutes and the RNA was then visualised under UV excitation. DNase treatment was carried out with Turbo DNA-free kit (Ambion) following manufacture's instruction, RNA was cleaned with RNA Clean & Concentrator™-5 kit (Zymo Research) following manufacture's instruction and was then quantified with a Qubit 3.0 Fluorometer and Agilent TapeStation. RNA depletion for both eukaryotic and prokaryotic ribosomal RNA was performed with the Ribo-Zero™ Magnetic Gold Kit Epidemiology (Epicentre) and mRNA was then concentrated using RNA Clean & Concentrator™-5 kit (Zymo Research), following manufacturer's instructions.

2.3.2 cDNA library preparation and RNA sequencing

RNA-Sequencing libraries from gills, digestive glands and caecum were prepared from each mRNA sample using the Ion Total RNA-Seq kit v2 (Thermo Fisher Scientific) following manufacturer's instructions and using an RNaseIII treatment time of 2.5–3.0 minutes.

Samples were barcoded using the Ion Xpress RNA-Seq Barcode kit (Thermo Fisher Scientific) following manufacturer's instruction. Yields and library sizes were assessed using the High Sensitivity D1K screentapes and reagents on a 2200 TapeStation Nucleic Acids System (Agilent Technologies). Appropriately diluted library aliquots were combined in pairs in equimolar amounts and used for template preparation using the Ion OneTouch 200 Template Kit v2 DL on a OneTouch system (Thermo Fisher Scientific), prior to loading onto a 318 chip, and they were then sequenced on an Ion Torrent PGM™ prepared as per the manufacturer's instructions (IonPGM200 Kit, Thermo Fisher Scientific).

The sequencing of the crystalline style sacs was performed at the Next Generation Sequencing Facility at the University of Leeds with HiSeq3000 using 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.

2.3.3 Contig assembly and transcriptome analysis

Quality control was performed on the raw reads with the programs FastQC (Schmieder and Edwards, 2011) and Bowtie2 (Langmead and Salzberg, 2012) and trimming of low quality reads was performed with the software Trimmomatic (Bolger *et al.*, 2014). The reads were assembled into contigs using the Trinity software (Grabherr *et al.*, 2011) and those shorter than 500 base pairs were discarded. After mapping the raw reads back on the contigs, gene expression levels were calculated as TPM values (Transcripts Per kilobase Million) for the best 10,000 transcripts using the following protocol:

- divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK);
- count up all the RPK values in a sample and divide this number by 1,000,000. This is the "per million" scaling factor;
- divide the RPK values by the "per million" scaling factor. This gives you the TPM value.

Annotation of the contigs was performed by BlastX searches against the non-redundant database of the National Center for Biotechnology Information (NCBI nrdb, https://blast.ncbi.nlm.nih.gov). Nucleotide sequences were translated into amino acid sequences using the ExPASY online tool (https://web.expasy.org/translate/). The online

software dbCAN (database for automated Carbohydrate-Active enzyme aNnotation, http://csbl.bmb.uga.edu/dbCAN/) was used to search for carbohydrate active domains after the contigs were converted into ORFs using the online tool Emboss (http://www.bioinformatics.nl/cgi-bin/emboss/getorf). Results with an e-value <$1e^{-10}$ and those with a CBM (Carbohydrate Binding Module) but no annotation were excluded, as well as CAZymes (Carbohydrate-Active enZYmes) belonging to the class of glycosyl transferases.

2.4 Proteomic analysis

2.4.1 Organs dissection and preparation

Gills, digestive glands and caeca from the same five individuals were dissected in 50 mM sodium phosphate buffer pH 7.0 containing EDTA-free protease inhibitors (1% v/v, Thermo Scientific) and kept on ice. They were pooled together by organ type, and the contents of the caeca were collected by placing them in 150 µl of the same buffer, cutting them open with a surgical blade and extracting the fluids containing enzymes and wood particles. Crystalline styles from 21 animals were also collected and pooled together, but they were not suspended in phosphate buffer. SDS loading buffer (1% SDS, 10% glycerol, 0.1% bromophenol blue and 100 mM 2-mercaptoethanol) was added and the samples were heated at 100 °C for 10 minutes, centrifuged, and the supernatant was run into a 4-20% polyacrylamide gel and stained with InstantBlue protein stain (Expedeon) to confirm the presence and amount of protein.

2.4.2 Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) analysis

In order to identify the protein content of the samples, in-gel tryptic digestion was performed with 200 ng of sequencing grade trypsin (Promega) and post reduction with dithioerythritol (1.5 mg/ml) and *S*-carbamidomethylation with iodoacetamide (9.5 mg/ml). Digestion was carried out overnight at 37 °C. The resulting peptides were analysed by label-free LC-MS/MS using an UltiMate 3000 RSLCnano HPLC system interfaced with an Orbitrap Fusion hybrid mass spectrometer (Thermo). Peptides were eluted from a PepMap, 2 µm, 100 Å, C18 EasyNano nanocapillary column (75 µm x 150 mm, Thermo) at 300 nl/minutes using a gradient elution of aqueous 1% (v/v) formic acid (solvent A) and aqueous 80% (v/v) acetonitrile containing 1% (v/v) formic acid (solvent B). The gradient was 3–10% of B over eight minutes, then 10–35% of B over 125 minutes, and then 35–65% of B over 50 minutes. Positive ESI-MS and MS2 spectra were acquired using Xcalibur software (version 4.0,

Thermo). Data-dependent acquisition was performed in top speed mode with a 1-s cycle. MS2 spectra were acquired in the linear ion trap with HCD activation energy of 32%.

2.4.3 Proteomic analysis

Protein identification was performed by searching tandem mass spectra against the assembled transcriptome of *L. pedicellatus* using the Mascot search program (http://www.matrixscience.com). Matches were passed through Mascot percolator to achieve a false discovery rate of <1% and further filtered to accept only peptides with expect scores of 0.05 or better and with a number of at least two significant sequences (unless stated in the results). 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 (Ishihama *et al.*, 2005). Proteins identified in the proteomics analysis were annotated by BlastX searches against the NCBI nrdb (https://blast.ncbi.nlm.nih.gov). Nucleotide sequences were translated into amino acid sequences using the ExPASY Translate online tool (https://web.expasy.org/translate/) and protein parameters calculated with the ExPASY ProtParam tool (https://web.expasy.org/protparam/). CAZy annotation was carried out using the online software dbCAN http://csbl.bmb.uga.edu/dbCAN/) and results with an e-value <$1e^{-10}$ and those with a CBM (but no annotation) were excluded, as well as CAZymes belonging to the class of glycosyl transferases. Putative N-terminal signal peptides were predicted with the online server SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/) or SignalP 3.0 (Petersen *et al.*, 2011).

2.4.4 Protein identification by MALDI/TOF-TOF tandem mass spectrometry

The identity of certain proteins separated by SDS-PAGE was investigated by MALDI/TOF-TOF tandem mass spectrometry. The protein band under examination was excised from the gel with a razor blade, cut into 3-4 pieces and placed in a 0.5 ml LoBind tube (Eppendorf). The band fragments were washed twice with 50% (v/v) aqueous acetonitrile containing 25 mM ammonium bicarbonate to remove the Coomassie stain and other contaminants, the cysteine residues were reduced with 10 mM DTE (100 mM ammonium bicarbonate, 10 mM dithioerythrietol) for one hour at 56 °C, and they were then alkylated and *S*-carbamidomethylated with 100 mM ammonium bicarbonate and 50 mM iodoacetamide for 30 minutes in the dark at room temperature. Following dehydration with acetonitrile for five minutes, gel pieces were dried in a SpeedVac, digested with the

addition of 0.2 µg sequencing-grade modified porcine trypsin (Promega) in 25 mM ammonium bicarbonate and incubated at 37 °C overnight. Peptides were extracted from the gel by washing three times with 50% (v/v) aqueous acetonitrile, dried in a SpeedVac and reconstructed in 0.1% (v/v) acqueous trifluoroacetic acid. Samples were analised by MALDI/TOF-TOF tandem mass spectrometry as described in detail by Alessi and colleagues (2017). Protein identification was performed by searching tandem mass spectra against the assembled transcriptome of *L. pedicellatus* or the NCBI nrdb using the Mascot search program (http://www.matrixscience.com).

2.5 Molecular biology techniques

2.5.1 cDNA production

RNA was extracted from the digestive glands and gills of *L. pedicellatus* using the Trizol method, and it was cleaned and concentrated as described in section 2.3.1. A poly(A) tail was added to the gills' RNA only using the poly(A) polymerase from Takara, following manufacturer's protocol (Sippel, 1973). cDNA was produced using the SuperScript II reverse transcriptase (Thermo Fisher Scientific). In short, 1 µl of oligo-dT primer 10 µM, 1 µl of dNTP mix (10 mM each), 1 to 5 ng of total RNA and RNAse free water to make up 12 µl were put in a microcentrifuge tube, heated at 65 °C for five minutes and quickly chilled on ice. After brief centrifugation, 4 µl of 5x First-Strand buffer, 2 µl of 0.1 M DTT and 1 µl of RNase free water were added, the contents were mixed gently and incubated at 42 °C for two minutes. SuperScript II reverse transcriptase (1 µl) was added, mixed by gentle pipetting and incubated at 42 °C for 50 minutes. The reaction was inactivated by heating at 70 °C for 15 minutes and the cDNA was purified with the Clontech NucleoSpin PCR Clean-up and gel extraction Kit.

2.5.2 Polymerase chain reaction (PCR)

PCR was performed to amplify the sequences encoding for the selected CAZymes from cDNA produced from the gills (for the bacterial enzymes) or the digestive glands (for the shipworm ones). The oligonucleotide primers used for annealing to the cDNA template (Table 2.1) were designed so that each primer had (when possible) a length of 18-22 bp, a melting temperature (T_m) between 52 and 58 °C with a difference of no more than 3 °C between the pair, GC content not exceeding 60%, and no repeats or runs longer than four

nucleotides. The T_m values and GC content were calculated using the Integrated DNA Technologies online calculator (https://www.idtdna.com/calc/analyzer). Amplification of the genes of interest was performed with the Phusion High-Fidelity DNA Polymerase (Thermo Fisher), using a PTC-200 Peltier Thermal Cycler (MJ Research). The reaction reagents and quantities, and the thermo-cycling conditions are listed in Table 2.2 and 2.3. 3 µl of the reaction products were analysed by agarose gel electrophoresis to confirm the presence and length of the DNA product.

Table 2.1. Primers developed in this study for the cloning of the bacterial proteins *Lps*GH5_8, *Lps*GH11, *Lps*GH134a and *Lps*GH134b. *Lps*AA10 could not be amplified from the cDNA and therefore a synthetic version of the gene was codon-optimised for *E-coli* expression.

Name of protein	Oligonucleotide primer sequence
***Lps*GH5_8**	Forward primer: GTTTCAGAGTAGTTAGCTGAACCAGCC
	Reverse primer: GAGTGTCGAGTTAGCACGGTGG
***Lps*GH11**	Forward primer: ACTAGCTACCTTTCACTCAAGTTGTG
	Reverse primer: CAGCCGTTGTTTTTAGTTACAGCT
***Lps*GH134a**	Forward primer: CACGGTACTCACACATACACAAATATAG
	Reverse primer: ATTGGCTTAACCAGGGTAAGC
***Lps*GH134b**	Forward primer: TATGGGTGCAAACTTAAGCTGC
	Forward primer: GTATCGTTGGATTGGCTTAACG

Table 2.2. Reaction reagents and quantities used for the PCR reactions.

Name of reagent	Concentration (µl)
Phusion buffer HF (5X)	10.0
Forward primer (20 µM)	1.0
Reverse primer (20 µM)	1.0
dNTP mix (dATP, dGTP, dCTP, dTTP) 10µM	1.0
Phusion High-Fidelity DNA Polymerase	0.5
Template (cDNA)	0.5
RNAse free water	33.5
DMSO (if needed)	2.5
Total	**50.0**

Table 2.3. The thermo-cycling conditions used for the PCR reactions.

Step	Temperature (°C)	Time (seconds)	Number of cycles
Initial denaturation	98	120	1
Denaturation	98	10	35
Annealing	55	30	35
Elongation	72	30 seconds/kb	35
Hold	4	infinite	1

2.5.3 3’ and 5’ Rapid Amplification of cDNA Ends (RACE)

Rapid amplification of cDNA ends was performed for the proteins *Lps*GH5_8 and *Lps*GH134a, whose 5’ and 3’ ends (respectively) were missing in the contigs generated from the RNA sequencing. The SMART RACE cDNA Amplification Kit from Clontech was used, following manufacturer’s instructions. The cDNA was obtained from *L. pedicellatus* gills and the primers were designed for RACE and nested RACE and are listed in Table 2.4. Touchdown PCR was performed for *Lps*GH5_8 only, since primers’ with T_m temperatures above 70 °C could be designed.

Table 2.4. RACE and nested RACE primers used for the cloning of the bacterial proteins *Lps*GH5_8 and *Lps*GH134a with the SMART RACE cDNA Amplification Kit.

Protein name	Oligonucleotide primer sequence
***Lps*GH5_8**	Forward 5’ RACE primer: CGCATCACATTCTCCCAATCCTGTCCCCAGTTCGGTGCA
	Forward 5’ nested RACE primer: GCAGCGCCTGCAATTGTTCCTGCAGCGCCTTCTTCTC
***Lps*GH134a**	Forward 3’ RACE primer: TGCCGGACACCGCAACGGCTCAAGTGG
	Forward 3’ nested RACE primer: TAACCCCAACACCGATGATATTCAACGG

2.5.4 DNA purification

The DNA obtained from the PCR reactions was purified using the Wizard SV gel and PCR clean-up System (Promega) following manufacturer’s instructions. Briefly, an equal volume of membrane binding solution was added to the PCR amplification products, and the mixture was transferred to a minicolumn, incubated for one minute and centrifuged at 16,000 g for one minute. After discarding the flowthrough, 700 µl of membrane wash solution were added and it was then centrifuged at 16,000 g for one minute. The step was repeated with 500 µl of membrane wash solution and centrifuging for five minutes. After allowing evaporation, the DNA was transferred into a new microcentrifuge tube, 50 µl of nuclease-free water were added, it was incubated at room temperature for one minute and centrifuged at 16,000 g for one minute. The clean PCR product was then used for downstream analysis or stored at -20 °C.

2.5.5 Nucleic acids quantification

The concentration of RNA, ssDNA or dsDNA was measured by spectrophotometry, using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). The samples (1.2 µl) were tested and the concentration was measured by the 260/280 ratio against a blank. A ratio

of 1.8 was considered pure for DNA and of 2.0 for RNA, with lower values indicating possible protein, phenol or other substances contamination. The 260/230 ratio was also used to check for other contaminants, with DNA and RNA considered pure if the ratio was in the range of 2.0-2.2.

2.5.6 Nucleic acid agarose gel electrophoresis

Agarose gel electrophoresis was used to visualise and separate DNA fragments of different sizes. The gels were prepared at a concentration of 0.8% agarose in 0.5x Tris-borate-EDTA buffer (TBE) buffer (45 mM Tris 45mM boric acid, 1 mM EDTA, pH 8.0) by heating in a microwave at full power for 1-2 minutes (depending on the gel size and thickness). The agarose-TBE solution was cooled at room temperature and 0.5 μg/ml ethidium bromide were added before letting the gel set in a cast with a well comb. Once solidified, the comb was removed and the gel was placed in an electrophoresis tank filled with 0.5x TBE buffer. The samples were prepared for analysis by mixing with 5x loading buffer (50 mM Tris, 5 mM EDTA, 50% (v/v) glycerol, 0.3% (w/v) orange G, pH 8.0) and they were loaded into the well alongside 5 μl of the molecular weight marker 1 kb Hyperladder (Bioline) to estimate size. The gels were normally 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).

2.5.7 StrataClone subcloning

The initial cloning of the genes of interest was performed using the StrataClone Blunt PCR cloning kit (Agilent Technologies) for improved efficiency and easier subsequent cloning into the expression vectors, and to verify the DNA sequence with Sanger sequencing. The cloning works by performing ligation of blunt-end PCR products by means of the re-joining activity of DNA topoisomerase I and the DNA recombination activity of the Cre recombinase. The vector used was pSC-A-amp/kan (Fig. 2.3), which is provided in the kit and is 4,269 base pairs long. It includes a *lacZ* α-complementation cassette for white/blue colony screening, kanamycin and ampicillin resistance, a multiple cloning site (MCS) with restriction sites, a *lac* promoter and operon, a *f1* origin of ssDNA replication and a pUC origin. Cloning into pSC-A-amp/kan was performed following manufacturer's instructions. In short, the DNA fragments obtained after performing PCR and DNA purification were quantified and diluted 10 times in nuclease free water if agarose gel electrophoresis

confirmed a strong and specific amplification. The ligation reaction was performed in PCR tubes by adding the following components in this order:

- 3 µl of StrataClone Blunt Cloning Buffer;
- 2 µl of the PCR product (diluted if necessary);
- 1 µl of StrataClone Blunt Vector Mix amp/kan.

After gentle mixing by repeated pipetting, the ligation reaction was left at room temperature for five minutes and then put on ice. 1 µl of the reaction was used to transform 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 colony screening for blue-white colonies (see next section). The plates were incubated at 37 °C overnight.

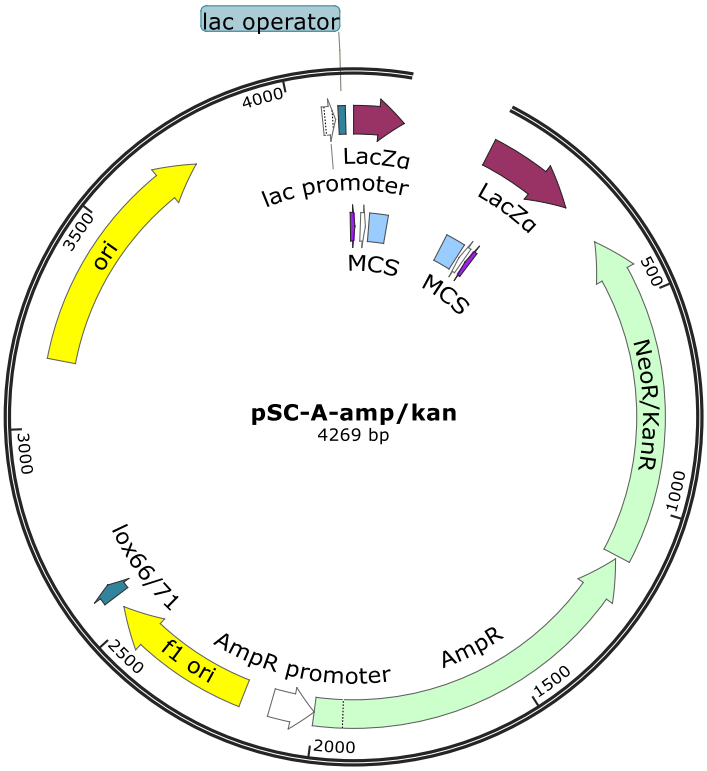


Figure 2.3. Map of the linearised StrataClone PCR Cloning Vector pSC-A-amp/kan. MCS=multiple cloning site, AmpR, NeoR and KanR=ampicillin, neomicin and kanamicin resistance.

2.5.8 Colony screening

In order to identify bacterial colonies containing the gene of interest after a transformation, colony screening was performed. The procedure involved picking a colony from a plate (only white colonies picked in case of white/blue screening) using a sterile pipette and dipping it into a PCR mixture prepared with the reagents listed in Table 2.5. Colony PCR was

performed by using the thermo-cycling conditions listed in Table 2.6. The primer pair used was the M13 forward and reverse primers for the pSC-A-amp/kan vector and the T7 promoter and terminator for the other vectors used. The PCR products (5-10 µl) were electrophoresed on an agarose gel and visualised under UV light to confirm the presence of the transformants in the picked colonies.

Table 2.5. Reaction reagents and quantities used for the colony PCR reactions.

Name of reagent	Concentration (µl)
DreamTaq Green PCR Master Mix (2x, Fermentas)	5.0
Forward primer (20 µM)	0.3
Reverse primer (20 µM)	0.3
RNAse free water	4.4
Total	**10.0**

Table 2.6. The thermo-cycling conditions used for the colony PCR reactions.

Step	Temperature (°C)	Time (seconds)	Number of cycles
Initial denaturation	95	120	1
Denaturation	95	30	35
Annealing	55	30	35
Elongation	72	60 seconds/kb	35
Final extension	72	300	1
Hold	4	infinite	1

2.5.9 Plasmid DNA extraction

The colonies identified as positive transformants (i.e. containing the plasmid with the gene of interest) were incubated overnight at 37 °C and shaking at 200 rpm in 10 ml of LB media containing the appropriate antibiotic/s. The bacterial cells were then harvested using the Wizard *Plus* SV Miniprep DNA Purification kit (Promega), according to the manufacturer's instructions.

2.5.10 DNA sequencing (Sanger sequencing)

The plasmid DNA obtained from the previous section was sent for sequencing at the DNA Sequencing and Services (Dundee University), following the instructions provided in their website (https://www.dnaseq.co.uk/). The primer pair used for sequencing was the M13 forward and reverse primers for the pSC-A-amp/kan vector and the T7 promoter and

terminator for the other vectors used. The obtained sequence chromatograms were analysed using the SnapGene software (GSL Biotech).

2.5.11 In-Fusion HD cloning

The In-Fusion HD cloning system (Clontech laboratories) was used to insert DNA from the genes of interest into the chosen expression vector. The In-Fusion system works by using the In-Fusion enzyme, which ligates DNA fragments by recognizing 15 bp overlaps at their ends. The procedure started by designing primers for amplification of the desired sequences into the expression plasmid. PCR was performed as described in section 2.5.2, but with an annealing temperature adjusted to the primers' T_m. An aliquot of the PCR product was analysed by agarose gel electrophoresis to confirm correct amplification, and it was then purified. The chosen expression vector was linearised by PCR to produce ends compatible with the designed primers.

The In-Fusion cloning reaction was performed in PCR tubes by adding the following components:

- 2 µl of 5x in-Fusion HD Enzyme Premix;
- 50-200 ng of the linearised vector;
- 10-200 ng of the purified PCR fragment to insert;
- nuclease-free water to make the volume up to 10 µl.

The reaction was incubated at 50 °C for 15 minutes and placed on ice. 1 µl was used to transform a suitable strain of competent cells.

2.6 Recombinant protein expression

2.6.1 Transformation of chemically competent cells

In order to insert the plasmid containing the gene of interest into the chosen competent cells, 1 µl of the reaction containing the transformed plasmid were added to 25-50 µl of competent cells that had previously been thawed on ice. The mixture was gently mixed by stirring and incubated on ice for 20 minutes. It was heat-shocked at 42 °C for 45 seconds, incubated on ice for two minutes, and 500 µl of pre-warmed LB or SOC medium were added. The competent cells were left to recover for at least one hour (two hours for better yields if selecting the transformants on kanamycin plates) at 37 °C with agitation, and they were then plated on LB agar plates containing the appropriate antibiotic/s. The plates were incubated overnight at 37 °C before proceeding with colony screening (section 2.6.7).

2.6.2 Expression systems

Different expression systems were used for the different attempts of heterologous protein production, which are listed in Table 5.2. In this section, only the systems that were used for successful protein production are described. They include *E. coli* cytoplasmic expression system using the plasmid pET52b(+) and the competent cells Rosetta-gami 2(DE3), *E. coli* periplasmic expression system with the plasmid pETFPP30 and the competent cells Rosetta 2(DE3) and the Baculovirus mediated insect cell expression system with the plasmid pOPINS3C and the insect cell line *Spodoptera frugiperda 9 (Sf9).*

2.6.2.1 *E. coli cytoplasmic expression system*

The expression of the bacterial proteins *Lp*sGH5_8, *Lp*sGH134s and *Lp*sGH134b was performed in the cytoplasm of *E. coli* using the plasmid pET52b(+) (Fig. 2.4). This expression vector is 5,227 base pairs long and contains a StrepTag II coding sequence at the N-terminus, which is cleavable with the human rhinovirus (HRV) 3C protease. It also encodes for a HisTag at the C-terminus and has a thrombin protease cleavage site. It contains a *lacI* promoter, T7 promoter, T7 transcription start and terminator sequences, a *f1* origin of ssDNA replication, an optimized ribosomal binding site (RBS), a MCS site containing multiple restriction enzyme sites and resistance to ampicillin. Rosetta-gami 2(DE3) (Novagen) competent cells were used as expression strains because the protein amino acid sequences were analysed with the GenScript rare codon analysis tool (https://www.genscript.com/tools/rare-codon-analysis) and the DiANNA 1.1 web server (http://clavius.bc.edu/~clotelab/DiANNA/), which highlighted the presence of numerous rare codons and disulphide bonds.

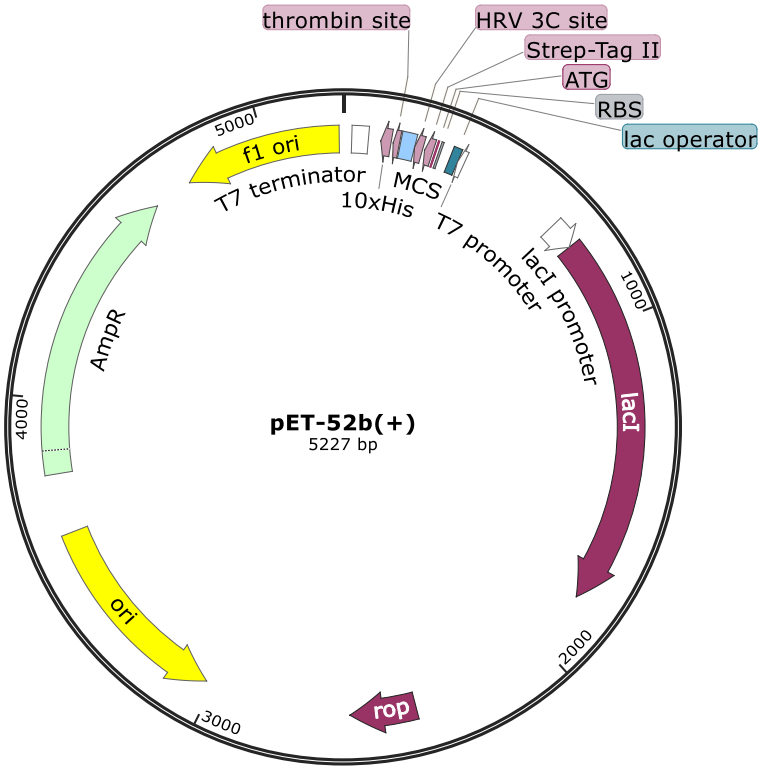


Figure 2.4. Map of the expression plasmid pET52b(+). HRV=human rhinovirus, ATG=start codon, RBS=ribosomal binding site, MCS=multiple cloning site, AmpR=ampicillin resistance.

2.6.2.2 *E. coli periplasmic expression system*

The expression of the bacterial protein *Lps*AA10 was performed in the periplasm of *E. coli* using the plasmid pETFPP30 (Fig. 2.5). This expression vector is 5,520 base pairs long and is specifically designed to guide the protein into the periplasmic space of the bacterial cells by means of the N-terminal pelB secretion sequence under control of the T7 promoter. The oxidising environment of the periplasm improves the protein correct folding by allowing the formation of disulphide bonds (Yoon *et al.*, 2010). pETFPP30 is a modification of the plasmid pET22b with a HRV 3C protease added just before the HisTag at the C-terminus. The other typical features of the pET plasmids are maintained (*lacI* operator, *f1* origin of ssDNA replication, optimized ribosomal binding site (RBS), MCS site containing multiple restriction enzyme sites and resistance to ampicillin). The competent cells used for expression were Rosetta 2(DE3) (Novagen).

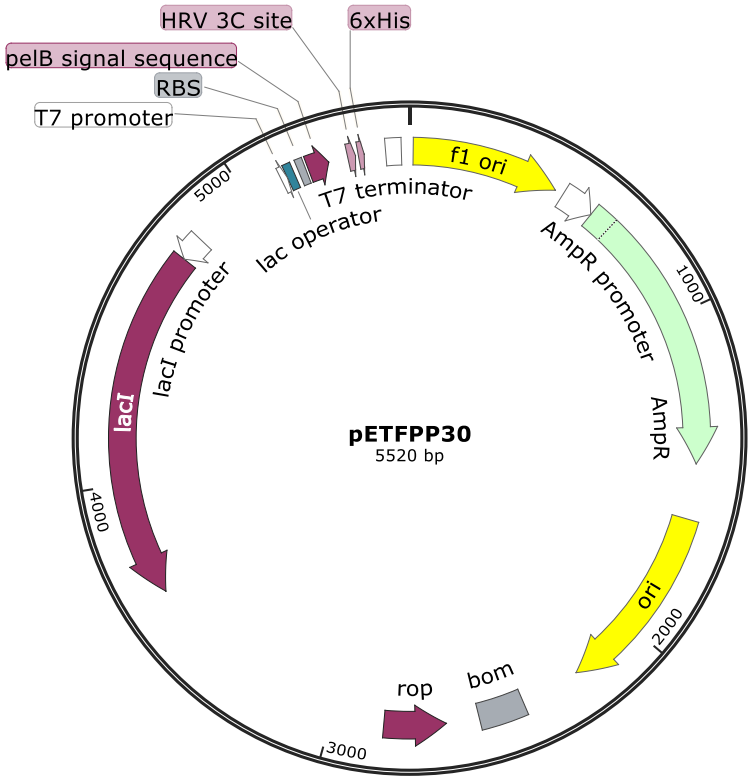


Figure 2.5. Map of the expression plasmid pETFPP30. RBS=ribosomal binding site, MCS=multiple cloning site, AmpR=ampicillin resistance, bom=basis of mobility.

2.6.2.3 Baculovirus mediated insect cell expression system

The expression of the bacterial protein *Lps*GH11 was performed in the cytoplasm of *S. frugiperda* insect cells using the plasmid pOPINS3C (Fig. 2.6). This expression vector was designed and constructed at the Oxford Protein Production Facility UK (OPPF-UK), is derived from the plasmid pUC and is 5,424 base pairs long (Bird, 2011). It includes a HisTag at the N-terminus, followed by a SUMO solubility tag, which can be cleaved with the HRV 3C protease. It also features a lacZ promoter, p10 promoter and bb-actin promoter, T7 transcription start and terminator sequences, CMV (human cytomegalovirus) enhancer, baculovirus recombinant regions and resistance to ampicillin.

Protein expression was performed using the baculovirus mediated insect cell expression system (Berger *et al.*, 2004), as explained in section 2.6.4.

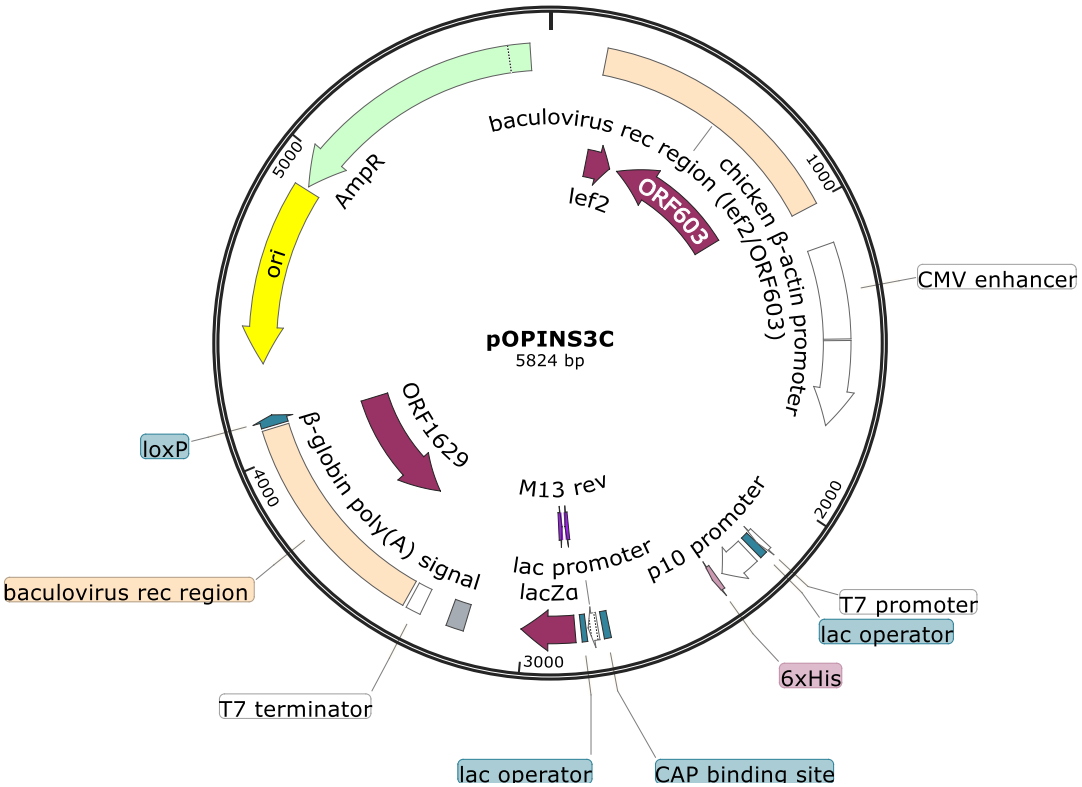


Figure 2.6. Map of the expression plasmid pOPINS3C developed at OPPF-UK. HRV=human rhinovirus, AmpR=ampicillin resistance, baculovirus rec region=baculovirus recombination region, CMV enhancer= human cytomegalovirus enhancer, CAP binding site=catabolite activator protein binding site.

2.6.3 Bacterial cells expression

Bacterial cells containing *Lps*GH5_8, *Lps*GH134a and *Lps*GH134b were grown in 100 ml of LB broth supplemented with carbenicillin (50 µg/ml) and chloramphenicol (34 µg/ml) at 37 °C until OD_{600}=0.7 and then induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) 1mM. They were grown overnight at different temperatures (16, 20, 30 and 37 °C) and 200 rpm to compare expression levels. Once the preferred temperature was identified, cultures of 1 litre were set up and grown overnight (around 18-20 hours) at that temperature. The cells were pelleted by centrifugation at 4,000 g for 15 minutes, resuspended in phosphate buffered saline (PBS) with 0.01 mM 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF). They were lysed by sonication (3 minutes: 3 second on, 7 seconds off), centrifuged at 4,000 g for 15 minutes and the supernatant (soluble fraction) was used for subsequent purification.

Bacterial cells expressing the *Lps*AA10 were grown in 500 ml of M9 Minimal Medium, containing 1% (w/v) glucose, carbenicillin (50 µg/ml) and chloramphenicol (34 µg/ml) at 37 °C until OD_{600}=0.7. The culture was induced with IPTG (0.1 mM) and grown overnight (18-

20 hours) at 20 °C and 200 rpm. The cells were harvested by centrifugation at 4,000 g for 15 minutes at 4 °C, resuspended in 50 ml/litre of 50 mM Tris-HCl with 20% sucrose (pH 8.0) and kept on ice for 30 minutes. After centrifugation at 8000 rpm for 10 minutes the supernatant was discarded, the cells were resuspended in 50 ml/litre of 5 mM $MgSO_4$ and kept on ice for 30 minutes. After centrifugation the supernatant, containing the periplasmic fraction, was equilibrated with 0.2 M Na phosphate buffer pH 7.6 to a final concentration of 50 mM and was used for subsequent purification

2.6.4 Insect cells expression

The expression of *Lp*sGH11 was performed at the Oxford Protein Production Facility UK (OFFP-UK) using insect cells and the baculovirus mediated expression system.

Bacmid preparation

Bacmid, a shuttle vector that can be propagated both in *E. coli* and insect cells, was inoculated from a glycerol stock belonging to OPPF-UK in 100 ml containing kanamycin (50 µg/ml) and chloramphenicol (35 µg/ml) and grown overnight at 37 °C. The BacMax kit (Cambio) was used to extract bacmid DNA according to manufacturer's instructions. The DNA was diluted to 100 ng/µl and distributed in 24 microcentrifuge tubes, where 1x NEB buffer 3 and 1x BSA were also added. The restriction enzyme Bsu361 (New England Biolabs) (1 µl) was added to each tube and it was incubated at 37 °C for two hours, heated at 72 °C for two minutes and stored at -20 °C until used.

Co-transfection and preparation of P1 virus stock

Sf9 cells were adjusted to a density of $5x10^5$ cells/ml, 500 µl were transferred into a 24-well culture plate and they were left to attach to the bottom of the well for one hour at room temperature. In the meanwhile, the transfection mixture was prepared by gently mixing 2.5 µl of bacmid, 100-500 of plasmid DNA and 50 µl of Sf900 II serum free medium (SFM, Invitrogen), pipetting 0.75 µl of FuGeneHD transfection reagent (Roche applied science) into the liquid and incubating at room temperature for 30 minutes to allow liposomes to form. The transfection mixture (25 µl) was added carefully without disturbing the monolayer, and the plates were swirled to distribute the mix equally, and incubated for five days at 27 °C after sealing with parafilm. After this period the supernatant (P0 virus stock) was stored in the dark at 4 °C, while the plate with the cells was stored at -80 °C.

P1 and P2 amplification

In a 24 well culture plate, 0.5 ml of Sf9 cells at the concentration of $1x10^6$ cell/ml were added to each well and they were left to attach to the bottom of the well for one hour at room temperature. 5 µl of P0 virus stock were added to each well, and the plate was incubated for six days at 27 °C. The supernatant was harvested to produce the P1 virus stock, which was kept at 4°C in the dark for immediate use and at -80 °C for long-term storage. The P2 virus stock for small scale expression screening was produced like the P1 stock but using 50 µl of P1 stock and storing the viral supernatant at day seven of the infection at 4 °C in the dark.

Small-scale protein expression screening

A lysis solution was prepared by mixing 20 ml of NPI-10 plus, 100 µl of DNase solution and 5 µl of beta-mercaptoethanol. The 24-well plate with the P0 virus was defrosted for 10 minutes at room temperature and 250 µl of lysis solution were added to each well. The plate was shaken at 500 rpm for 15 minutes at room temperature, the suspension was transferred to a 96-deep well block and 20 µl were pipetted into a PCR pate and mixed with 20 µl SDS loading buffer (1% SDS, 10% glycerol, 0.1% bromophenol blue and 100 mM 2-mercaptoethanol). The PCR plate was centrifuged at 6,500 g for 30 minutes and the supernatant was placed in a flat-bottom 96-well plate containing, in each well, 20 µl Ni-NTA agarose beads. It was shaken for 30 minutes at 600 rpm at room temperature and then placed on a type B Quiagen magnet for five minutes. The liquid was discarded, the plate was removed from the magnet and 150 µl of NPI-20 were added gently in each pipetting 2-3 times well to resuspend the beads. The plate was re-placed on the magnet for two minutes, the liquid discarded, and 100 µl of NPI-20 were added. The procedure was repeated with 60 µl of NPI-250, the plate was shaken for 10 minutes at 500 rpm at room temperature and re-placed on the magnet for three minutes. Finally, 25 µl were transferred to a PCR plate containing in each well 25 µl of SDS loading buffer and blots were electrophoresed for both the whole cells and the soluble product. Once the expression of *Lps*GH11 was confirmed, scale up was performed in 500 ml of Sf9 cells with a virus dilution of 1:1,000.

2.6.5 Sodium Dodecyl Sulphate PolyAcrylamide Gel Electrophoresis (SDS-PAGE)

Gels used for SDS-PAGE were the Mini-Protean TGX Precast gels (Bio-Rad, USA) with a polyacrylamide concentration depending on the analysed protein size or at a gradient from

4 to 20%. The electrophoresis was carried out using a Mini Protean II apparatus (Bio-Rad, USA). The tanks were filled with running buffer (25 mM Tris, 192 mM glycine, 0.1% SDS, pH 8.3). The samples 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 five minutes to denature the proteins. The wells were loaded with 15, 20 or 30 µl of sample (depending on gel and wells size) alongside with 5 µl of pre-stained 1kb HyperLadder (Bioline). The gels were electrophoresed at 200 V until the marker dye-front ran off the gel (around 45 minutes). The staining was performed with Instant*Blue* Coomassie stain (Expedeon) and the de-staining by washing various times with deionised water (dH_20). Pictures of the gels were taken with the Syngene PXi gel documentation imaging system.

2.6.6 Schiff's fuchsin-sulfite glycoprotein stain

The Schiff's fuchsin-sulfite glycoprotein stain was used to detect glycoproteins. The protocol involved performing SDS-PAGE with the samples of interest and then fixing the gel for one hour in 30% (v/v) methanol and 7% (v/v) acetic acid. After three washes for five minutes in ultrapure water, the gel was left for one hour in 15% (v/v) periodic acid in 5% (v/v) acetic acid to oxidise the cis-diol sugar groups in glycoproteins. After a series of washes with ultrapure water, it was placed for one hour in Schiff's fuchsin-sulfite solution (Sigma-Aldrich) in the dark to allow the formation of aldehyde groups, which are detected through the formation of Schiff-base bonds with a reagent that produces magenta bands. The gel was immersed in 0.5% sodium metabisulfite for one hour and subsequently stored in 7% acetic acid. Pictures of the gels were taken with the Syngene PXi gel documentation imaging system.

2.6.7 Western blot analysis

To perform western blot analysis, the proteins under study were subject to SDS-PAGE as described in section 2.6.5. They were blotted into a nitrocellulose membrane using a iBlot 2 dry blotting system (Thermo Fisher Scientific) and iBlot transfer stacks (Thermo Fisher Scientific) according to the manufacturer's instructions, after immersion of the gel in 20% ethanol for five minutes. The successful transfer of the protein bands was visualised by staining with Ponceau solution (0.1% (w/v) Ponceau S in 5% (v/v) acetic acid). After washing with deionised water until clean, the membrane was incubated with blocking buffer (1xBPS (15 mM K_2HPO_4, 80 mM Na_2HPO_4, 140 mM NaCl, pH 7.5) with 5% (w/v) skimmed milk

powder) for one hour with gentle rocking. After blocking, the membrane was washed three times for five minutes and gentle rocking with PBS containing 0.05% Tween 20, to decrease non-specific background staining. It was incubated for two hours at room temperature (or at 4 °C overnight) in a solution of PBS containing 0.05% (v/v) Tween 20, 1% (w/v) BSA supplemented with anti-Strep II antibody horseradish peroxidase (HRP) conjugate (Novagen) or anti-His antibody HRP conjugate (Sigma Aldrich) both at 1:5,000 dilution. After a second washing step (three times for five minutes and gentle rocking with PBS containing 0.05% Tween 20), the membrane was incubated for five minutes in 2 ml of 50% stable peroxidise solution and 50% luminol/enhancer solution using the SuperSignal West Pico PLUS Chemiluminescent Substrate kit (Thermo Scientific), it was briefly dried and the HRP activity was visualised with the Syngene PXi gel documentation imaging system.

2.6.8 Protein quantification (Bradford assay)

Proteins were quantified using the Bradford method described in Bradford (1976). In short, 5 µl of samples and standards (bovine serum albumin (BSA) ranging from 0.03125 to 1.5 mg/ml in pure water) were added to 250 µl of Coomassie Plus Protein Assay Reagent (Thermo Fisher Scientific) in a 96 well optical microtitre plate. Blanks were also prepared containing the protein reagent in the appropriate buffer. After incubation at room temperature for five minutes, the absorbance at 595 nm (OD_{595}) of the samples was read using a Sunrise plate reader (Tecan). A linear standard curve was produced using the BSA standards and it was used to calculate the concentration of the samples.

2.6.9 Protein concentration

The fractions obtained from protein purification 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. The samples were centrifuged until reduced to the desired volume following manufacturer's recommendations.

2.6.10 Media

The media used for recombinant protein expression were LB, SOC, auto-induction and M9 minimal medium.

2.6.10.1 Lysogeny Broth

LB medium (Lysogeny Broth, also improperly called Luria Bertani broth, from its creator Giuseppe Bertani) is a nutritionally rich medium that is routinely used to grow bacteria. It was prepared by dissolving 25 g of LB Broth Miller (Fisher BioReagents) in 1 liter of dH_20, adjust the pH to 7.0 and autoclaving at 121 °C for 15 min. LB-agar, used for making plates, was prepared by dissolving 40 g of LB-agar Miller (Formedium) in 1 liter and autoclaving.

2.6.10.2 Super Optimal broth with Catabolite repression

Super Optimal broth with Catabolite repression (SOC) is a nutrient-rich bacterial growth medium normally used to improve efficiency of transformations. It consists of Super Optimal Broth (SOB) with added glucose. It is prepared by dissolving (in 1 liter of dH_20) 20 g tryptone, 5 g yeast extract, 0.5 g NaCl, 0.186 g KCl, 3.603 g glucose, 0.952 g anhydrous $MgCl_2$ and 1.204 g heptahydrate $MgSO_4$. The pH was then adjusted to 7.0 and it was autoclaved at 121 °C for 15 minutes.

2.6.10.3 Auto-induction medium

Auto-induction medium was at times used as an alternative to LB when expressing proteins, to test expression levels. The recipe is complex and included preparing the ZY medium base by dissolving 10 g tryptone and 5 g yeast extract in 1 litre of dH_20. Salts were prepared by adding 24.65 g $MgSO_4$ heptahydrate to 100 ml of dH_20. 50x5052 (5052=0.5% glycerol, 0.05% glucose, 0.2% α-lactose) were prepared by adding 25 g glycerol, 2.5 g glucose and 10 g α-lactose to 73 ml of dH_20. 20xNPS (NPS=nitrogen, phosphorus and sulfur) were prepared by adding 6.6 g $(NH_4)2SO_4$, 13.6 g KH_2PO_4 and 14.2 g Na_2HPO_4 to 90 ml of dH_20. 1000x metals were prepared by adding to 36 ml of dH_20 50 ml 0.1M $FeCl_3$ hexahydrate (in 01.M HCL), 2 ml 1 M $CaCl_2$, 1 ml 1 M $MnCl_2$ tetrahydrate, 1 ml 1 M $ZnSO_4$ heptahydrate, 1 ml 0.2 M $CoCl_2$ exahydrate, 2 ml 0.1 M $CuCl_2$ exahydrate, 1 ml 0.2 M $NiCl_2$ exahydrate, 2 ml 0.1 M Na_2MoO_4 pentahydrate, 2 ml 0.1 M Na_2SeO_3 pentahydrate and 2 ml 0.1 M H_3BO_3.

The salts and ZY medium were separately autoclaved at 121 °C for 15 minutes and then to 185.4 ml of the ZY base were added 200 µl $MgSO_4$, 200 µl 1000x metals, 4 ml 50x5052, 10 ml 20xNPS and 200 µl of the appropriate antibiotics.

2.6.10.4 M9 minimal medium

M9 minimal medium was used only for the expression of the protein *Lp*sAA10. The medium was prepared by first making up the salts 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. The salts were autoclaved at 121 °C for 15 minutes. The M9 minimal medium was prepared by adding 200 ml salts, 2 ml 1M sterile $MgSO_4$, 20 ml 20% glucose and 100 µl 1 M CaCl2 (sterile) to 700 ml of dH_20.

2.7 Protein purification

2.7.1 Affinity chromatography

*Lp*sGH5_8, *Lp*sGH134a and *Lp*sGH134b were purified by affinity chromatography after cell harvest, sonication and centrifugation (section 2.6.3). After addition of 5 mM $MgCl_2$ and DNaseI (0.025 U/µl, NEB) and filtering through a 0.45 µm filter, the supernatant was run through a 5 ml StrepTrap HP column (GE Healthcare Life Science) using an ÄKTA purifier or ÄKTA start (GE Healthcare Life Science) following manufacturer's instructions. Washes were performed with PBS and elution with 2.5 mM desthiobiotin in PBS. The eluted fractions containing absorbance peaks were analysed by SDS-PAGE to confirm the presence of the recombinant protein, combined and concentrated (quantification done by Bradford assay), and a thermal shift assay (section 2.7.3) was performed to verify correct folding. The strep-tag was removed by adding the HRV Turbo protease (Abnova) at a ratio of 1:100 and leaving the sample overnight at 4°C and gentle shaking. The sample was purified by gel filtration (see next section).

The periplasmic fraction obtained from the expression of *Lp*sAA10 was equilibrated in 0.2 M sodium phosphate buffer pH 7.6 to a final concentration of 50 mM. It was applied to a 5 ml StrepTrap HP column, with the same tools and protocols described above for the other proteins. After purification, the eluted fractions containing absorbance peaks were analysed by SDS-PAGE to confirm the presence of the recombinant protein, combined, concentrated (quantification done by Bradford assay), and a thermal shift assay (section 2.7.3) was performed to verify correct folding. Fivefold excess copper was added as $CuSO_4$ to provide the metal for the protein active centre, since the protein was expressed in minimal medium not containing metals. Then unbound copper and desthiobiotin were removed by gel filtration (see next section).

After insect cell lysis with the lysis buffer I-PER insect cell protein extraction reagent (Thermo Fisher Scientific) according to manufacturer's instructions, the supernatant containing *Lp*sGH11 was equilibrated with wash buffer to a final concentration of 300 mM NaCl and 30 mM imidazole. Nichel affinity purification was performed by adding 5 mM $MgCl_2$ and DNaseI (0.025 U/µl, NEB), filtering through a 0.45 µm filter and passing the supernatant through a HisTrap HP 5 ml column (GE Healthcare Life Science) using an ÄKTA start (GE Healthcare Life Science) and following manufacturer's instructions. Equilibration was performed with wash buffer (20 mM sodium phosphate pH 7.4, 300 mM NaCl, 30 mM imidazole) and elution with elution buffer (20 mM sodium phosphate pH 7.4, 300 mM NaCl, 500 mM imidazole) using a gradient of 30 to 500 mM imidazole. Fractions were collected and analysed by SDS-PAGE. Fractions containing the protein were pooled and concentrated (quantification done by Bradford assay) before being subjected to gel filtration (see next section). After this purification step, the fractions containing the protein were combined and concentrated and the SUMO tag was removed using the HRV Turbo protease (Abnova) at a ratio of 1:100 and incubation overnight at 4 °C and gentle shaking. Purification was carried out manually using a 5 ml HisTrap column (GE Healthcare Life Science) and a gradient of 20-500 mM imidazole. The fractions were electrophoresed by SDS-PAGE to confirm tag cleavage and protein presence, were pooled together, concentrated and subject to thermal shift assay.

2.7.2 Gel filtration

Gel filtration was performed for *Lp*sGH5_8, *Lp*sGH134a and *Lp*sGH134b to clean them from the desthiobiotin and the strep-tag after its removal with the HRV, and for *Lp*sAA10 to eliminate the unbound copper and desthiobiotin. 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 10 mM sodium phosphate buffer pH 7.0 for *Lp*sAA10, and with PBS for the other proteins. The relevant peaks were verified by SDS-PAGE, protein-containing fractions were combined and concentrated (quantification done by Bradford assay) and a thermal shift assay was repeated to confirm that the protein was still correctly folded.

2.7.3 Thermal shift assay

Thermal shift assays were conducted on purified proteins to verify correct folding using a Mx3005P qPCR System (Agilent Technologies) and SYPRO™ Orange Protein Gel Stain (Life Technologies) as a dye. This was diluted 1/100 in dH_20 and 3 µl were added to the sample, which also contained 10 µl of the protein to be analysed and 17 µl of the appropriate buffer. The intensity of the fluorescence was measured against a temperature gradient of 25–95 °C and the obtained values were plotted to determine the melting temperature (T_m). This was done by curve fitting using a five parameter sigmoid equation with the T_m measured as the midpoint at http://paulbond.co.uk/jtsa (Schulz *et al.*, 2013).

2.8 Enzyme activity assays

2.8.1 DNS (3,5-dinitrosalicylic acid) reducing sugars assay

DNS (3,5-dinitrosalicylic acid) reducing sugars assays were used to assess activity of the expressed recombinant proteins and of gills, digestive glands, crystalline style and caecum contents of *L. pedicellatus* on a range of polysaccharides (see section 2.8.5 for the a description of the substrates used). The recombinant proteins were used in their purified form and at the concentration of 0.5 mg/ml. The shipworm organs (5 caeca, 12 digestive glands, 6 gills, 10 crystalline styles) were dissected and gills and digestive glands were homogenised in a non-reducing buffer (EDTA-free protease inhibitors (1% v/v, Thermo Scientific) 50 mM sodium phosphate pH 7.6) to preserve enzymatic activity. The obtained material was sonicated (10 pulses of two seconds) to disrupt the cells and release the contents, it was centrifuged at maximum speed for one minute on a benchtop centrifuge and the supernatant containing the soluble protein was filtered through a 0.22 µm porous membrane and quantified with a Bradford assay. The proteins were diluted at a protein concentration of 0.5 mg/ml for each organ. For the caeca, only the contents were used (not the organ's tissue), which were obtained by cutting opening the caecum with a blade and suspending the contents in the same non-reducing buffers as the other organs; the contents were not sonicated. The crystalline styles were not homogenised but directly solubilised in the buffer by heating at 37 °C for one hour.

The reactions (50 µl) were performed in triplicates and were set up in 2 ml microcentrifuge tubes, containing 2.5 µg of protein extract or recombinant protein, 0.1 % of substrate, 12.5 µl of 50 mM sodium phosphate pH 7.6 (7.8 for the recombinant proteins) and 27.5 µl of pure water. They were incubated at 30 °C for 20 hours (two hours for the recombinant

proteins) with shaking at 320 rpm. After brief centrifugation to pellet the substrate, an aliquot of 9 µl of the supernatant was placed into a PCR tube together with 31 µl of DNS reagent (9.75 g of dinitrosalicylic acid, 1.4 g of NaOH, 21.6 g of sodium potassium tartrate tetrahydrate, 0.53 ml of phenol and 0.59 g of sodium metabisulfite in 100 ml of pure water), and it was heated at 100 °C for five minutes. After cooling at room temperature, 160 µl of pure water were added to the tube and 180 µl of the sample were used to measure the absorbance at 540 nm in a 96 well optical microtitre plate using a Sunrise plate reader (Tecan). Glucose standards were analyse together with the samples and were used to produce a standard curve to calculate the nanomoles of reducing sugars released.

2.8.2 Heat maps

To produce heat maps of the expressed recombinant proteins, a citrate-phosphate buffer (CPB) system was used. To make the CPB, two solutions were prepared: solution A (0.1 M citric acid) and solution B (0.2 M disodium hydrogen phosphate). Solution A and B were mixed according to table 2.7 to obtain buffers at different pHs.

Table 2.7 Table used to calculate the amount of solution A and B to use to make up the citrate-phosphate buffer at different pHs.

pH	Solution A (ml)	Solution B (ml)	Final volume (ml)
4	30.8	19.3	50
4.5	27.3	22.7	50
5.0	24.3	25.8	50
5.5	21.6	28.5	50
6.0	18.4	31.8	50
6.5	14.5	35.5	50
7.0	8.9	41.2	50
7.5	3.9	46.1	50
8.0	1.4	48.6	50

Reactions were performed in triplicates in 96-wells PCR plates, at 24 different temperatures in the range from 30 to 74 °C and using a Tetrad 2 Peltier thermal cycler (BIO-RAD) with a gradient mode. The reactions (100 µl) were placed in each well, containing 50 µl of 1% (w/v) glucomannan substrate, 10 µl of enzyme and 40 µl of buffer. Enzyme concentration was 10 µg/ml for *Lp*sGH5_8 and *Lp*sGH134a, and 500 µg/ml for *Lp*sGH134b, while *Lp*sGH11 was not tested due to lack of protein. Blank reactions not containing the enzymes were also performed. The plates were incubated in the thermal cycler for two hours, cooled to room

temperature, 100 µl of DNS reagent (see section 2.8.1 for the recipe) were added and the plate was heated at 100 °C for five minutes. The absorbance at 545 nm was measured in a 96 well optical microtitre plate using a Sunrise plate reader (Tecan). The raw data from the microplate reader was transferred into an Excel sheet and blank values were subtracted from the actual reactions. Values below zero were set to zero and mean values among the three plates were determined for each well. The data was transformed into relative activity by setting the highest value from the average to 100%. Phyton 3.6 was used for the analysis, with the module SciPy (Jones *et al.*, 2014) to interpolate the data in and Matplotlib (Hunter, 2007) to plot the map.

2.8.3 Polysaccharide analysis using carbohydrate gel electrophoresis (PACE)

PACE analysis for *Lp*sGH5_8, *Lp*sGH134a and *Lp*sGH134b was performed by mixing the purified enzyme at 20 µg/ml with 0.5% galactan, glucomannan, galactomannan, mannan or locust bean gum (LBG) or with 40 mg/ml of milled Scots pine wood (pre-treated in 0.5 N NaOH for 30 minutes at 90 °C and rinsed five times in 50 mM $NaPO_4$ buffer) in 50 mM $NaPO_4$ buffer pH 6.5 and incubated overnight at 30 °C with shaking. The samples were centrifuged, the supernatant was transferred to a new tube and the undigested polysaccharides were removed by precipitation with 80% ethanol. Following centrifugation, supernatants were transferred to a new tube and dried.

For *Lp*sGH11, *Miscanthus* stem AIR (alcohol insoluble residues) was pre-treated in 4 M NaOH for one hour at room temperature and neutralized with HCl. The resultant substrate at 1 mg/ml (of initial untreated AIR) was digested overnight at room temperature with various amounts of xylanase (3- 40 µg/ml).

All samples were purified on Nanosep 10k and dried. Dried digestion products and manno-oligosaccharide and xylo-oligosaccharide standards and appropriate controls were labelled with 8-aminonaphthalene-1,3,6-trisulfonic acid (ANTS, Invitrogen) and separated by polyacrylamide gels, as described by Bromley and colleagues (2013). PACE gels were visualized using a G-box (Syngene). Experiments were carried out in triplicates.

2.8.4 Product identification by MALDI/TOF-TOF tandem mass spectrometry

In order to analyse the digestion products of *Lp*sAA10 on different substrates, reactions were carried out on microcrystalline cellulose (Avicel), phosphoric acid swollen cellulose (PASC), squid chitin, shrimp chitin, colloidal chitin, glucomannan, barley β-glucan and beech

wood xylan (Sigma). The 50 µl reactions contained 50 mM ammonium acetate buffer pH 6.0, 4 mM ascorbic acid or gallic acid in 0.2 M ammomium acetate buffer pH 6.0, 0.4% (w/v) substrate, 3 µM of LpsAA10 and pure water to make up the final volume. The reactions (and the appropriate blanks) were incubated overnight at 30 °C with shaking at 300 rpm. The samples were centrifuged and 1 µl of the supernatant was mixed with an equal volume of 20 mg/ml of 2,5-dihydroxybenzoic acid (DHB) in 50% acetonitrile and 0.1% TFA on a SCOUT-MTP 384 target plate (Bruker). The spotted samples were dried in a vacuum desiccator and permethylation was carried out according to Ciucanu and Kerek (1984). Spotted samples were analyzed by MALDI/TOF-TOF tandem mass spectrometry using 2,5-DHB matrix with 0.1% TFA on an AB-Sciex 4700 and Ultraflex III MALDI/TOF-TOF instrument (Bruker). Data were collected using a 2-kHz smartbeam-II laser and acquired on reflector mode (mass range 300–3000 Da) for MS analysis and on LIFT-CID for MS/MS analysis using argon as collision gas. FlexControl and FlexAnalysis softwares were used for data acquisition and analysis. On average, about 10,000 shots were used to obtain high-enough resolution. MS/MS fragmentation patterns were named according to Domon and Costello (1988).

2.8.5 Substrates

The substrates that were used for the DNS assay or for PACE are: barley β-glucan (β-D-1,3-1,4-glucan), mannan (borohydride reduced), konjac glucomannan (β-D-1,4), galactomannan (CAROB), larch arabinogalactan, wheat arabinoxylan, tamarind seed xyloglucan, potato galactan, galactan (Gal:Ara:Rha:Xyl:GalUA=91:2:1.7:0.3:5) all purchased from Megazyme; LBG, carboxymethyl-cellulose (CMC) and beech wood xylan, all purchased from Sigma-Aldrich. PASC was prepared by moistening 5 g of Avicel PH-101 (Sigma-Aldrich) in pure water and treating with 150 ml of ice cold 85% (v/v) phosphoric acid on an ice bath for one hour. Cold acetone (500 ml) was added while stirring and the swollen cellulose was filtered on a glass-filter funnel, washed three times with 100 ml ice cold acetone and twice with 500 ml of water. It was then suspended in 500 ml of pure water and blended to homogeneity.

Chapter 3: Anatomy and physiology of the digestive organs of *Lyrodus pedicellatus*

3.1 Introduction

Shipworms have been known for a long time, as they have been documented in the works of Pliny, Ovid, Aristophanes and even Homer, mainly because of their devastating effect on wooden boats (Nair and Saraswathy, 1971). The first proper scientific report, prompted by the shipworms' fascinating ability to live in timber, was produced by Sellius, who in 1733 described the species *Teredo navalis*. Since then, numerous studies have approached different aspects of Teredinids biology, and the fact that they are indeed able to feed on wood – not just live in it - is now widely accepted. Despite the multitude of publications, there is limited information on the mechanisms of wood digestion by these animals. One of the main subjects of discussion is the symbiosis with the endosymbiotic bacteria found in the bacteriocytes in the Glands of Deshayes. The bacteria location (the gills and not the digestive system as in other wood-eating animals) and importance related to wood digestion (production of fixed nitrogen, lignocellulolytic enzymes, antibiotics and other metabolites) is clear; however, the mechanism by which the digestive CAZymes produced by the bacteria are selected among other secreted bacterial proteins and are able to travel to the digestive system, and eventually end up in the distant caecum for wood digestion are still open to debate. In this thesis, the food groove is proposed as a route by which the enzymes could be transported to the caecum. The food groove (described in section 1.3.5.4) is a stream of mucus propelled by the movement of cilia, which transports food particles from the gills surface to the animal's mouth and the rest of the digestive system (Gosling, 2015). The food groove is routinely used by bivalve molluscs for transporting to the mouth the organic material that has been captured by filter feeding (Orton, 1912; Morton, 1970). It is the only known connection between the gills, a respiratory organ, and the digestive system, and is the most likely route for the bacterial enzymes to enter the caecum.

The caecum, despite being nowadays recognised as not just the site for wood storage, but as a key organ for lignocellulose digestion and absorption (Betcher *et al.*, 2012; O'Connor *et al.*, 2014; Shipway, 2013), is not well characterised. Indeed, the biochemical evidence of its ability to digest wood is dated, imprecise, or uncomplete. Furthermore, the role of caecum tissue in the production of lignocellulolytic enzymes or other elements important

for the digestion process (such as sugar transporters dedicated to move glucose from the caecum lumen to the circulatory system or antibiotics aimed at reducing the bacterial population) has not been elucidated. Finally, there is a need for scanning electron micrographs to confirm the presence in the caecum lumen of wood fragments, and of the microvillar border and the typhlosole, which are structures that assist in wood digestion.

The shipworm digestive glands are also of interest, because of their dual nature (production of enzymes and lignocellulose digestion) and because of the presence of the phagocytes (amoebic cells able to incorporate and digest wood fragments). These have previously only been evidenced by drawings, and require high-resolution imaging to confirm their existence and features in *L. pedicellatus* or other shipworm species.

Finally, possibly the most intriguing but still rather mysterious organ of the shipworm is the crystalline style. This gelatinous rod, known to aid extracellular digestion in bivalves, has never been investigated in shipworms, in term of its composition and whether it plays a role in wood digestion.

3.2 Aims of the chapter

This chapter aims to explore the anatomy and the physiology of wood digestion in *L. pedicellatus*, to identify the main organs involved in the process and the underlying mechanisms. The work presented here builds on previously published work by further investigating unexplored areas and relevant aspects with novel techniques that will allow to draw a more precise picture of how wood is ingested, processed and digested by shipworms.

The chapter will focus on the following aspects:

- confirm the presence of bacteriocytes in the gills;
- investigate the hypothesis that the food groove is the organ through which the bacterial CAZymes are transported from the gills to the caecum;
- assess which component/s of the wood (cellulose, lignocellulose or lignin) are utilised for nutrition;
- study the enzymatic properties of the gills, digestive glands and caecum content;
- investigate the absorptive structures of the caecum (typhlosole and microvilli);

- confirm the dual structure of the gland (digestive/excretory) and the presence of phagocytes in the digestive portion and their role in intracellular wood digestion;
- examine the structure of the crystalline style and confirm its involvement in extracellular wood digestion;
- investigate the structure of the crystalline style sac;
- explore the enzymatic properties of the crystalline style.

3.3 Gills and food groove

3.3.1 Investigation of gills and food groove

3.3.1.1 Microscopy and in vivo assays of the gills and food groove

The presence of bacteriocytes in the Teredinids gills is well established. However, at the start of this work it was necessary to confirm their presence in the gills of *L. pedicellatus*, to study their distribution along the hemi-ctenidia and to determine the path by which wood-degrading enzymes move from the gills to the caecum. Adult shipworms gills were dissected and longitudinal sections were stained with toluidine blue and observed by visible light microscopy, as shown in Figure 3.1.

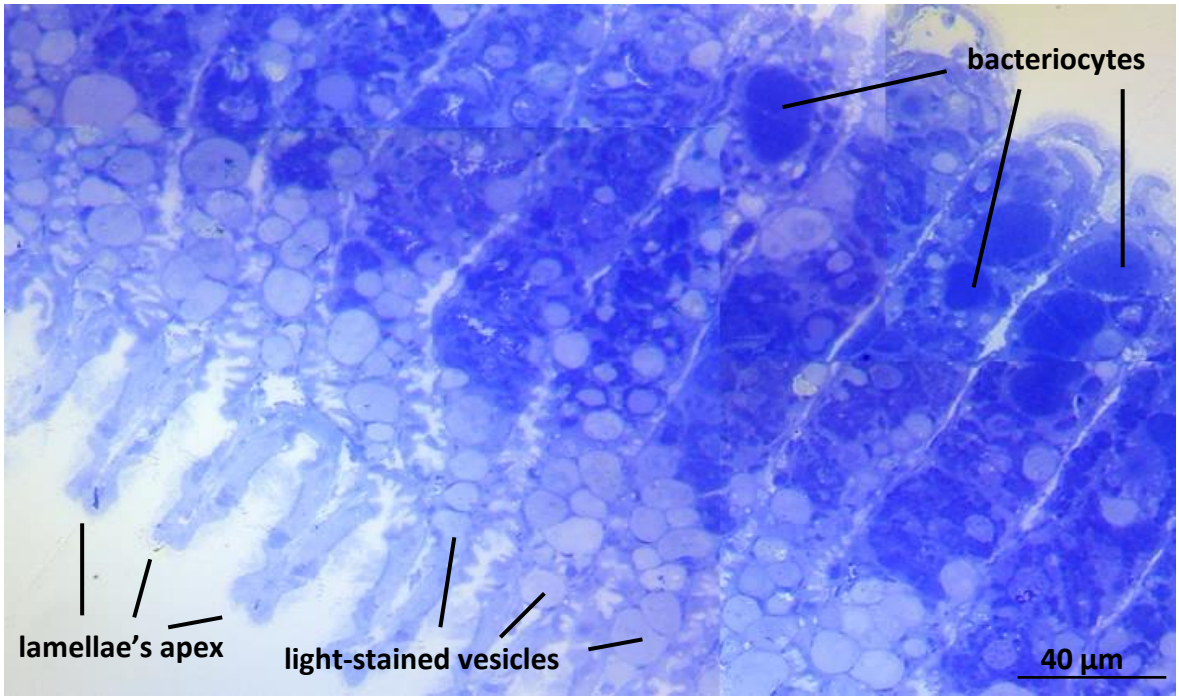


Figure 3.1. *Lyrodus pedicellatus* bacteriocytes location. Light microscopy (100x magnification) of a toluidine blue stained 0.5 µm longitudinal section of paraffin-embedded gill tissue. The bacteriocytes are visible along the single lamellae and white-stained vesicles mainly towards their apex. Picture by Clare Steele-King.

Bacteriocytes of different sizes (10-40 µm) are visible along the gill lamellae, but are not seen towards the gill apex or near the ventral area, where the food groove is formed. Structures similar to bacteriocytes in size and shape, which show lighter staining, are found along the lamellae, and appear to accumulate towards the gill apex. This area near the lamellae apex is normally filled with bacteriocytes, as can be seen in Figure 1.7. This suggests that the lightly stained vesicles could also be bacteriocytes, which stain in a different colour due to a different chemical composition.

Figure 3.1 provides visualisation of the distribution of bacteriocytes along the gill lamellae; however, it does not add information on their movements within the gills and the digestive system. To enable this, shipworms were partially dissected (removing the mantle to expose the main organs) and incubated in a solution containing chromogenic substrates used for the histochemical determination of the presence of active glycosyl hydrolases. In the presence of an active enzyme, the chromogenic substrate – in this case x-cellobiose (5-bromo-4-chloro-3-indolyl β-D-cellobioside) – is oxidised into an insoluble blue colour by the action of enzymes with cellulase activity (Horwitz *et al.*, 1964).

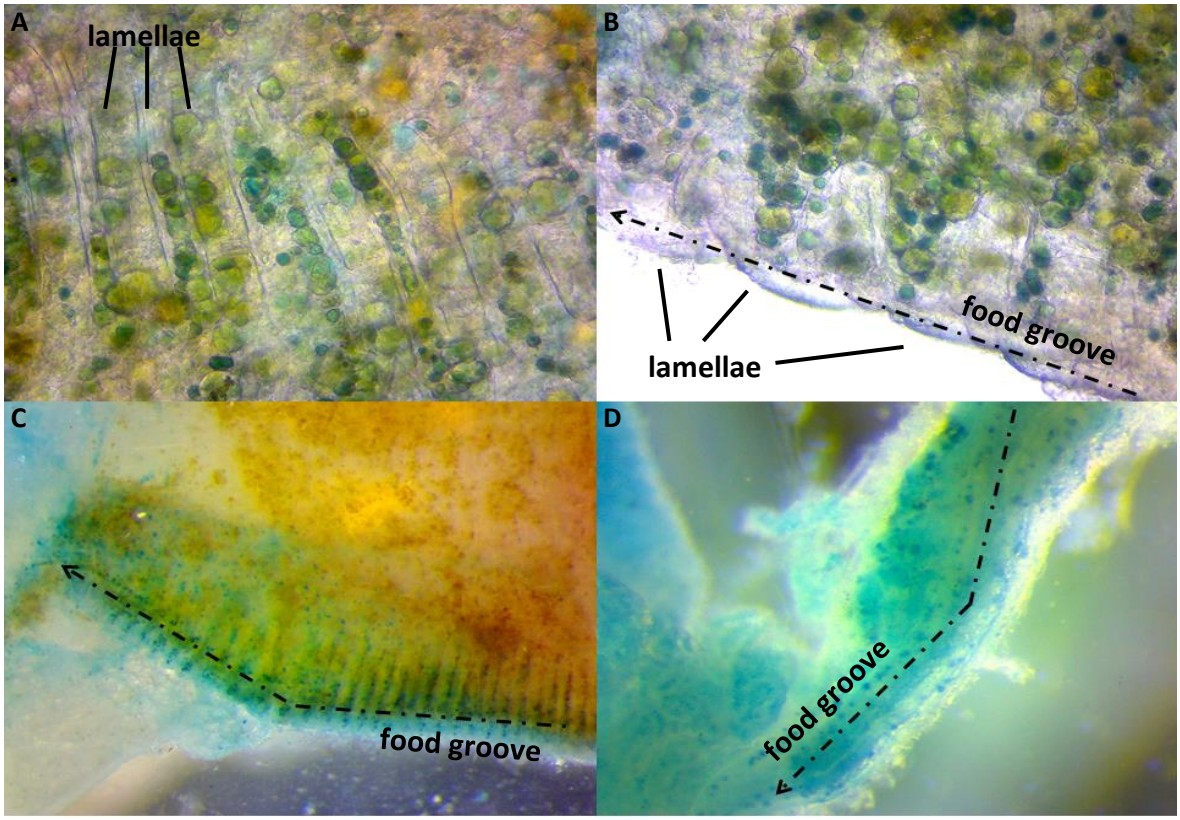


Figure 3.2. Gills and food groove stained with x-cellobiose. Sections of the gill lamellae (**A**) and their apical area (**B**) showing the presence of green-blue vesicles along the lamellae and towards the food groove. **C.** Visualisation of part of the gills showing the vesicles within the lamellae and in the food groove, a line of blue dots along the gills. **D.** Picture focusing on the food groove, showing the stream of vesicles moving towards the mouth. Magnification 20x.

The tissues were examined under visible light microscopy (Figure 3.2) and yellow-green-blue vesicles were observed in the gills lamellae, characterised by being of the same size (10-40 μm in diameter) and shape as the bacteriocytes seen with the toluidine blue staining, and covering the same area. These coloured vesicles seemed to accumulate towards the apex of the lamellae and were seen along the food groove, both proximally and distally to the gills, and close to the mouth area, therefore confirming the presence of cellobiohydrolase activity in these regions. These results suggest that the vesicles that are found in the gills filaments and contain the bacterial wood-degrading enzymes move to the apex of the lamellae to join the food groove, and are then translocated to the shipworm mouth by the action of the cilia. The movement of the vesicles appear to follow a similar route to food particles that have been captured by filter feeding, as reported in other bivalve mollusc (see section 1.3.5.4 for details).

To gain a better understanding of the nature of these vesicles, and confirm the hypothesis that they are indeed bacteriocytes, gills and food groove were observed by transmission electron microscopy (TEM). In order to understand whether these vesicles travel as far as the mouth, an area of the food groove close to the gills (proximal) and one near the mouth (distal) were sectioned. The results are shown in Figure 3.3 for the gills and 3.4 for the food groove.

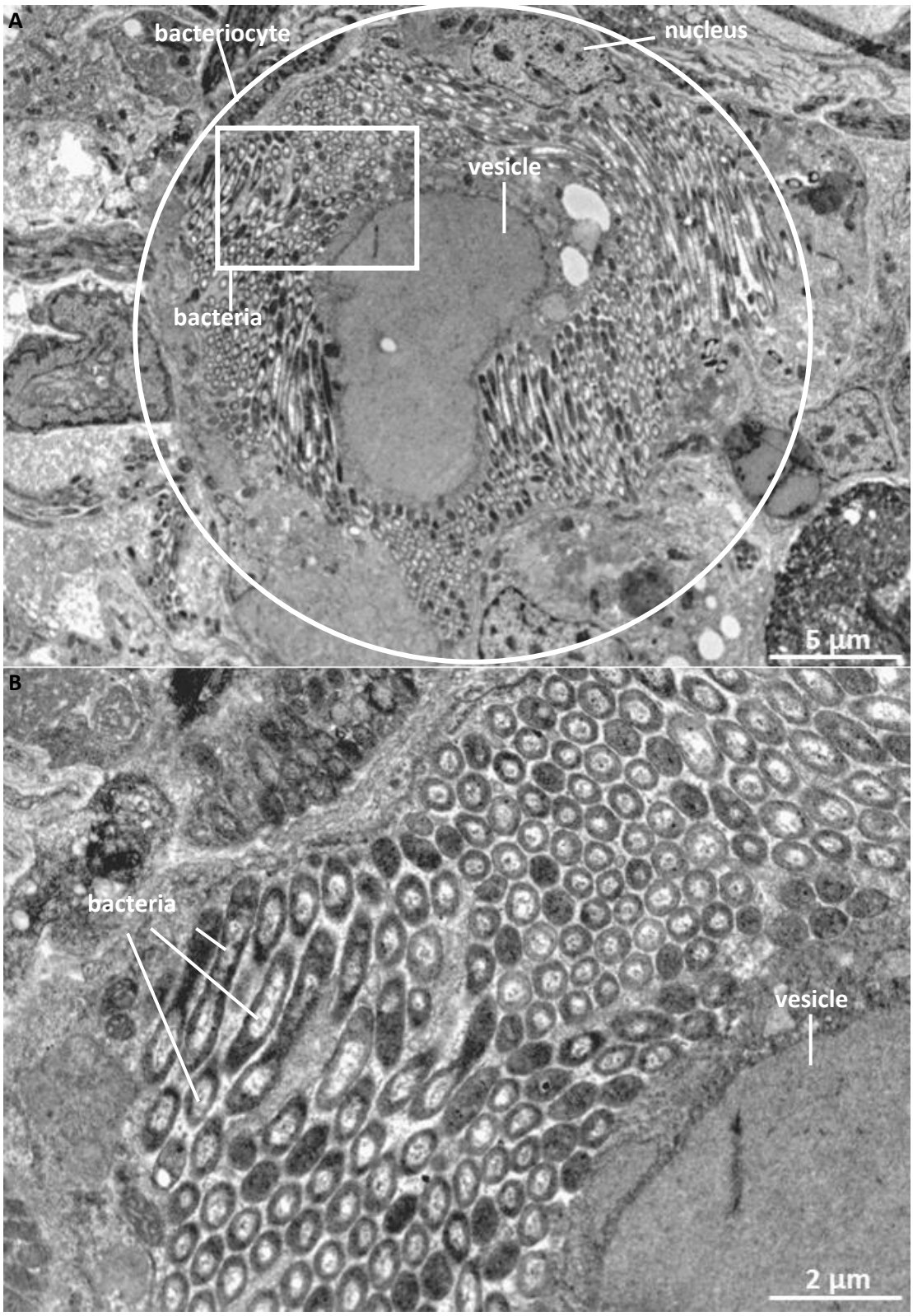


Figure 3.3. TEM of a section of gill lamina. **A.** Bacteriocyte containing numerous endosymbiotic bacteria and a big vesicle of unknown nature. **B.** Magnification of the area circled in the top figure to give a better view of the bacteria. Images by Meg Stark.

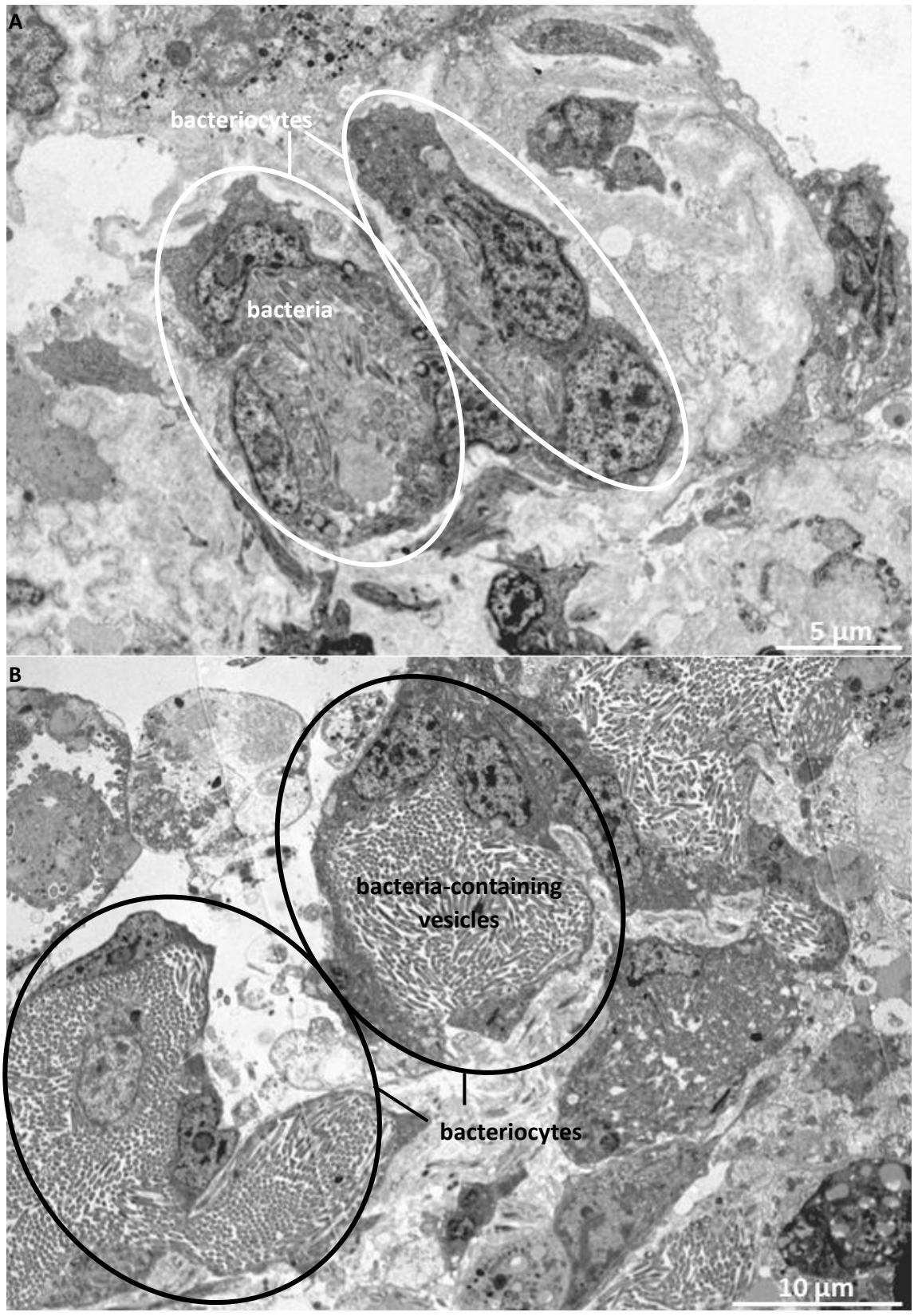


Figure 3.4. TEM of a section of the food groove. Bacteriocytes containing several endosymbiotic bacteria are clearly visible both in the proximal (**A**) and distal (**B**) food groove. Images by Meg Stark.

The TEM pictures of the gills confirm the presence of bacteriocytes packed with endosymbiotic bacteria in the area of the glands of Deshayes. Similar bacteriocytes are also

seen in the food groove, both in the proximal and distal regions, indicating that they are sloughed off the gills laminae and transported along the food groove in a mucus stream to the to the animal's mouth, together with the CAZymes produced by the bacteria. The size, shape and path of coloured vesicles containing cellobiohydrolase observed in the in-vitro experiment with x-cellobiose is comparable to that of the bacteriocytes, indicating that they could indeed be symbionts-containing vesicles.

3.3.1.2 Immunogold labelling of the gills and food groove

The results presented in the previous sections show that vesicles containing active glycosyl hydrolases are found in the gill lamellae, at their apex and along the food groove all the way toward its distal section, suggesting that the food groove could indeed be the organ through which these enzymes are moved towards the digestive system. Intact bacteriocytes are also seen following the same path. However, the experiments did not manage to determine whether the activity (cellobiohydrolase) was due to an enzyme produced by the bacteria or by the shipworm, nor whether the enzymes are found within the bacteriocytes. Furthermore, these observations do not confirm whether the structures observed by TEM are the same as the vesicles that contain cellobiohydrolase activity. Cellobiohydrolases found in the caecum of *L. pedicellatus* are mainly transcribed by the bacteria in the gills and belong to the CAZymes classes GH5 and GH6, as will be explained in chapter 4, which presents the results of the transcriptomic analysis. In order to assess whether the vesicles containing bacteria in the food groove are associated with glycosyl hydrolases that might be responsible for the oxidation of the x-cellobiose, immunogold labelling was performed using an antibody raised against a recombinant bacterial GH5 (details of its characteristic and expression are found in section 5.4), following the methods presented in chapter 2.2.4. Sections of the shipworm gills, food groove and caecum were used for the experiment, and the results are shown in Figure 3.5.

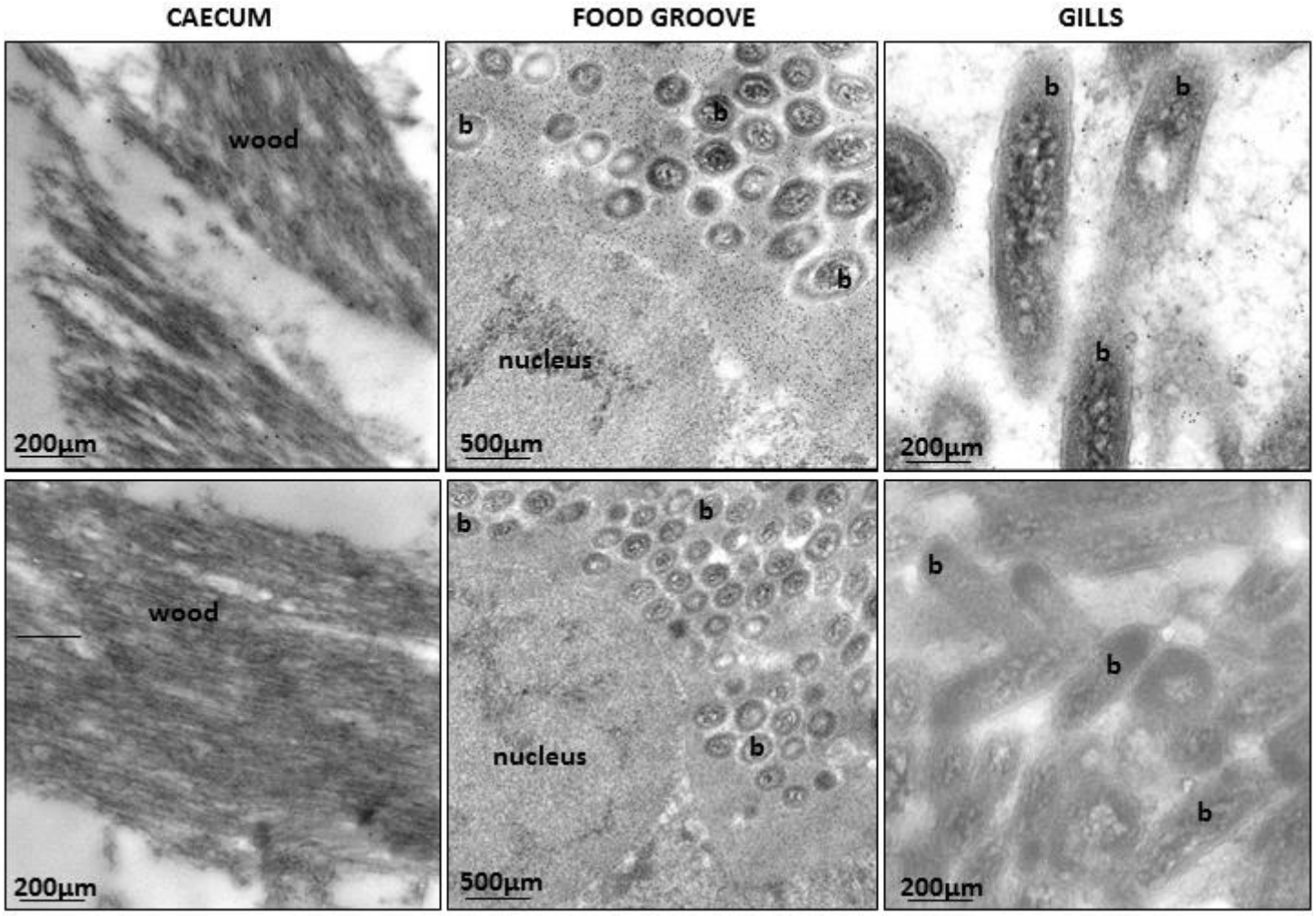


Figure 3.5. Immunogold labelling of the gills, food groove and caecum. Staining (small black dots) representing the location of a bacterial GH5 (the recombinant *Lps*GH5_8 of chapter 5) is seen in proximity of the bacteria in the gills and food groove and over and near the wood fragments in the cecum. The top pictures represent the actual experiment, while the bottom pictures are the negative controls performed with pre-immune serum to show the absence of gold particles. b=bacteria. Images by Clare Steele-King.

Specific labelling against the bacterial GH5 was observed inside bacteriocytes visualised in gills and food groove sections, where gold staining was seen associated to the bacteria only, indicating the presence of the secreted protein. In the caecum section, labelling was observed on and around the wood fragments. Negative controls using pre-immune serum showed no labelling in the same tissues, indicating high specificity for the antibody. These experiments indicate that the food groove is the most probable route by which bacterial enzymes are transported from the gills to the caecum in *L. pedicellatus*. The enzyme appears to be secreted by the bacterial cells, since it was only seen in proximity of the bacteria, but it was not possible to determine whether it was also found outside the endosymbionts containing vesicles or outside the bacteriocytes. Indeed, the tissues embedding stage – which was attempted many times - was extremely difficult due to the presence of large amount of mucus, and therefore the sections obtained were of poor quality, containing many holes, not allowing a clear visualisation of the fine details of the tissues.

3.3.1.3 In vitro activity assay of the gills

The ability of the bacteria to produce lignocellulolytic enzymes was first shown by Waterbury in 1983, when he was able to isolate *Teredinibacter turnerae* – one of the main shipworm endosymbionts – from various shipworms species, including *L. pedicellatus*, and demonstrated that it can digest cellulose. Further progress was made when other authors succeeded in purifying and characterising an extracellular endoglucanase secreted from bacteria cultured from the shipworm's Gland of Deshayes (Greene *et al.*, 1988) and recombinantly expressed a multifunctional cellulase from *T. turnerae* (Ekborg *et al.*, 2007). In 2007 the complete genome of *T. turnerae* was published (Yang *et al.*, 2009) giving access to the full range of CAZymes encoded by this bacterium, which number more than 100. The sequenced genome also showed how *T. turnerae* is well adapted to the digestion of terrestrial wood material, in contrasts to its close relative *Saccharophagus degradans*, which have enzymes to digest also plant, fungal and algal polysaccharides. Finally, O'Connor in 2014 sequenced the metagenome of endosymbionts cultivated from the shipworm *B. setacea*, and identified numerous CAZymes modules and associated CBMs, consistent with their role in the degradation of plant cell wall polysaccharides. However, the above publications might show only part of the CAZymes that are actually produced by the endosymbionts, as the authors relied solely on those bacteria that could be cultivated

from the shipworm gills, which represent only 10% of the whole community (Luyten et al, 2006). Indeed, *T. turnerae* is so far the only shipworm endosymbiont whose different strains can be successfully and reproducibly cultivated. The rest of the consortium, which is thought to be composed of closely related but distinct bacteria (Distel *et al.*, 2002a), have not been cultured or studied so far.

With the view of further investigating the ability of the endosymbiotic bacteria to digest different types of substrates, *in vitro* assays with a range of polysaccharide substrates were performed, with the release of reducing sugars detected with 3,5-dinitrosalicylic acid (DNS), which records the amount of reducing sugars produced (Miller, 1959). The gills of six adult *L. pedicellatus* were dissected and homogenised in a non-reducing buffer containing protease inhibitors, to preserve enzymatic activity. The obtained material was sonicated to disrupt the cells and release the contents, and after centrifugation the supernatant was measured for protein levels (Bradford, 1976). The DNS reducing sugars assay was performed using thirteen different substrates, which included pachyman, phosphoric acid swollen cellulose (PASC), carboxy methyl cellulose (CMC), xylan, xyloglucan, arabinoxylan, arabinogalactan, mannan, glucomannan, locust bean gum (LBG), galactan, lichenan and avicel. The results are shown in Figure 3.6.

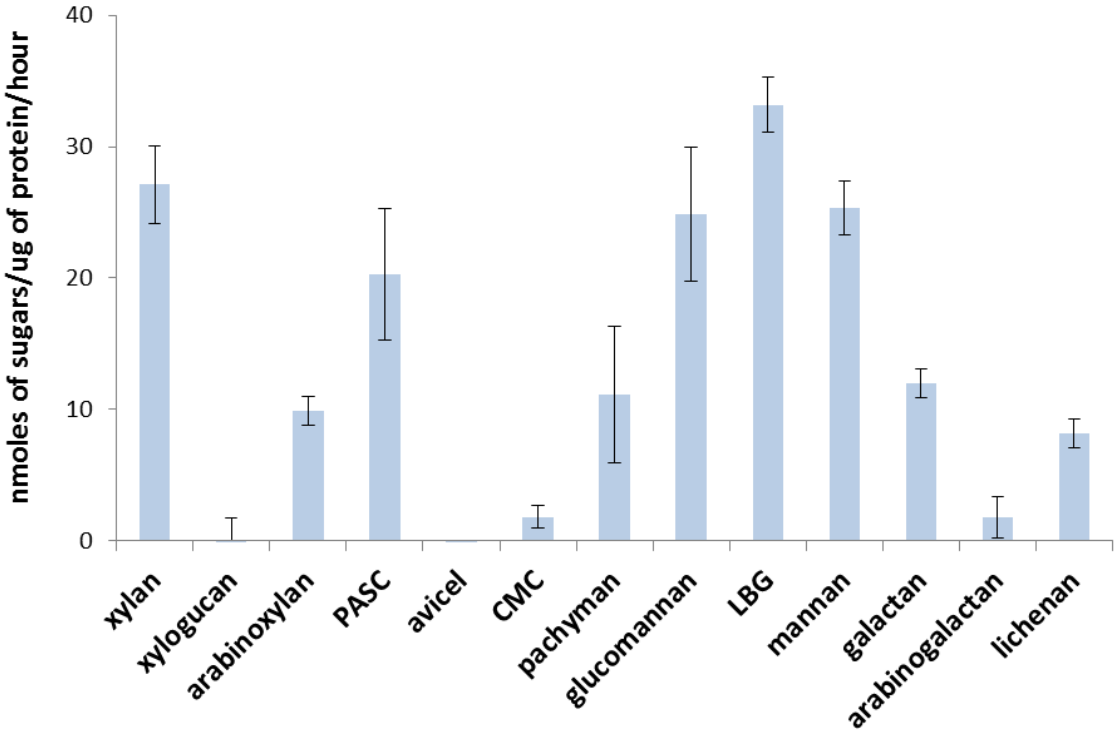


Figure 3.6. Enzymatic activity of gill extracts on a range of polysaccharide substrates. The 50 µl reactions contained 2.5 µg of gill soluble protein, substrates at a concentration of 0.1%, 50 mM sodium phosphate buffer pH 7.6 and were incubated for 20 hours.

These assays show that the gills contain a range of CAZyme activities on a variety of different polysaccharides. The substrates that produced the highest amount of reducing sugars are LBG, xylan, mannan and glucomannan, indicating that the gills enzymes are particularly suited for the digestion of glucans, mannans and xylans. Activity on PASC shows that there is substantial cellulase activity, although it is low against the more crystalline avicel substrate. These results support the known lignocellulose degrading ability of the gills endosymbiotic bacteria, which are adapted for the degradation of woody material. Little activity was recorded on xyloglucan or CMC, indicating little endoglucanase activity, suggesting that the cellulases present are likely to be processive cellobiohydrolases. There was little activity on arabinogalactan and lichenan.

3.4 Caecum

3.4.1. The wood-degrading properties of the caecum

The published biochemical evidence that the caecum has a role in wood digestion is rather sparse, circumstantial and dated. One of the first authors who investigated this problem in a chemical rather than biological way was Harington (1921), who dissected and homogenised what he called “liver” of *Teredo norvegica*, mixed it with various substrates and observed the presence of reducing substances every 24 hours, concluding that an enzyme of some nature was able to reduce sawdust into glucose. However, it is not clear whether the “liver” he used for his experiments was the digestive gland or the caecum, giving therefore no further information on the role of this organ. Another experiment to study the digestive abilities of the same shipworm species was performed a few years later by comparing the chemical composition of the shipworm faeces (called frass) to that of the original wood that the animals bored into, with the view of comparing qualitatively and quantitatively the changes that occur to the wood by passing through the digestive system (Dore and Miller, 1923). The results indicated that shipworms utilise 80% of the cellulose found in wood and part of the hemicellulose (from 15 to 56%); however, the work was not able to point out which organ was responsible for the digestion of those components of the woody material.

Miller and Boyton (1926) performed a reducing sugar assay with the Benedict-Ostenberg method (Thomas and Dutcher, 1924) on the caecum and the original wood in which the shipworms of the species *Bankia setacea* were grown, and showed a four-fold increase in

the reducing sugars content, indicating wood digestion activity in the caecum. It is not clear though whether the whole caecum was used or rather just the contents and the way the organs were dissected and prepared.

Reducing sugar assays on cellulose following the Park and Johnson method (Park and Johnson, 1949) were also used with *Teredo navalis*, whose internal organs were dissected and divided in a pre-caecal portion - anterior to the caecum but excluding the shells and muscles - and a post-caecal one - caecum, intestine, gonads and shreds of the mantle - (Greenfield and Lane, 1953). The results showed that the post-caecal portion contained twice the activity seen in the pre-caecal one, suggesting that either the cellulolytic enzymes are produced there, or they are secreted in the anterior portion and then concentrated in the posterior one. However, as the author themselves pointed out, the organs were not tested separately; therefore, it is difficult from this work to understand precisely where the cellulolytic activity was localised.

Liu and Townsley (1968) used labelled glucose to follow the sugar degradation in the caecum of the shipworm species *Bankia setacea*. They found high concentration of free glucose (19.4 mg/g of caecum dry weight), soluble proteins and the presence of cellulase and other enzymes, as well as various metabolic intermediates resulting from the degradation of glucose, indicating that the caecum might be the place were nutrient uptake and absorption take place. Later work reported the presence of 23 different enzymes associated with the digestion and synthesis of glucose, including a cellobiase, all of which were thought to be endogenous and not of bacterial origin (Liu and Walden, 1970).

There is a wide gap in the literature from the experiments just described and recent research aimed at confirming the role of the caecum in lignocellulose digestion. Betcher and colleagues in 2012 used FISH (Fluorescence *In Situ* Hybridisation) and laser scanning confocal microscopy with 16S rRNA directed oligonucleotide probe to show that the caecum of five different species of shipworms (including *L. pedicellatus*) hosts very few bacteria, which therefore probably contribute very little to lignocellulose degradation in this organ. It was later confirmed, using genome-enabled proteomic methods (O'Connor *et al.*, 2014), that the bacterial enzymes found in the caecum come from the endosymbionts found in the gills, and that up to 98% of the prokaryotic proteins found in the caecum are categorised as “plant cell wall polysaccharides active enzymes”, thus demonstrating the involvement of both the bacteria and the caecum in wood digestion. The same authors

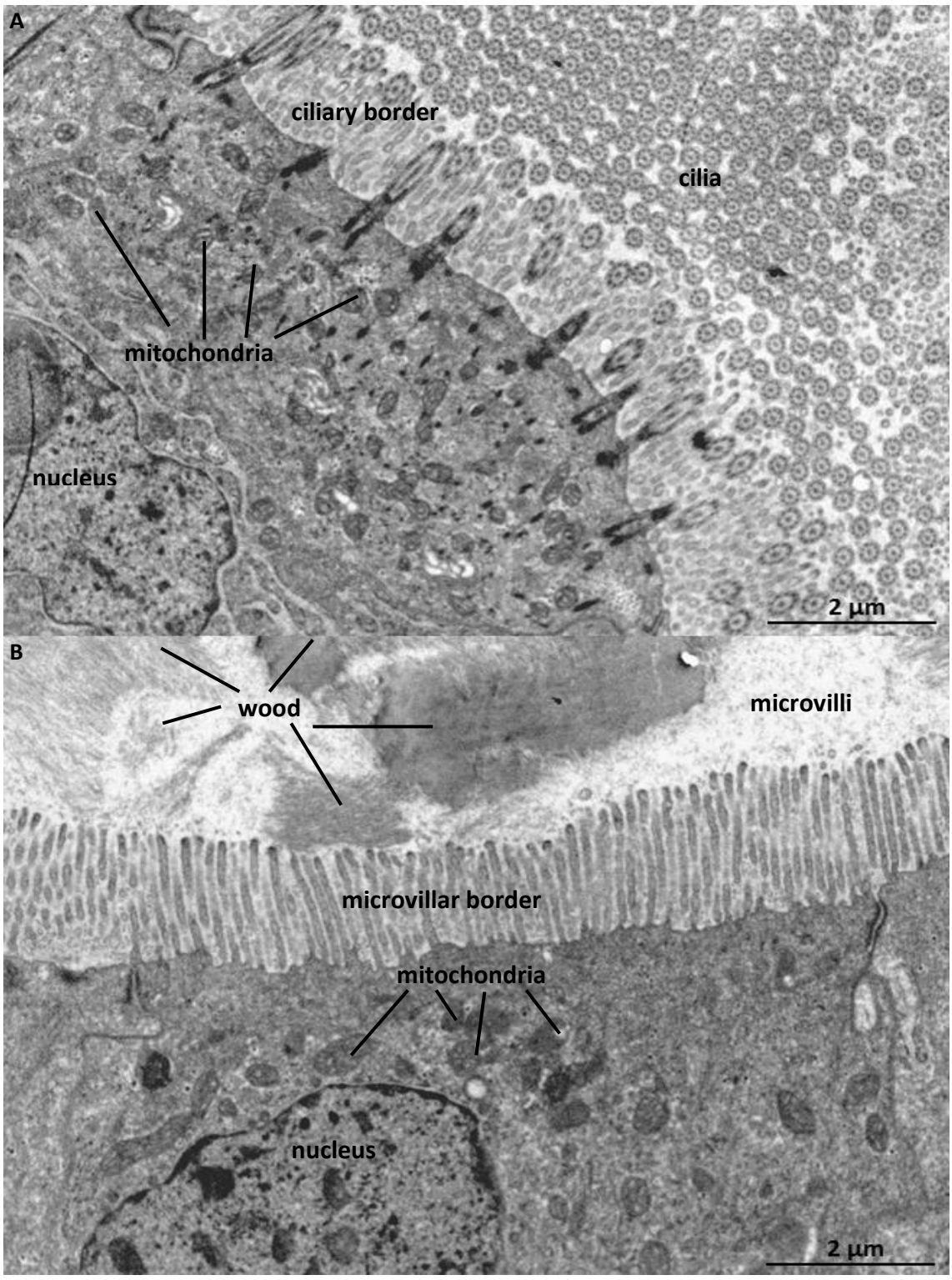
performed ultrafiltration of the caecum extract, and fractionated the proteins, managing to isolate some lignocellulolytic proteins, though they focused solely only bacterial proteins. The presence of shipworm encoded wood-degrading enzymes in the caecum of *L. pedicellatus* was also investigated using RNA-sequencing methods (Shipway, 2013), showing that there was only a limited role of the caecum tissue in lignocellulolytic enzymes production.

3.4.2 Investigation of the lignocellulolytic activities of the caecum

In order to close the gaps in our current knowledge and describe in detail the precise role of the caecum in the digestion of wood in the shipworm *L. pedicellatus*, a series of experiments were performed, which employed different biological and chemical techniques ranging from compositional analysis of wood versus frass, to transmission electron microscopy (TEM) of the caecum tissues, enzymatic assays and protein identification of SDS-PAGE gel bands of the caecum contents and liquid chromatography of the digestion products of a mixture of wood and caecum extracts. In addition, the protein composition of the caecum content was analysed by trypsinolysis followed by liquid chromatography and tandem mass spectrometry (LC-MS/MS), and the level of transcription of CAZymes in the caecum analysed by sequencing the RNA isolated in its tissues. The results obtained from these last two techniques will be presented in the chapter 4, which is dedicated to the proteomic and transcriptomic analysis of the shipworm digestive organs.

3.4.2.1 Transmission electron microscopy of the caecum

Our initial approach was to analyse the caecum tissues by TEM in order to understand in detail its ultrastructure, which in turn may provide support for its functional role. The caeca of three adult *L. pedicellatus* were freshly dissected and were embedded and sectioned for TEM analysis (details of the materials and methods are in section 2.2.3). The most representative pictures obtained are shown in Figure 3.7.


A
ciliary border
cilia
mitochondria
nucleus
2 µm
B
wood
microvilli
microvillar border
mitochondria
nucleus
2 µm

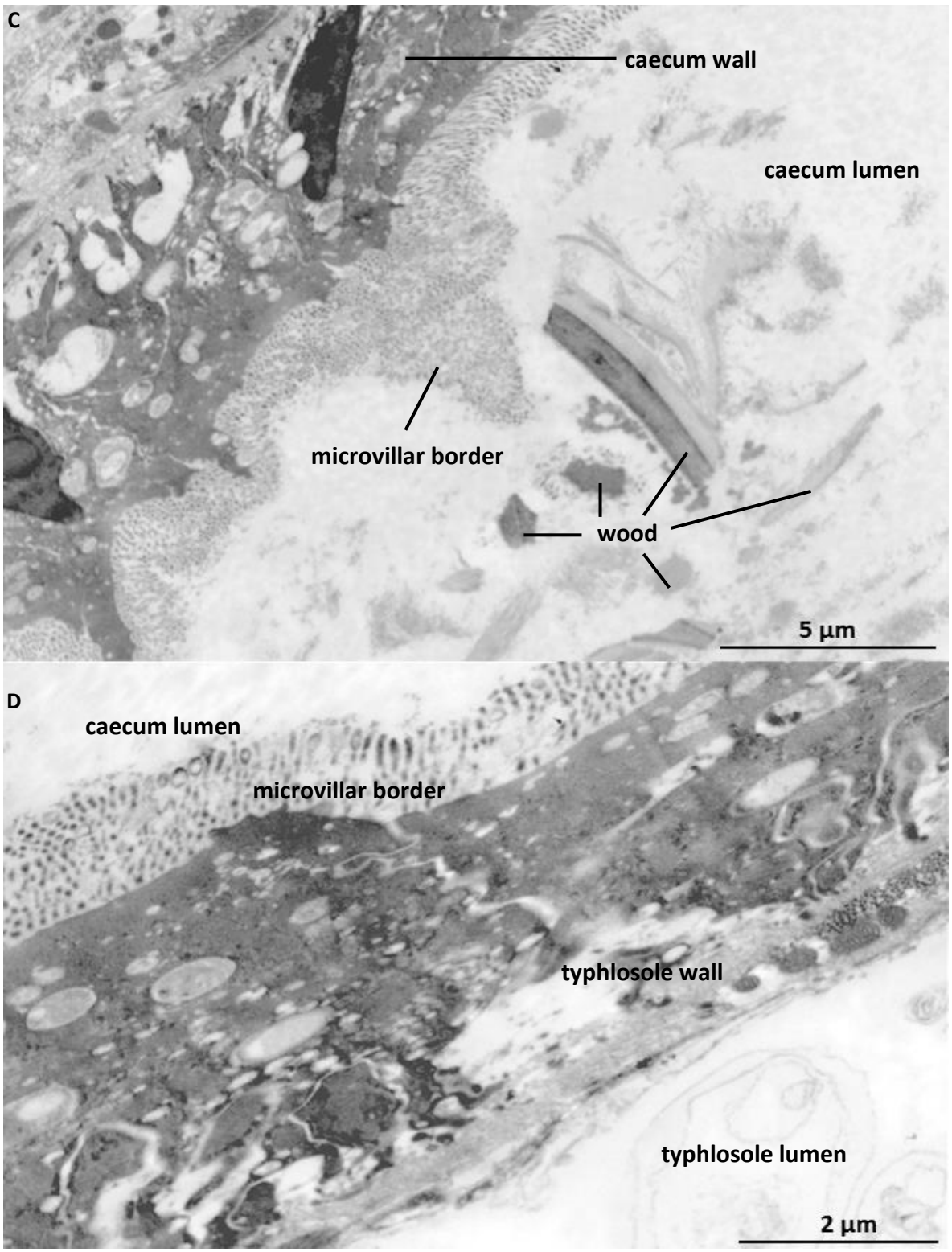


Figure 3.7 Caecum and typhlosole in *L. pedicellatus* visualised by TEM. **A.** Section of the caecum showing numerous mitochondria and large amount of cilia projecting from its epithelium. **B.** Section through the microvillar epithelium, with wood on top of the microvilli. **C.** Different section of the caecum showing its lumen on the right side of the picture, where fragments of woods are clearly visible. The caecum wall is on the left side, with a thick layer of microvilli covering it. **D.** Section of the caecum typhlosole, showing the caecum lumen on the left top and the typhlosole lumen on the bottom right. The typhlosole wall has a microvillar border only on the caecum lumen side. Images A and B from Anna Simona, C and D from Clare Steele-King.

These micrographs are the first TEM images of the caecum of *L. pedicellatus*. They show very clearly the presence and extent of the microvillar and ciliary borders along the caecum walls, with details of their structure and information about their thickness and area coverage. Previously only SEM pictures were available, obtained from other shipworm species. Picture 3.7B and C show the presence of wood fragments in the caecum lumen and on top of the microvilli; this is not a new finding, since wood has been reported in the caecum since Sigerfoos's description of it in 1908, however photographic evidence had not been available until now. The picture of the typhlosole with the layer of microvilli on the caecum lumen side is useful in confirming the role of this structure in the absorption of nutrients.

3.4.2.2 Compositional analysis of wood versus frass

In order to understand which components of the wood are utilised by *L. pedicellatus* for nutrition, and what type of changes occur in the wood while passing through its digestive system, the compositional analysis of wood versus frass performed by Dore and Miller in 1923 was repeated. Indeed, this information is very relevant when collecting information about the caecum, as we now know that it is the main organ involved in the degradation of the wood and in the absorption of sugars, therefore it can be assumed that the changes in the wood composition are mainly due to the caecum. The compositional analysis was performed using more recent protocols compared to the ones used 1923 (see materials and methods in chapter 2 for details), allowing a more detailed view of lignocellulose composition, analysing the different components of the hemicellulosic fraction and quantifying ash and silica. The wood that was used for the analysis was softwood from Scots pines, which was the same on which the shipworms used for this thesis were grown. The frass was collected by regularly filtering the water in the tank on filter paper and drying it at 30 °C until all the moisture was lost. Only shipworms – no other woodborers - were grown on the wood. The results obtained are shown in Figure 3.8 and 3.9.

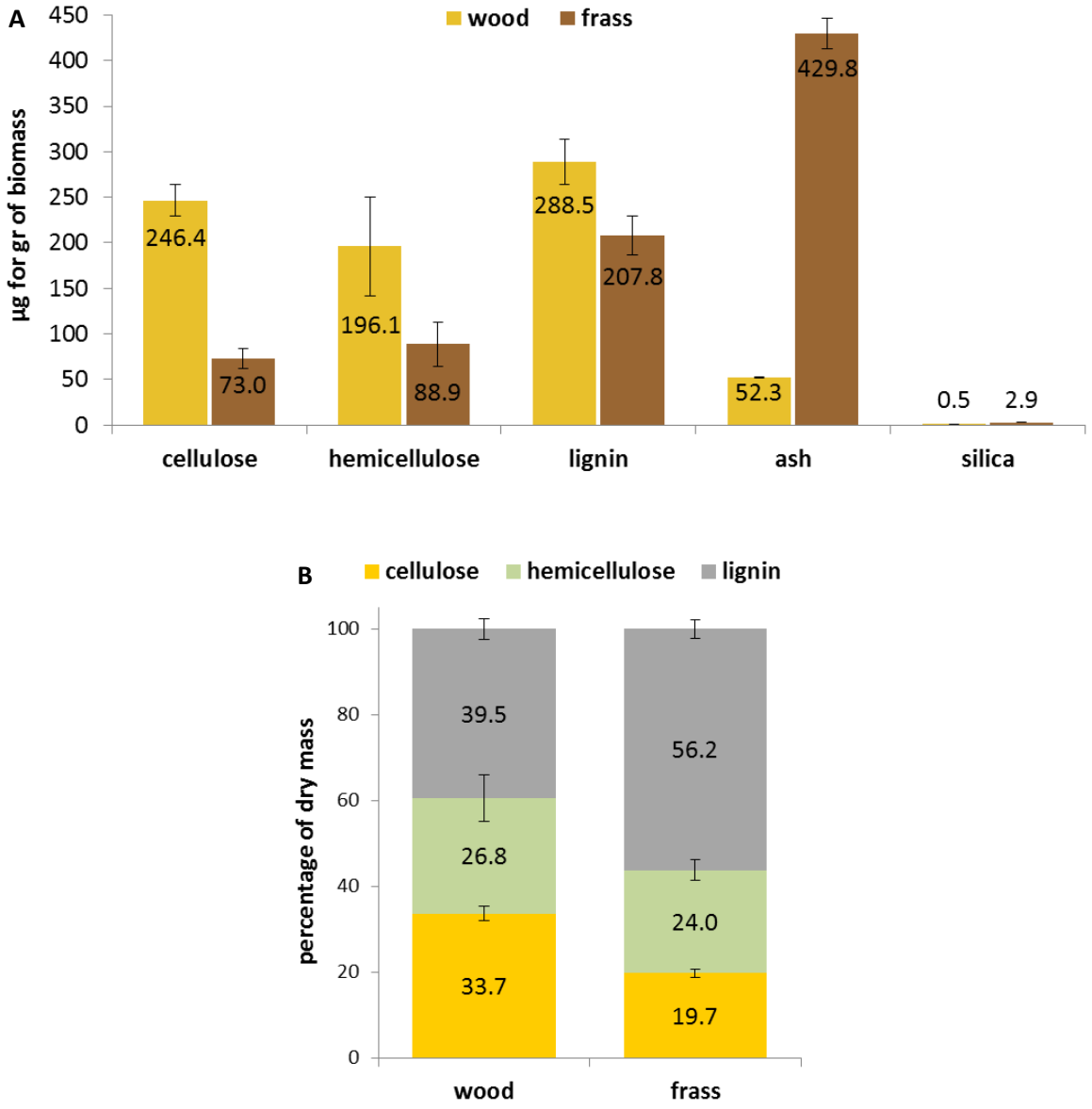


Figure 3.8. Compositional analysis of wood versus frass in *L. pedicellatus*. **A.** The graph shows the amounts (as µg/g of biomass) of the different fractions of wood (cellulose, hemicellulose, lignin, ash and silica) before (wood) and after (frass) passing through the shipworm's digestive system, giving a picture of which fractions have mainly been digested. **B.** Re-analysis of data from A considering only cellulose, hemicellulose and lignin, presented as percentage of the total of the three fractions. The wood analysed was Scots pine. The error bars indicate the standard deviation of the five replicates.

The results of the compositional analysis are similar but more detailed to what Dore and Miller published in 1923. Their analysis showed that 80% of the cellulose was digested by the shipworm, which is a very high amount and indicates that Teredinids are almost as efficient as higher termites in digesting the cellulose they ingest (Sun *et al.*, 2013). Our analysis showed (Figure 3.8A) that *L. pedicellatus* is able to utilise 70.4% of the cellulose that is available in the wood they ingest. The amount is lower but still surprisingly similar to that obtained 95 years ago. The hemicellulose fraction is also degraded by *L. pedicellatus*,

with almost 54.7% of it being lost after digestion, a similar but more precise result from Dore and Mille, who could only approximate the amount of hemicellulose lost to a range from 15 to 56%. Lignin also appears to decrease by about 28% during the digestion process. There is a large increase in ash content in the frass, but it is not clear what the origin of this is, and a compositional analysis of this would be beneficial in the future. In order to focus on the lignocellulosic fraction, the data were reanalysed, this time only considering the lignocellulosic fraction of wood and therefore excluding ash and silica, and calculating the percentage of each fraction among the total. The results, presented in Figure 3.8B, highlight how cellulose is indeed the component most utilised by the shipworm, that hemicellulose is only marginally digested and that lignin instead accumulates in the frass sample and therefore appears not to be utilised by the shipworm (Oevering *et al.*, 2003), indicating that the other component of wood are preferentially degraded. The data though should take into account of the amount of mass that is lost after the digestion of wood by the shipworm, and should be normalised to that value; unfortunately, no mass loss experiment are available for any species of shipworm at present, therefore it was not possible to take this into account.

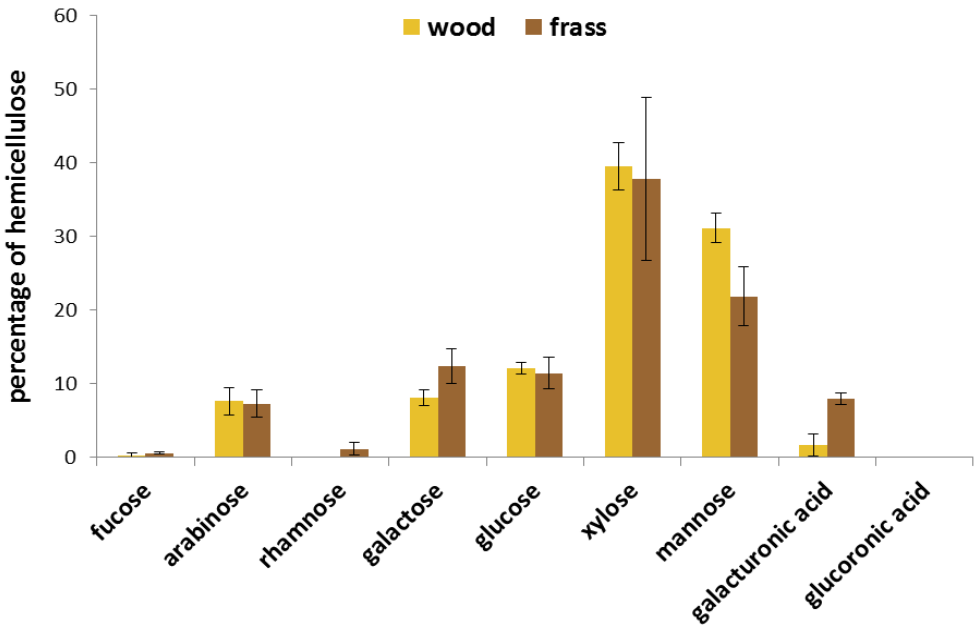


Figure 3.9. Compositional analysis of the hemicellulosic fraction of wood versus frass in *L. pedicellatus*. The wood analysed was Scots pine. The error bars indicate the standard deviation of the five replicates.

Since the hemicellulose fraction seems to be at least partially digested by the shipworm, it was examined in detail, comparing the original wood with the frass. The results, presented

in Figure 3.9, show how mannose is the main sugar in hemicellulose targeted by *L. pedicellatus,* decreasing of 29.9% percent in the frass (paired t-test, p=0.0093). Xylose decreases of 4.3% compared to the original wood, but the result is not statistically significant (paired t-test, p=0.3660), possibly due to the variability of the original data for the frass, hence the significant standard deviation. Glucose and arabinose seem to reduce only very slightly in the frass, with the results not being statistically significant (paired t-test, p=0.2222 and 0.3333, respectively). Other sugars, such as galactose and galacturonic acid, accumulate in the frass (paired t-test, p=0.0024 and 0.0002, respectively) and, therefore, are probably not utilised for nutrition by the shipworm.

3.4.2.3 In vitro activity assay of the caecum fluids

The results above give an overview of *L. pedicellatus'* ability to obtain sugars from lignocellulose, as well as details of which ones are utilised. To be able to efficiently break down wood into single sugars the shipworms must be endowed with a wide range of lignocellulolytic enzymes in the caecum. The ability of these enzymes to act on different types of substrates was investigated previously by some authors (Harington, 1921; Miller and Boynton, 1926; Greenfield and Lane, 1953), which however were only able to report an increase in the presence of reducing sugars in the caecum or posterior part of the digestive system, but not in identifying the organ where the activity is located or to quantify the release of reducing sugars and compare it among different substrates. To further investigate the lignocellulolytic activity of the enzymes of the caecum of the shipworms, the caeca of ten adult *L. pedicellatus* specimens were dissected and the caecum fluids were extracted in a non-reducing buffer, to preserve enzymatic activity (details of the materials and methods are in section 2.8.1). After centrifugation, the presence and abundance of proteins in the combined supernatant was assessed with the Bradford method, and the soluble proteins were tested for cellulolytic activity with the dinitrosalicylic acid (DNS) colorimetric method. Thirteen different polysaccharides were chosen for the assay, in order to have a range of different substrates, including pachyman, PASC, CMC, xylan, xyloglucan, arabinoxylan, arabinogalactan, mannan, glucomannan, LBG, galactan, lichenan and avicel. The results are shown in Figure 3.10.

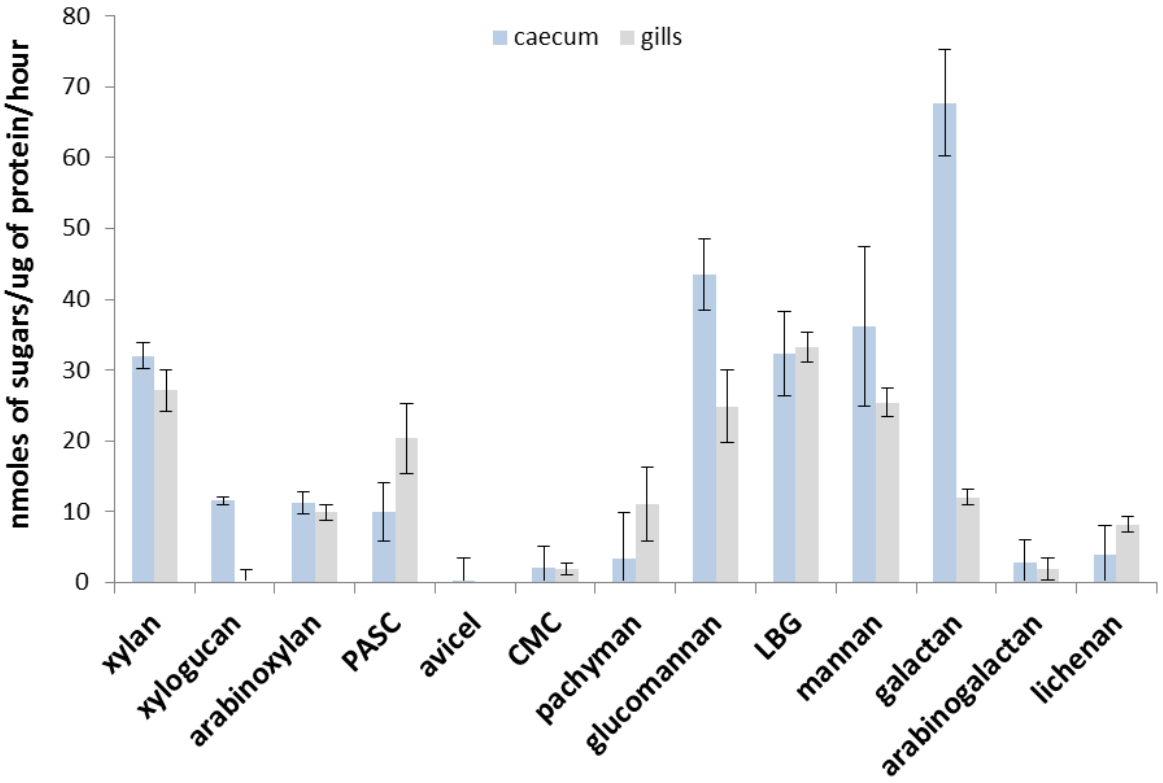
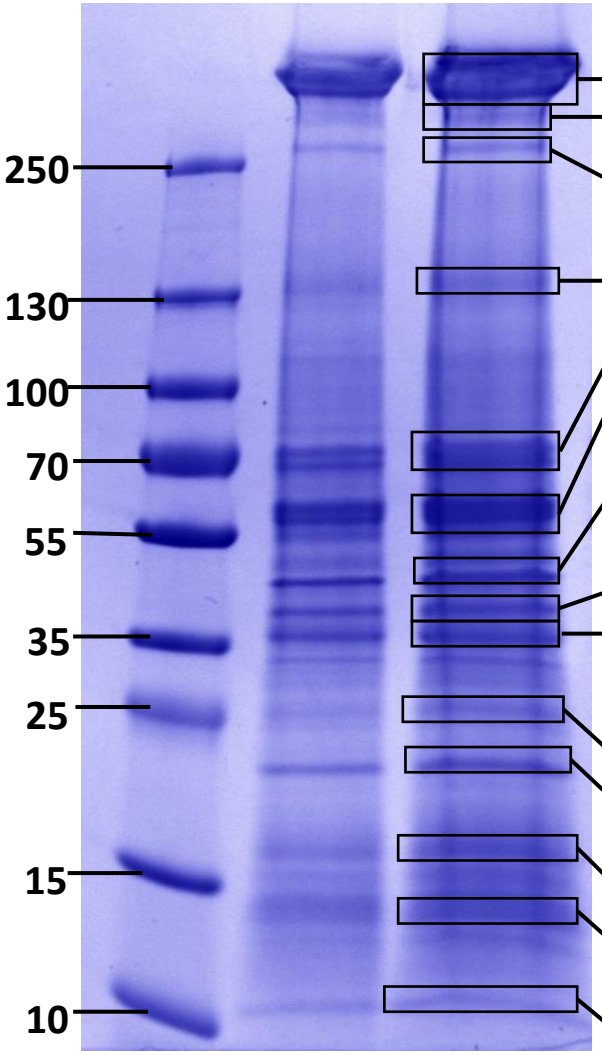


Figure 3.10. Enzymatic activity of soluble caecum fluids on a range of polysaccharide substrates. The 50 µl reactions contained 2.5 µg of digestive caecum fluid protein, substrates at the concentration of 0.1%, 50 mM sodium phosphate buffer pH 7.6 and were incubated for 20 hours. The results for the caecum are shown in light blue, while the ones in grey are from the gills and have been added to the graph for comparison.

The *in vitro* assays show that the soluble proteins found in the caecum of *L. pedicellatus* have activity against a range of polysaccharides associated with lignocellulosic biomass. Galactan, mannan, glucomannan, LBG and xylan give the highest amount of reducing sugars release, indicating that the shipworm enzymes are very well suited for digesting glucans, mannans and xylans, which are the most abundant polysaccharides in gymnosperm wood such as Scots pine, the substrate on which they were feeding (Fengel and Wegener, 1989; Moreira and Filho, 2008). These results support the data obtained from the compositional analysis, indicating that mannan and xylan are the hemicellulose components mostly digested in the caecum. The enzymes found in the caecum do not seem to be able to act on crystalline cellulose, as incubating on Avicel did not produce any reducing sugars, and the results are similar for CMC, pachyman, lichenan and arabinogalactan. The caecum seems to have the ability to degrade a wider panel of substrates compared to the gills (results presented in section 3.3.1.4), indicating that the lignocellulolytic capability of the endosymbiotic bacteria does not account for the full range of degradation seen in the caecum fluids.

3.4.2.4 Caecum protein identification

The results obtained show the wood degrading ability of *L. pedicellatus* and indicate which substrates it is more adapted to digest. However, details on the nature and relative abundance of the enzymes found in the shipworm caecum and responsible for lignocellulose breakdown are sparse. Previous publications have shed some light on the nature of these enzymes, highlighting the presence of glycosyl hydrolases and LPMOs (O'Connor *et al.*, 2014; Shipway, 2013), but these works were skewed either towards the bacterial or shipworm enzymes and did not include a proper estimation of the abundance of the identified proteins in the main organ involved in wood digestion, the caecum. In order to try to identify and measure the abundance of CAZy proteins in the shipworm caecum, the caeca of five animals were dissected and the fluids were extracted. 1% SDS, 10% glycerol, 0.1% bromophenol blue and 100 mM 2-mercaptoethanol were added, the sample was boiled for 10 minutes at 100 °C and centrifuged, and the supernatant was subjected to SDS-polyacrylamide gel electrophoresis in order to separate the caecum total proteins on the basis of molecular mass (Figure 3.11). The most intense and sharp bands obtained were excised from the gel and subjected to trypsinolysis and MALDI-TOF/TOF tandem mass spectroscopy, and the obtained spectra were matched to the EST nucleotide database obtained from sequencing the RNA extracted from the main digestive organs of *L. pedicellatus*: caecum, digestive glands and gills (details on the transcriptomics analysis are presented in Chapter 4)



Band	Contig name	Annotation	Species	Domain	CAZy modules	#
1	c178040_g3_i1	**lactase-phlorizin hydrolase-like**	*Crassostrea virginica*	Eukaryotes	**GH1+GH1+GH**	**3**
	c178040_g1_i1	**lactase-phlorizin hydrolase-like**	*Mizuhopecten yessoensis*	Eukaryotes	**GH1+GH1**	**2**
	c175047_g4_i1	**lactase-like protein**	*Larimichthys croceas*	Eukaryotes	**GH1**	**1**
2	c176344_g1_i4	**cellulase**	*Corbicula japonica*	Eukaryotes	**GH9**	**2**
	c124310_g1_i1	collagen alpha-4(VI) chain	*Mizuhopecten yessoensis*	Eukaryotes	none	1
	c178040_g3_i1	**lactase-phlorizin hydrolase-like**	*Crassostrea virginica*	Eukaryotes	**GH1+GH1+GH1**	**1**
3	c176344_g1_i4	**cellulase**	*Corbicula japonica*	Eukaryotes	**GH9**	**1**
	c178040_g3_i1	**lactase-phlorizin hydrolase-like**	*Crassostrea virginica*	Eukaryotes	**GH1+GH1+GH1**	**1**
4	c179345_g2_i11	**beta-mannosidase-like**	*Mizuhopecten yessoensis*	Eukaryotes	unknown	5
	c178089_g2_i1	**beta-mannosidase**	*Crassostrea gigas*	Eukaryotes	**GH2**	**2**
	c172204_g1_i1	**uncharacterized family 31 glucosidase KIAA1161**	*Crassostrea gigas*	Eukaryotes	**GH31**	**1**
5	c165240_g2_i2	histone H2A.V	*Diaphorina citri*	Eukaryotes	none	1
	c165524_g1_i1	cytoplasmic aconitate hydratase-like	*Crassostrea virginica*	Eukaryotes	**GH9**	**1**
6	c176917_g2_i1	**glycosyl hydrolase family 5_53**	*Alteromonadaceae bacterium*	Prokaryotes	**GH5_53**	**2**
	c165524_g1_i1	cytoplasmic aconitate hydratase-like	*Crassostrea virginica*	Eukaryotes	**GH9**	**2**
7	c142397_g1_i1	actin, alpha skeletal muscle	*Crassostrea gigas*	Eukaryotes	none	1
	c167657_g1_i1	beta actin	*Hippoglossus hippoglossus*	Eukaryotes	none	1
	c167657_g3_i1	beta-actin	*Clarias gariepinus*	Eukaryotes	none	1
	c169024_g2_i1	**mannan endo-1,4-beta-mannosidase**	*Mizuhopecten yessoensis*	Eukaryotes	**GH5_10**	**1**
8	c88022_g1_i1	**mannan endo-1,4-beta-mannosidase**	*Mizuhopecten yessoensis*	Eukaryotes	none	2
	c166651_g1_i1	serine protease	*Hyriopsis cumingii*	Eukaryotes	none	1
	c177415_g1_i10	aldose 1-epimerase-like	*Lingula anatina*	Eukaryotes	none	1
	c177885_g3_i1	glyceraldehyde-3-phosphate dehydrogenase	*Bifiguratus adelaidae*	Eukaryotes	none	1
	c178818_g1_i1	glutamine synthetase	*Tridacna squamosa*	Eukaryotes	none	1
9	c165315_g1_i2	**glycosyl hydrolase family 5_8**	*Alteromonadaceae bacterium*	Prokaryotes	**GH5_8+CBM10+CBM2**	**1**
	c177001_g2_i1	**glycosyl hydrolase family 5_8**	*Alteromonadaceae bacterium*	Prokaryotes	**GH5_8+CBM10+CBM2**	**1**
	c173057_g5_i1	growth factor-beta-induced protein ig-h3	*Crassostrea gigas*	Eukaryotes	none	1
	c168880_g1_i1	beta-Ig-H3/fasciclin	*Rhodopirellula sp. TMED11*	Prokaryotes	none	1
	c157067_g1_i1	serine protease	*Hyriopsis cumingii*	Eukaryotes	none	1
	c166928_g1_i1	growth factor-beta-induced protein ig-h3	*Crassostrea gigas*	Eukaryotes	none	1
10	c173837_g2_i2	**auxilliary activity family 10**	*Alteromonadaceae bacterium*	Prokaryotes	**AA10+CBM10**	**2**
	none	**acetylxylan esterase / xylanase**	*Teredinibacter turnerae*	Prokaryotes	**GH10+CBM5+CBM10+CBM50**	**1**
11	c169869_g1_i2	**1,4-beta-xylanase**	*Cellvibrio sp. OA-2007*	Prokaryotes	**GH11+CBM10+CBM5**	**2**
	c169869_g1_i1	**1,4-beta-xylanase**	*Cellvibrio sp. OA-2007*	Prokaryotes	**GH11+CBM10+CBM5**	**2**
	c86197_g1_i1	**endoglucanase-like**	*Crassostrea gigas*	Eukaryotes	GH45	2
12	c177344_g8_i1	transgelin-like protein-3	*Mytilus coruscus*	Eukaryotes	none	1
	c176807_g3_i1	**plant cell wall polysaccharide active protein**	*Alteromonadaceae bacterium*	Prokaryotes	**GH134**	**1**
13	c164057_g1_i1	galectin	*Hyriopsis cumingii*	Eukaryotes	none	2
	c168473_g1_i2	universal stress protein A-like protein	*Crassostrea gigas*	Eukaryotes	none	2
	c176446_g1_i1	uncharacterised	*Lottia gigantea*	Eukaryotes	none	2
	c145250_g1_i1	epididymal secretory protein E1	*Monopterus albus*	Eukaryotes	none	1
14	no match	chain A, Crevice-Forming Mutants Of Bpti	*Bos taurus*	Eukaryotes	none	1

Figure 3.11. SDS-PAGE of the total protein of *L. pedicellatus* caecum fluids. The black rectangles on the gel on the left show the bands that were cut out for analysis and the table on the right presents the results of the protein identification. The annotation was performed against the *L.pedicellatus* RNAseq database and shows, for each band, the contigs that were identified (contig name) with their annotation, species, domain they belong to (domain) and CAZy modules. One match could not be identified in our database and was therefore matched against the NCBI nrdb. The CAZymes are shown in bold. # denotes the number of matches found for the corresponding peptide.

The proteins identified above probably represent only a fraction of those present in the caecum fluids, as some are not abundant enough to be singled out by MALDI-TOF/TOF tandem mass spectroscopy or the band is too faint to be seen on the gel. However, they probably are a good representation of the most abundant enzymes found in the caecum contents. This protein fraction was eluted with SDS and may have released enzymes not present in the fraction assayed for activity (which was extracted without SDS), since this solvent elutes proteins bound to the wood fragments in the caecum. The protein identification results show that most of the bands contain endogenous rather than bacterial proteins even when looking only at CAZymes, though precise abundance estimates cannot be inferred from gel bands, as one band could contain more than one protein. The shipworm CAZymes dominate the upper part of the gel, as they are proteins with a higher molecular weight, including the multi-modular GH1 that has a molecular mass of around 300 kDa (Sabbadin *et al.*, 2018b). Notable is also a cellulase with two GH9 domains that appears in a band over 250 kDa, though the nucleotide sequence encoding for that protein could not be cloned during this project and the band could represent a dimer or multimer. Most of the CAZymes present in the caecum, particularly those of bacterial origin, are multimodular proteins that are characterised by the presence of GH, AA and CBM domains, which allow the proteins to bind to the substrate, thus enhancing the protein catalytic efficiency (Hashimoto, 2006).

In order to identify the proteins in the caecum that are tightly bound to the wood from those that are present in the fluids, some caecum fluids were centrifuged before adding SDS, and the denaturing buffer was applied to the obtained pellet. The supernatant was also used to check for proteins not bound to the wood by adding denaturing buffer to it. After centrifuging, both samples were placed at 100° C for 10 minutes and then subjected to SDS-gel electrophoresis (Figure 3.12). The bands were not subjected to trypsinolysis and MALDI-TOF/TOF tandem mass spectroscopy, but were identified based on the molecular mass from the gel in Figure 3.11.

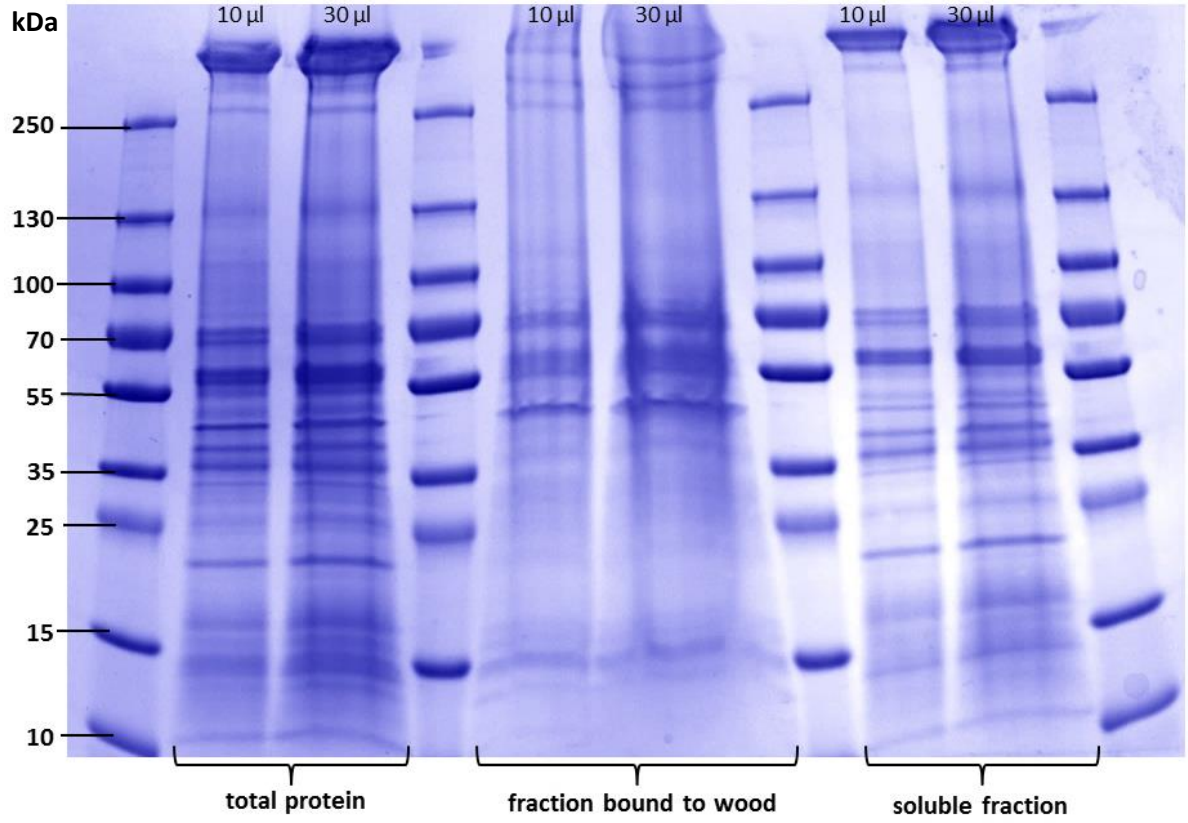


Figure 3.12. SDS-PAGE of the proteins isolated from *L. pedicellatus* caecum. The figure shows on the left the bands found in the total protein sample (same as figure 3.16), in the middle those of the sample with only the proteins attached to wood and on the right those for the proteins associated to the fluids (soluble fraction).

The figure shows that the vast majority of the bands identified in the total protein sample are also seen in the soluble fraction. The only exemption is the third band from the top, which contains two endogenous proteins, a multimodular GH1 and a cellulase with two GH9 domains, both mentioned above. These two proteins must be tightly bound to the wood and can only be separated from it by using strong denaturing agents. The fraction with proteins bound to wood presents fewer bands than the other two fractions, indicating that only some proteins bind tightly to the timber. Most of the proteins located in this fraction are also seen in the total protein sample, suggesting that they are at least partially soluble. The forth band from the top of the soluble fraction, containing two GH2 and a GH31, all of eukaryotic origin, does not appear in the sample bound to wood.

3.5 Digestive glands

3.5.1 Investigation of the digestive glands

3.5.1.1 Microscopy of the digestive glands

The only evidence for the presence of wood containing phagocytes in the digestive glands is at present the drawing presented by Potts in her publication in 1923 (see chapter 1.3.5.4). In order to substantiate this work and to find out whether the phagocytes exist also in *L. pedicellatus*, posterior digestive glands from adult specimens were dissected and prepared for light microscopy and TEM. Figures 3.13-3.15 present the images obtained of the "digestive" portion of this digestive gland, and confirm the presence of the phagocytes containing wood fragments.

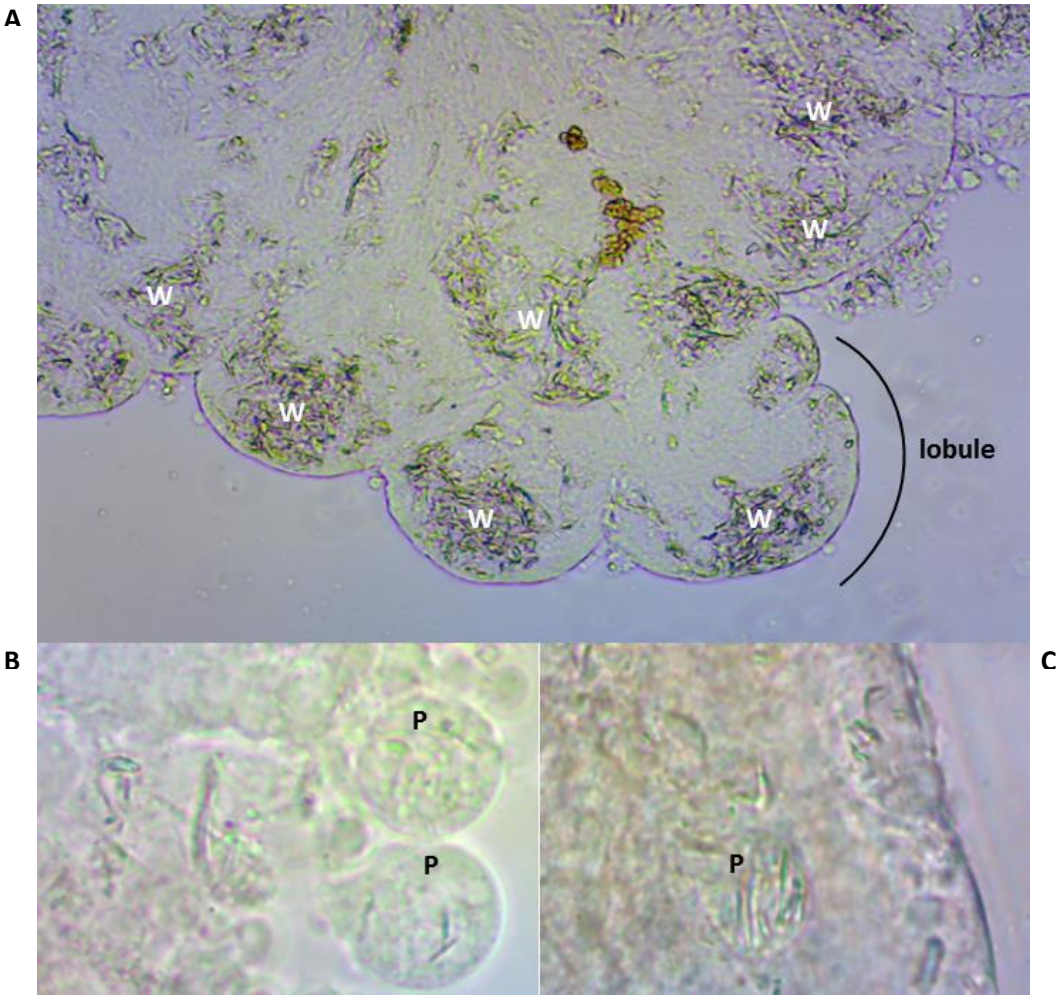


Figure 3.13. Light microscopy of the "digestive" portion of the digestive glands of adult *L. pedicellatus.* **A.** Section of the gland showing wood (W) in the lobules. The picture was obtained by pressing the gland against the coverslip, so that the lobules would break and release the phagocytes, similar to the procedure used by Potts (1923). Magnification 20x. **B.** and **C.** Zoom into the lobules showing phagocytes (P) containing wood. Magnification 100x. Pictures by Federico Sabbadin.

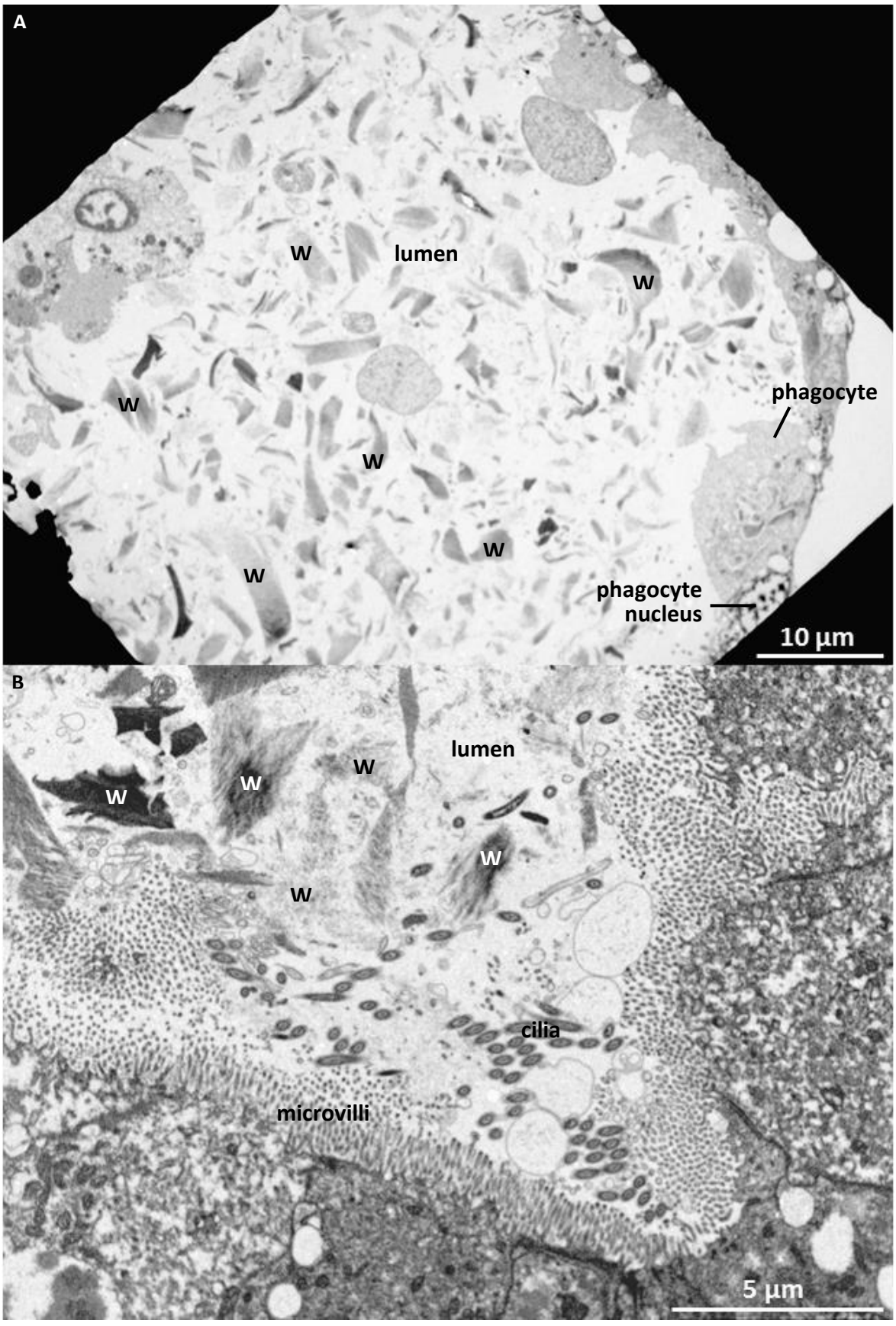


Figure 3.14. TEM image from the "digestive" portion of the digestive glands of *L. pedicellatus.* Both micrographs present an overall view of a gland lobule, with the digestive epithelium and the lumen containing wood fragments (W). A phagocyte emitting pseudopodia is visible on the right bottom corner of **A**. Microvilli and cilia are present on the epithelium shown in **B**. Images by Meg Stark.

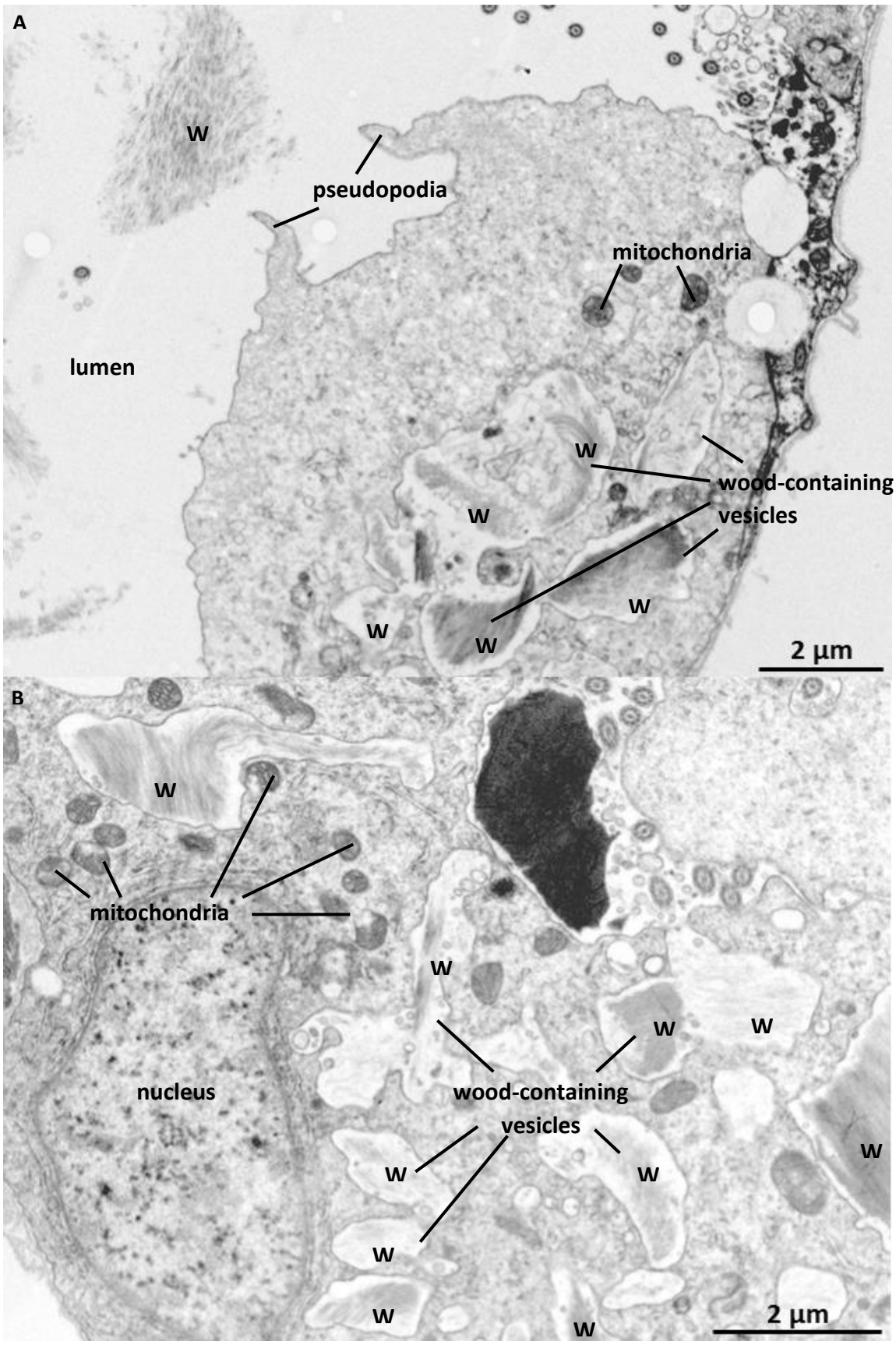


Figure 3.15. TEM image of phagocytes found in the "digestive" portion of the digestive glands of *L. pedicellatus.* **A.** Micrograph showing a phagocyte from the digestive epithelium, emitting pseudopodia and containing wood-filled vesicles. Wood is also visible in the lumen. **B.** Close-up onto a phagocyte, showing numerous wood-containing vesicles and mitochondria. Images by Meg Stark.

These images show that the lumen of the “digestive” portion of the digestive glands is filled up with wood fragments. As described by Potts, the epithelium of this portion is thin and characterised by the presence of amoebic cells, the phagocytes, which can detach from the epithelium and engulf wood particles by the action of pseudopodia. Picture 3.15 shows one of these phagocytes with projected pseudopodia, while numerous fragments of wood can be seen in vesicles within the cell. Microvilli cover the apical epithelium to help increasing the surface area, and cilia are found in the glands duct and partially in the lumen to assist the movement of wood fragments.

The “excretory” portion of the posterior digestive gland was also observed under TEM, and the results are shown in Figure 3.16.

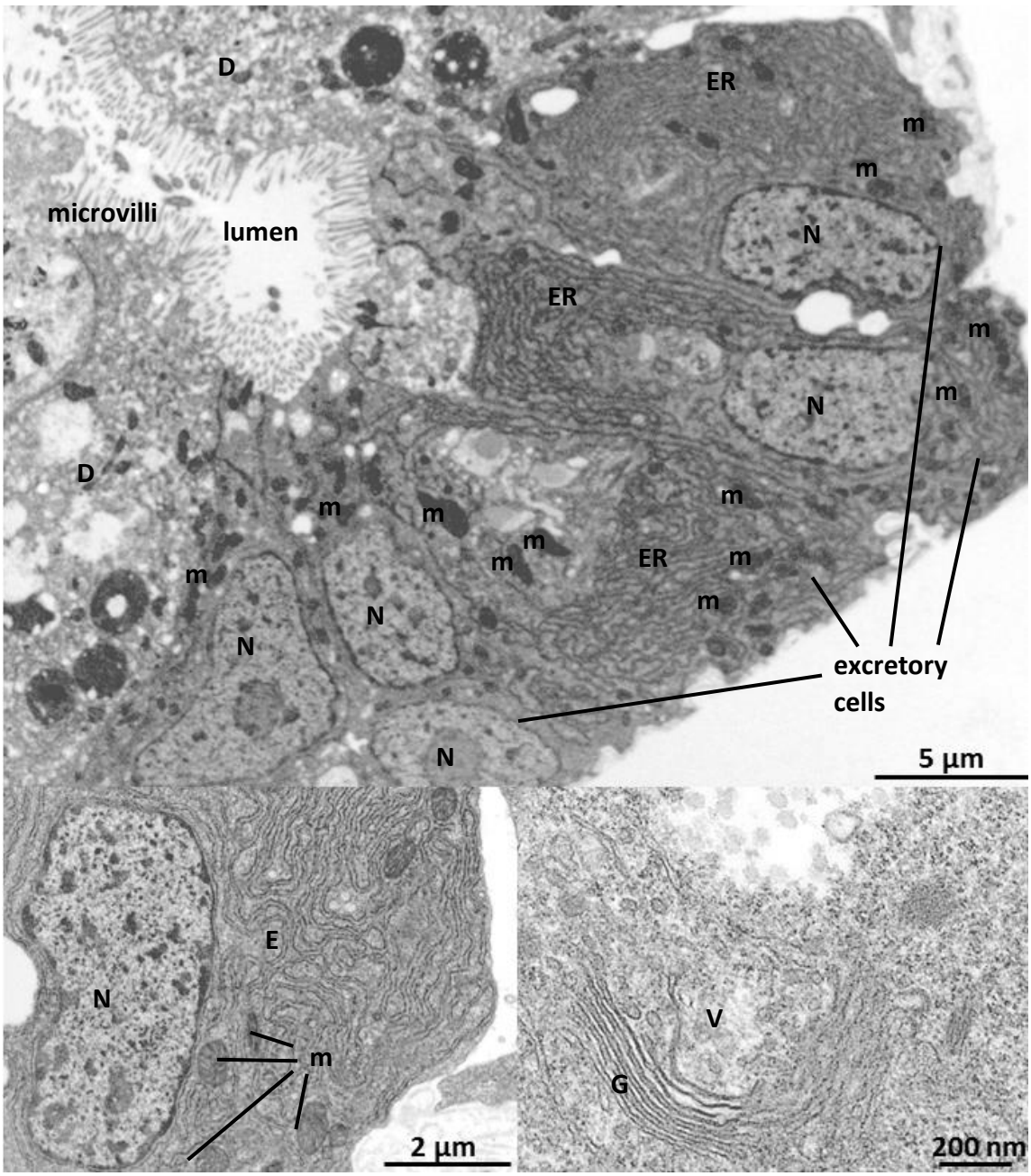


Figure 3.16. TEM images of the "excretory" portion of the digestive glands of *L. pedicellatus.* The top image shows a lobule of the gland, with excretory cells on the right side and digestive cells (D) on the left. The excretory cells present an extensive endoplasmic reticulum (ER) and numerous mitochondria (m) and microvilli on their epithelium. The bottom image shows close ups of the excretory cells, showing in detail the endoplasmic reticulum (left picture) and the Golgi apparatus (G) with secretory vesicles (right picture). Images by Karen Hodgkinson.

This portion of the gland is characterised by cells that present the typical cytology of secretory cells. Indeed, the epithelium is columnar and the cells present a well-developed endoplasmic reticulum, coupled with Golgi apparatus and secretory vesicles that transport the products of the cells to the external environment. Mitochondria provide the energy for

the secretion of high amount of enzymes. Microvilli are also present on the apical epithelium.

3.5.1.2 In vitro activity assay of the digestive glands

In order to assess the ability of the digestive gland to degrade different polysaccharide substrates, and to compare it to the gills and the caecum, an *in vitro* DNS reducing sugar assay was performed with the soluble contents. The digestive glands of five adult *L. pedicellatus* were dissected and homogenised in a non-reducing buffer containing protease inhibitors. The obtained material was sonicated to disrupt the cells and release the contents, and after centrifugation the supernatant was tested with the Bradford method to estimate protein abundance. The DNS reducing sugars assay was performed on thirteen different substrates, including pachyman, PASC, carboxy methyl cellulose (CMC), xylan, xyloglucan, arabinoxylan, arabinogalactan, mannan, glucomannan, LBG, galactan, lichenan and avicel. The results are shown in Figure 3.17.

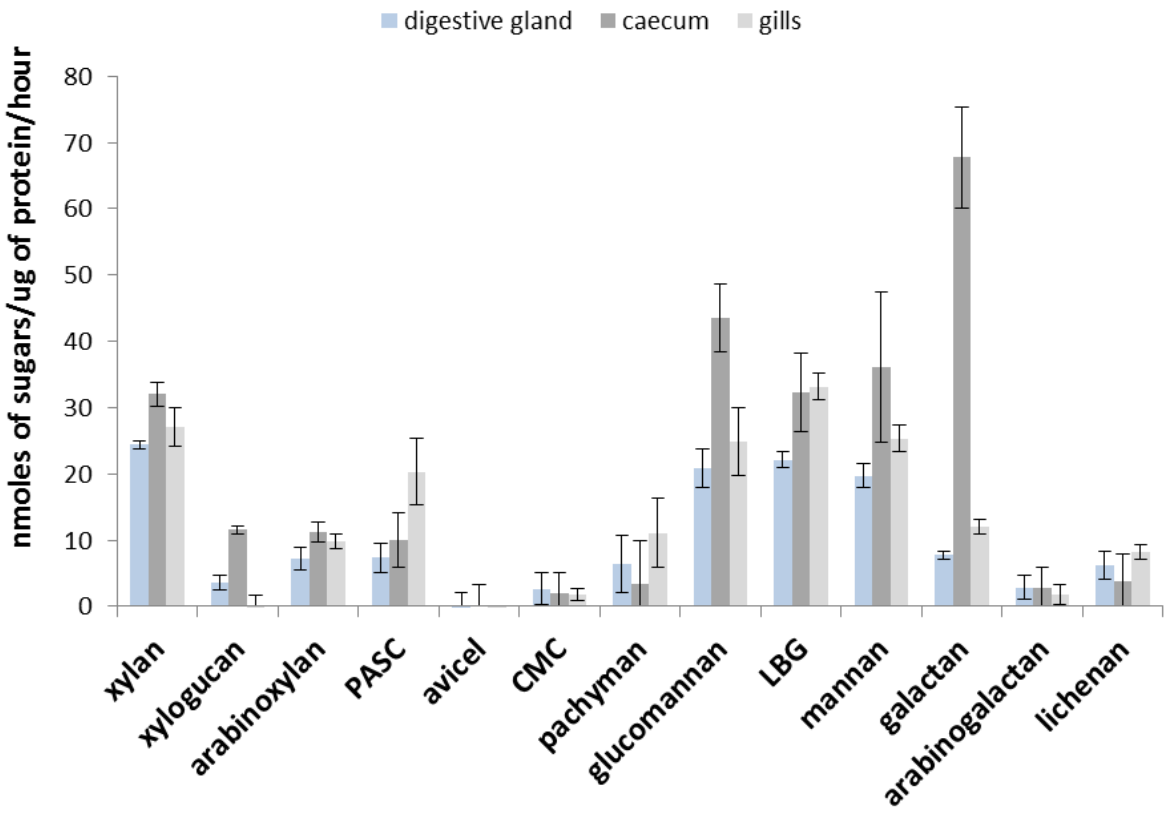


Figure 3.17. DNS reducing sugars assays of the digestive glands after incubation for 20 hours on a range of substrates. The 50 µl reactions contained 2.5 µg of digestive gland soluble protein, substrates at the concentration of 0.1% and 50 mM sodium phosphate buffer pH 7.6. The results for the digestive gland are shown in light blue, while the ones in grey are from the gills and caecum and have been added to the graph for comparison.

The graph shows that the enzymatic activities of the digestive glands have a similar pattern to those of the caecum, with galactan, glucomannan, mannan, LBG and xylan giving the highest amount of sugar release. This indicates that the enzymes produced by the digestive glands are suited for the digestion of glucans, mannans and xylans. No or very little activity was recorded on Avicel or CMC. The difference between gland and caecum is mainly in the amount of sugars released, with the caecum producing much more sugars than the gland. This can be explained, because while in the caecum the lignocellulolytic digestive enzymes are concentrated, in the glands many different enzymes are found, including those that are not related to wood digestion; this would give a lower signal compared to the caecum, since the same quantity of protein was used overall, but probably a lower quantity of digestive enzymes. Furthermore, the caecum includes both bacterial and endogenous enzymes while the gland should only contain enzymes produced by the shipworm.

3.6 Crystalline style

3.6.1. Investigation of crystalline style and its sac

3.6.1.1 Light and electron microscopy of the crystalline style and its sac

Despite being identified in 1686 (Heide, 1892), and being an unusual organ, the literature on the crystalline style is far from extensive and its function still remains largely unknown. The studies published so far focus mainly on the crystalline styles of bivalves, while publications on gastropods are rare. The structure and histology of the style and its sac rely mainly on drawings and descriptions of what was observed under the microscope, being most the publications from the 1970s or earlier. Only a few publications present light or electron microscopic pictures of the style or the sac, for a very limited number of species of non-shipworm bivalves only (McQuiston, 1970; Kristensen, 1972; Judd, 1979; Tall and Nauman, 1981). No work on the crystalline style of shipworms has been previously published.

The style of molluscs varies between species, even among bivalves, which is a consequence of the varied habitats in which they live (Alyakrinskaya, 2001). As a general rule, the style can be described as a hyaline cylindrical rod, whose posterior end is round, while its anterior side is more elongated and often presents a thinner tip. Food detritus is often seen

on the apex of the style, followed by a more bulbous area and a sort of collar (Figure 3.18). The core of the style has a granular appearance (Kristensen, 1972) and is described to have a concentric layered structure (Edmondson, 1920; Alyakrinskaya, 2001).

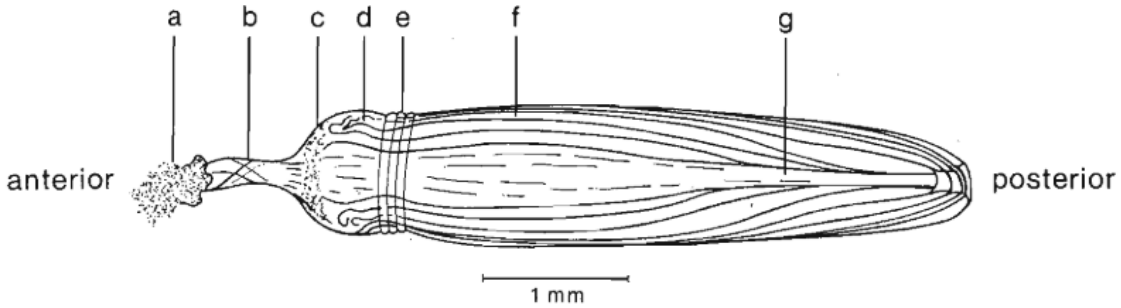


Figure 3.18. Schematic representation of a typical bivalve crystalline style as described by Kristensen in 1972. a) lump of food detritus, b) tip, c) fine particles embedded in mucus, d) bulbous area, e) collar-like structure, f) coarse lamellae and g) central granular area.

In order to find out whether the style of shipworms is similar to that of other bivalves, styles were dissected from adult *L. pedicellatus* individuals and were observed under light microscopy, after a wash with pure water to eliminate any attached debris resulting from the dissection. The results are shown in Figure 3.19 and illustrate how the style of shipworms is structurally similar to those of other bivalves, showing a round border on its posterior side and a thin tip on the anterior end, a bulbous area after the tip and a collar-like structure. The posterior side appears more opaque that the rest of the style and some structure is visible in the core area of the organ, whose nature is unknown (it is possible that they could be artefacts resulting from the phase contrast technique). The central area with a granular appearance described by Kristensen is not evident in this picture.

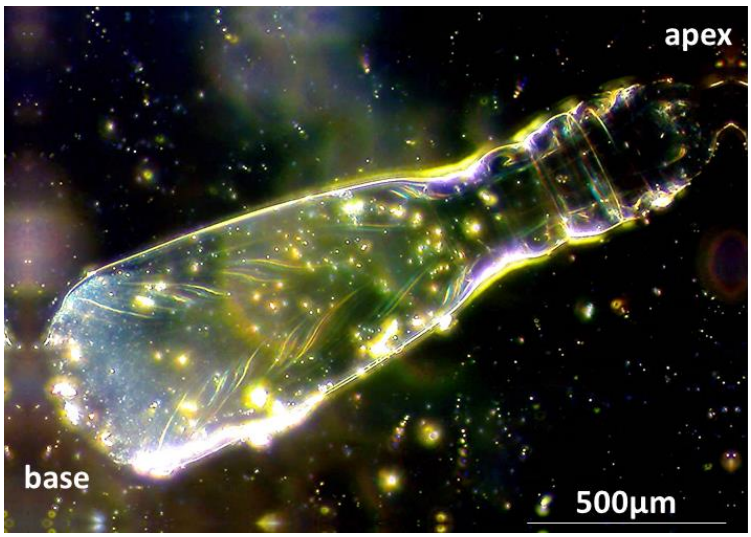


Figure 3.19. Light microscopy phase contrast image of the crystalline style of *L. pedicellatus* once the sac has been removed. Picture by Federico Sabbadin. Magnification 20x.

In eulamellibranchs, the crystalline style sac can either be connected to the intestine via a slit, or be completely separate from the intestine (Judd, 1979); the latter is the case for *L. pedicellatus*, as observed during the dissections. The fine structure of the sac of bivalves has been studied in detail by various authors, with the most complete publication being that produced by Judd in 1979, which includes drawings as well as good quality electron microscopy images. According to Judd, the cells that line the sac are mainly divided into A and B type, which are found at different heights along the style length. The A cells are predominant around the middle and anterior area of the style and they are columnar cells characterised by being tall, regular and covered with dense cilia; they contain numerous large mitochondria, and granules of lipofuscin are found throughout the cell. These cells are not thought to have a secretory nor absorptive function, but instead to provide the energy (from the mitochondria) and the structure (the cilia) to rest and rotate the crystalline style (Mackintosh, 1925). B cells are found only at the base of the sac and are also tall and slender, but have less and longer cilia and also present microvilli. They are characterised by the presence of an extensive system of rough endoplasmic reticulum as well as Golgi apparatus, and they contain numerus vesicles that can be seen as being secreted between the cilia. They contain fewer mitochondria and lipofuscin is almost absent. These cells have clearly a secretory nature and they are thought to be the ones involved in the production of the crystalline style itself.

With the view of exploring the structure of the crystalline style sac of the shipworms and in particular of *L. pedicellatus*, crystalline styles still contained in the style sac were dissected, fixed and prepared for visualisation by TEM (details in section 2.2.3). Images of the analysed sections are shown in Figure 3.20-3.21.

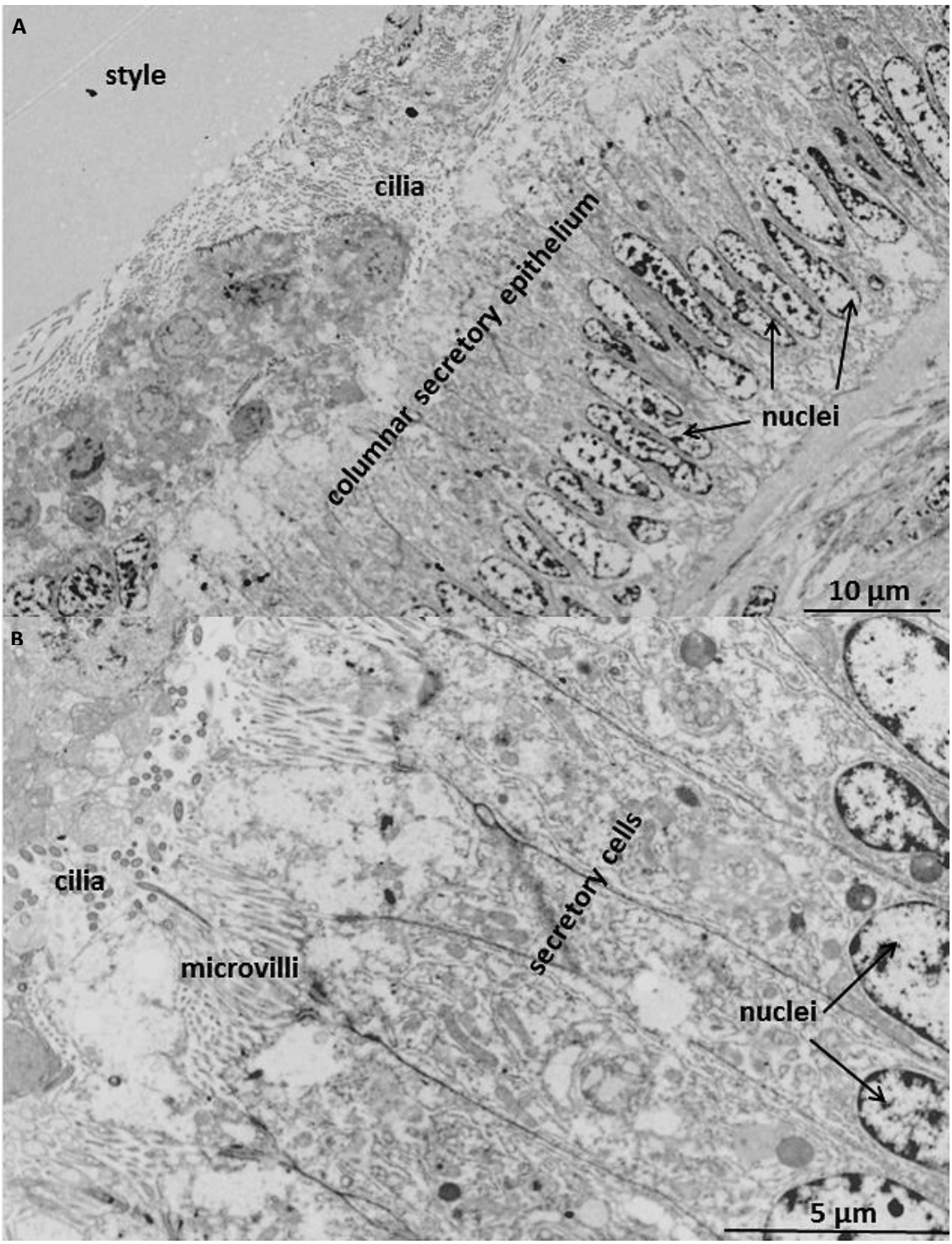


Figure 3.20. TEM images of the basal part of the crystalline style sac of *L. pedicellatus.* **A.** Image showing the columnar secretory epithelium, with the crystalline style on the top left corner lying on a layer of cilia. **B**. Close up of the secretory cells, showing the abundant endoplasmic reticulum and secretory vesicles. Images by Clare Steele-King.

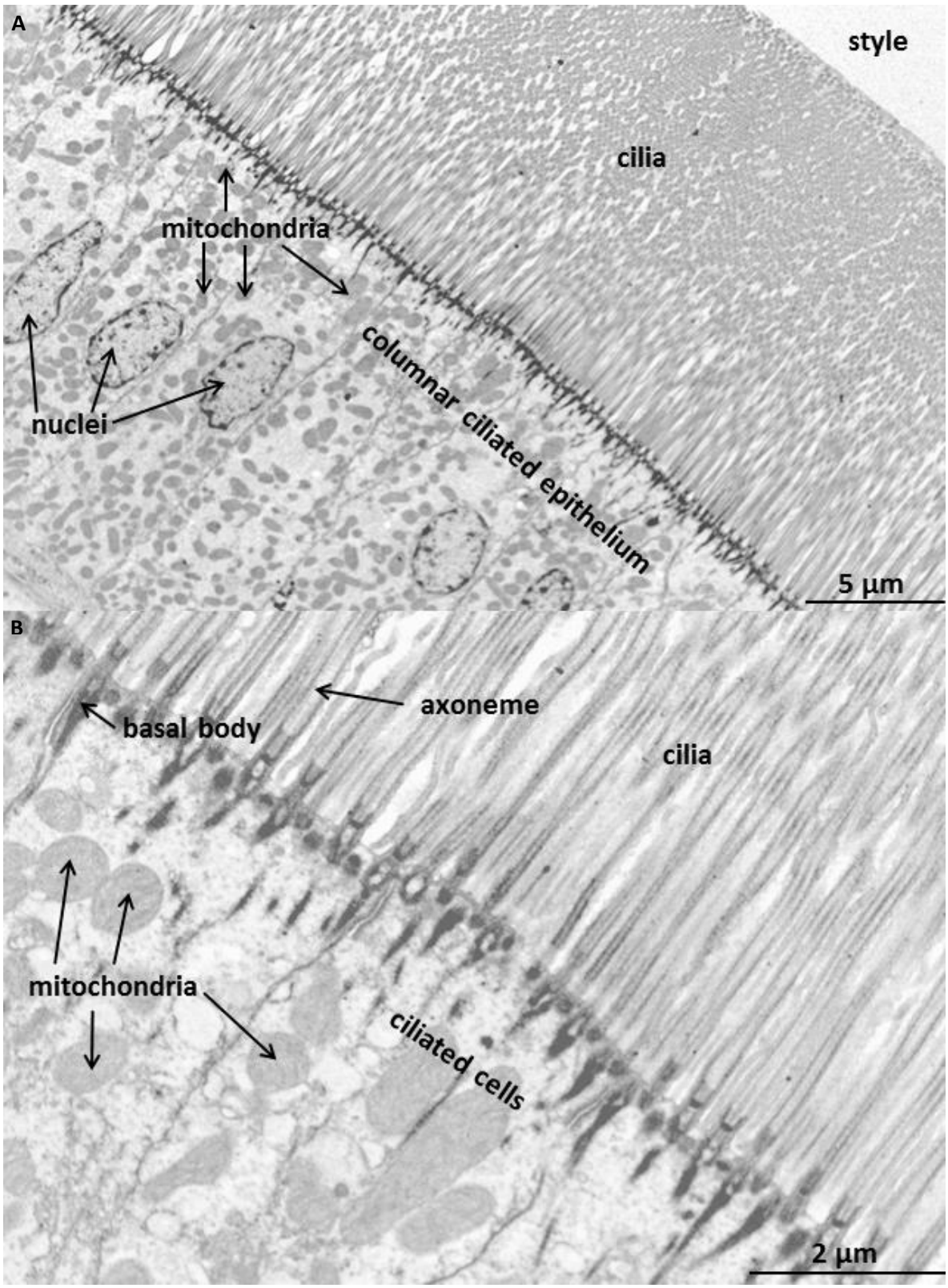


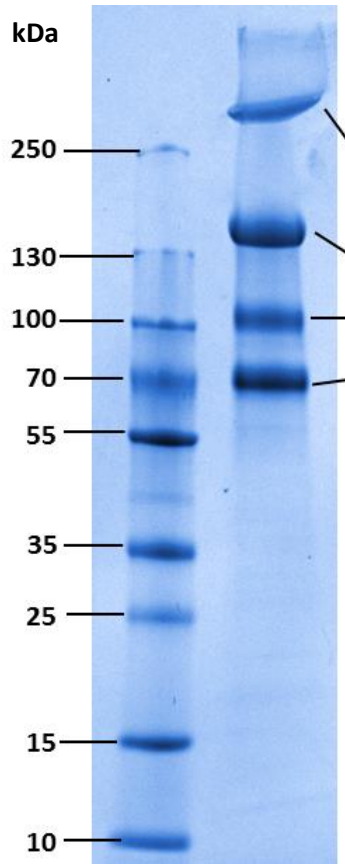
Figure 3.21. TEM images of the middle part of the crystalline style sac of *L. pedicellatus.* **A.** Image showing the columnar ciliated epithelium, with the crystalline style on the top right corner lying on a very thick layer of cilia. A vast amount of mitochondria is visible in the cells. **B.** Close up of the ciliated cells, showing in detail the structure of the cilia and the numerous mitochondria. Images by Clare Steele-King.

The TEM images presented here are the first of the crystalline style sac of a shipworm. They show that in *L. pedicellatus* there is the same cell arrangement seen in most other bivalves. Indeed, the basal part of the style contains secretory cells in a well-structured columnar epithelium, similar to those described as cell type B by Judd in 1979. Some cilia cover the epithelium but they are not particularly dense, and microvilli are also present. The cells contain rough endoplasmic reticulum as well as Golgi apparatus, and numerus secretory vesicles. The cells at the middle and top of the sac of *L. pedicellatus* are clearly similar to the type A cells of bivalves, being covered by an extensive layer of cilia and containing numerous mitochondria to provide energy for the movement of the cilia. In contrast to other bivalves' cells, no lipofuscin granules were seen in these samples.

3.6.2.2 Proteins of the crystalline style

The amount, nature and characteristics of the proteins of the crystalline style have been studied by researchers using a variety of different methods. The overall protein content varies according to different publications, but ranges from around 12 to 14% of the total weight; indeed, the style is mainly composed of water (Nelson, 1917; Bailey and Worboys, 1960; Alexander *et al.*, 1978). The first visualisation of the proteins of bivalves was made by Bedford (1969), who carried out flat-plate acrylamide gel electrophoresis on the styles of different species to show that the protein patterns are constant among individuals of the same species but varies between different species, and can therefore be used as a complement to taxonomical identification. Gel electrophoresis of other species was used to compare the composition of the style along its length (Alexander *et al.*, 1978). The most complete work so far on the crystalline style proteins of a bivalve shows that they can be classified into medium molecular weight proteins (around 40-50 kDa) and high molecular weight ones (95-170 and >300 kDa), which act together in forming the gel that constitutes the style. These proteins are highly glycosylated, with the medium-sized ones showing less associated carbohydrates. The carbohydrates allow the retain of water and therefore help in the formation of the gel by the crosslinking of disulphide bonds (Mackenzie and Marshall, 2014). Most of the proteins so far discovered in the style were identified by their activity, while only a handful of authors managed to either purify a style protein (Sova *et al.*, 1970; Wojtowicz, 1972) or to identify them by 2D gel electrophoresis followed by MALDI-TOF tandem mass spectroscopy of selected spots (Mackenzie and Marshall, 2014).

To explore the characteristics and properties of the protein of the crystalline style of *L. pedicellatus*, SDS-gel electrophoresis of the style of some adult individuals was firstly performed, to compare the pattern to other bivalve species. The crystalline styles were dissected from 10 different shipworms, pulled together and dissolved in 1% SDS prior to loading. The obtained gel (Figure 3.22) shows a remarkably low number of bands, with only four being clearly visible and distinct. These bands were excised from the gel and were subjected to trypsinolysis and MALDI-TOF/TOF tandem mass spectroscopy, and the obtained spectra were compared to the EST database obtained from sequencing the RNA extracted from caecum, digestive glands gills and crystalline style of *L. pedicellatus.*



Band	Contig name	Annotation	Species	Domain
1	cap3_Contig92120	collagen alpha-4(VI) chain-like	*Lingula anatina*	Eukaryotes
	cap3_Contig80246	collagen alpha-4(VI) chain	*Crassostrea gigas*	Eukaryotes
2	o_r_c154355	cartilage-matrix protein like	*Crassostrea virginica*	Eukaryotes
3	cap3_Contig90328	cartilage matrix protein	*Mizuhopecten yessoensis*	Eukaryotes
4	cap3_Contig6858	MACPF domain-containing protein 4	*Mytilus galloprovincialis*	Eukaryotes

Figure 3.22. SDS-PAGE of *L. pedicellatus* crystalline style. The well was loaded with 10 µl of protein. The black arrows show the bands that were cut and the table on the right presents the results of the protein identification performed on them. The annotation was performed against the *L. pedicellatus* RNAseq database and shows, for each band, the contigs that were identified (contig name) with their annotation, species and domain.

The SDS PAGE analysis was useful in identifying some of the main proteins of the crystalline style, which were annotated as similar to collagen or cartilage. Probably these proteins form the style matrix and give it structure and rigidity. No mucin, albumin or globulins (which were reported by some authors) were identified in the analysed bands, nor any

protein that can be accounted responsible for the reported digestive capabilities of the style. However, only the main bands of the gel we analysed, while other proteins might be present in the style but not be particularly visible in the gel. Furthermore, the MALDI-TOF/TOF tandem mass spectroscopy technique is not sensitive enough to recognise all the proteins present in a gel band, but only identifies the most abundant; some digestive enzymes could have been present in the samples but have gone undetected. One of the proteins found in band four was annotated as having a MACPF (membrane attack complex/perforin) domain. This is an interesting and unexpected as these proteins have antimicrobial properties and are part of the defence system in vertebrates, where they act by causing lysis of gram-negative bacteria (Rosado *et al.*, 2008). Antimicrobial properties have been recorded for the crystalline style of a few bivalve species (McHenery and Birkbeck, 1979; Ito *et al.*, 1999; Nilsen *et al.*, 1999; Miyauchi, 2000; Olsen *et al.*, 2003; Xue *et al.*, 2007). These activities were thought to be due to the presence of lysozyme-like proteins in the crystalline style, which were not identified in the protein-identification analysis for the *L. pedicellatus* crystalline style.

In order to investigate the glycosylation pattern of the crystalline style of *L. pedicellatus*, a similar experiment to that performed by Judd (1987) and Mackenzie and Marshall (2014) was carried out. Two SDS-polyacrylamide gels were electrophoresed in denaturing conditions, but while one was normally stained with Coomassie blue to identify the proteins, the other one was dyed using the Schiff's fuchsin-sulfite glycoprotein stain, which highlights the presence of carbohydrates. The resulting gels depicted in Figure 3.23 show that overall most of the proteins found in the style are highly glycosylated. The lowest molecular weight band is the one presenting less glycosylation, while the band with molecular weight of around 140 kDa is the one presenting the highest amount of carbohydrates. Some proteins were found to remain in the well and have hardly entered the resolving gel, which are possibly those of extremely high molecular weight that were identified by Mackenzie and Marshall (2014); these also seem to be carrying high quantities of carbohydrates.

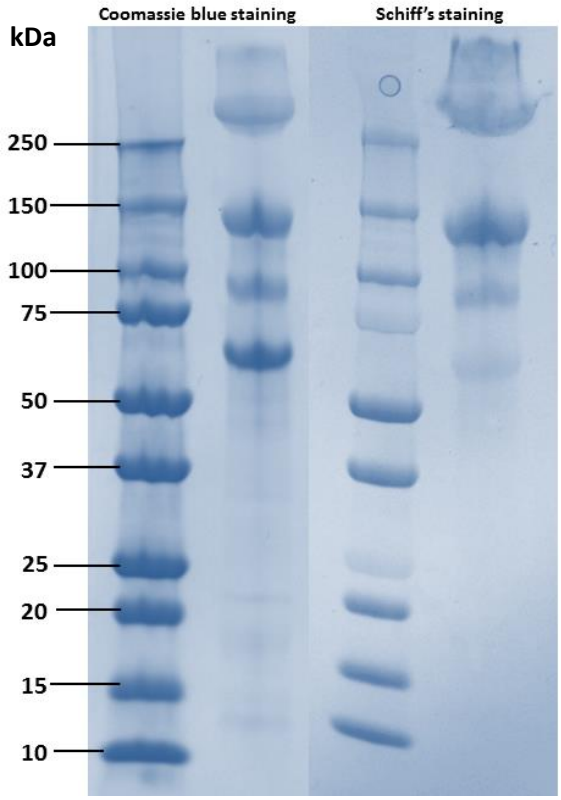


Figure 3.23. Carbohydrate staining of the crystalline style of *L. pedicellatus*. SDS-PAGE with Coomassie (left) and Schiff's fuchsin-sulfite glycoprotein stain (right). The well was loaded with 10 µl of protein.

3.6.2.3 Enzymatic activities of the crystalline style

One of the main roles identified for the crystalline style is that of contributing to the extracellular digestion of the ingested food. The way it is thought to perform this task is to release the digestive enzymes it contains while the style consumes itself by scraping against the gastric shield. Most of the literature on the crystalline style focuses on identifying the digestive enzymes contained in this organ, and therefore the substrates that it can degrade. The first carbohydrase identified in the style was an α-amylase, whose presence in a species of cockle was demonstrated by Coupin in 1900. Since then, many enzymatic activities have been identified throughout the years, some recorded by many authors, others only by few. Table 3.1 summarises the enzymatic activities so far highlighted by these papers.

Table 3.1. Summary of the enzymatic activities of the crystalline style of bivalves and gastropods.

Enzymatic activity	Class	Reference
amylase	bivalve, gastropod	Coupin, 1900; Yonge, 1931; Lavine, 1946; Bailey et al., 1960; Horiouchi and Lane, 1966; Kristensen, 1972; Wojtowicz, 1972; Alexander et al., 1978; Seiderer and Newell, 1979; Brock and Kennedy, 1992; Alyakrinskaya, 2001
cellobiase	gastropod	Horiouchi and Lane, 1966
cellulase	bivalve, gastropod	Lavine, 1946; Bailey et al., 1960; Horiouchi and Lane, 1966; Alexander et al., 1978; Hammed and Paulpandian, 1987; Brock and Kennedy, 1992; Alyakrinskaya, 2001; Sakamoto, 2008; Mackenzie et al., 2014
chitinase	bivalve	Smucker and Wright, 1984; Smucker and Wright, 1986
chitobiase	bivalve	Wojtowicz, 1972
fucoidanase	gastropod	Alexander et al., 1978
galactosidase	gastropod	Horiouchi and Lane, 1966; Kristensen, 1972; Wojtowicz, 1972; Payne et al., 1972; Wojtowicz, 1972; Alexander et al., 1978
glucanase	bivalve	Mackenzie et al., 2014
glucosidase	Bivalve. gastropod	Lavine, 1946; Horiouchi and Lane, 1966; Kristensen, 1972; Wojtowicz, 1972; Hammed and Paulpandian, 1987
glycogenase	bivalve	Hammed and Paulpandian, 1987
laminarinase	bivalve, gastropod	Bull and Chesters, 1966; Hammed and Paulpandian, 1987; Sova et al., 1970; Kristensen, 1972; Wojtowicz, 1972; Alexander et al., 1978; Brock and Kennedy, 1992
lipase	bivalve	Bailey et al., 1960; Hameed and Paulpandian, 1987; Alyakrinskaya, 2001
lysozyme	bivalve	McHenery and Birkbeck, 1979; Seiderer et al., 1984; Olsen et al., 2003; Xue et al., 2007
maltase	bivalve, gastropod	Horiouchi and Lane, 1966; Kristensen, 1972
oxidase	bivalve	Lavine, 1946; Berkeley, 1923
sucrase	bivalve, gastropod	Horiouchi and Lane, 1966; Kristensen, 1972
trehalase	gastropod	Horiouchi and Lane, 1966
xylanase	gastropod	Alexander et al., 1978

With the aim of assessing the carbohydrate degrading abilities of the crystalline style of shipworms, 10 crystalline styles from adult specimens of *L. pedicellatus* were dissected and solubilised in 50mm sodium phosphate buffer pH 7.6 containing protease inhibitors, to preserve enzymatic activities. A Bradford assay was performed to estimate protein abundance and a DNS reducing sugars assay was carried out on thirteen different substrates, including pachyman, PASC, CMC, xylan, xyloglucan, arabinoxylan, arabinogalactan, mannan, glucomannan, LBG, galactan, lichenan and avicel. The results are shown in Figure 3.24.

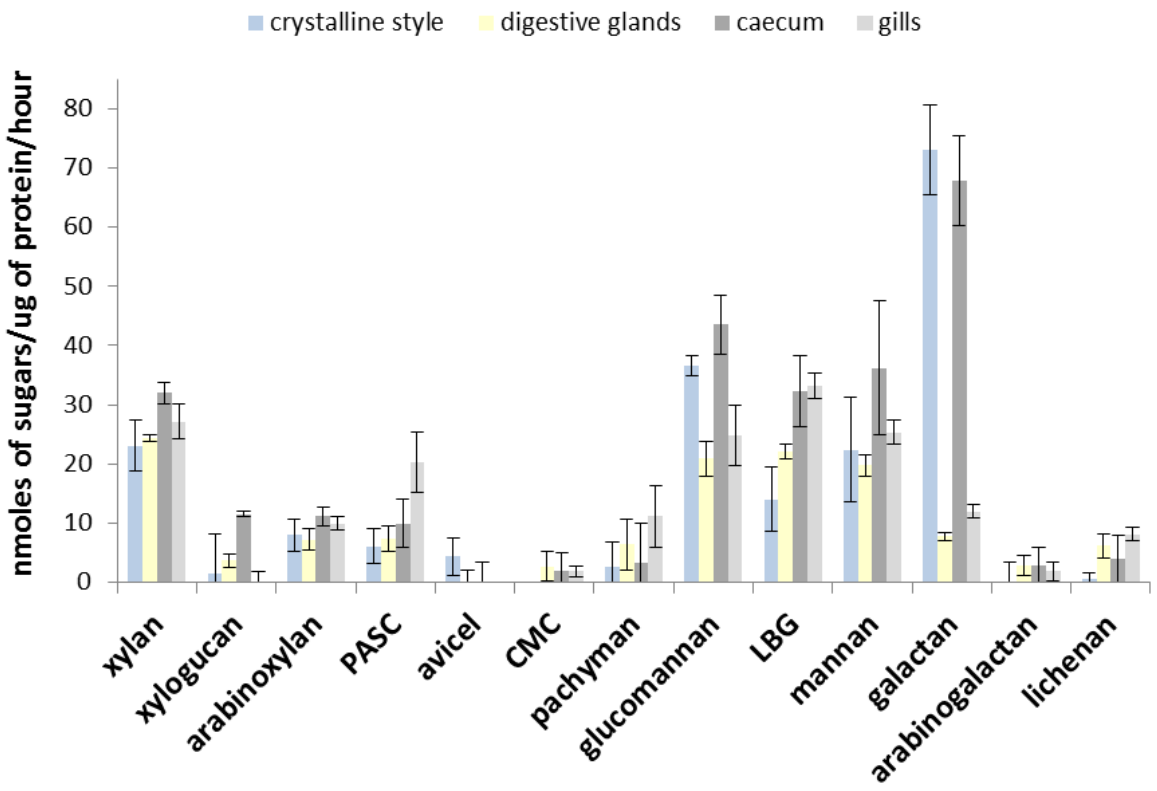


Figure 3.24. Enzymatic activity of the crystalline style on a range of substrates. The 50 µl reactions contained 2.5 µg of crystalline style soluble protein, substrates at the concentration of 0.1%, 50 mM sodium phosphate buffer pH 7.6 and were incubated for 20 hours. The results for the crystalline style are shown in light blue, while the ones in grey and yellow are from the gills, caecum and digestive glands and have been added to the graph for comparison.

This assays shows that enzymes from the crystalline style of *L. pedicellatus* are able to degrade a wide range of polysaccharides, with a preference for mannans, glucans and xylan. The ability of the shipworm's style to degrade mannan has not been previously reported for molluscs, and a xylanase was only previously recorded in a gastropod (Alexander *et al.*, 1978). These two enzymatic activities might be a consequence of the diet of the animals, which were grown in Scots pine, a gymnosperm wood rich in glucomannans and xylans. Similarly to the other organs of the digestive system, the crystalline style does not seem to possess enzymes able to digest crystalline cellulose or Avicel, and very little for CMC, pachyman, lichenan and arabinogalactan. Compared to the other organs, the carbohydrate degrading properties of the style seem very similar to those of the caecum.

3.7 Discussion

Shipworms of the species *L. pedicellatus* have been shown to be able to live on a diet of wood (Becker, 1959; Gallager *et al.*, 1981). The means by which they succeed to survive on such a diet, which lacks in nitrogen and other nutrients, is by the establishment of a symbiosis with bacteria that live in bacteriocytes in the shipworms gills (Popham and Dickson, 1973). The bacteria not only fix the atmospheric nitrogen and provide it to the host, but also produce enzymes able to digest lignocellulose, as well as antibiotics and other secondary metabolites beneficial to the shipworm (Distel *et al.*, 2002b; Elshahawi *et al.*, 2013; Haygood *et al.*, 2015). The existence of this type of endosymbiosis is nowadays well accepted, however its details are not completely known. In particular, there is much debate on the mechanism by which the lignocellulolytic enzymes produced by the bacteria in the gills (a respiratory organ) reach their final destination (the caecum, a digestive organ), which is physically distant and does not appear to have an obvious connection to the respiratory system. The work presented in this chapter shows that the most likely mechanism for the transfer of the bacterial enzymes to the caecum is the food groove, a structure typical of bivalves that is normally used in connection with filter feeding. This mucous channel captures food particles on the gills surface and transports them to the mouth by the action of cilia (Gosling, 2015). The food groove is the only connection between the gills and the digestive system in shipworms and it is hypothesised to be the source of the movement of the bacterial CAZymes. Electron microscopy, *in vivo* staining and immunogold labelling of the caecum, gills and food groove – both proximally and distally to the gills – have shown the presence of the bacteria-containing bacteriocytes and of active glycosylases of prokaryotic origin in the three tissues, suggesting that the bacteria are not just limited to the gills lamellae but travel along to the food groove all the way to the mouth, and that enzymatic activity related to wood digestion is found along the same route. This means that the bacteria – and the enzymes they produce – can therefore reach the digestive system and eventually the caecum by travelling via the food groove. Here the prokaryotic lignocellulolytic enzymes contribute to the digestion of wood in synergy with the enzymes produced by the shipworm. The utilisation of the food groove as a mean of transporting bacteria and enzymes from the gills to the digestive system is an interesting adaptation, which suggests that the shipworm might have adopted an organ that was already found in bivalve for a new function, the transport of wood-degrading enzymes. This new transport system could have allowed them to exploit the lignocellulolytic enzymes

produced by the bacteria in the gills, making wood digestion more efficient and permitting the shipworms to take full advantage of the abundant carbohydrates found in wood; this in turn could have allowed them to fill an ecological niche in the marine ecosystem that was only partially utilised by other organisms. It is also possible that the utilisation of the food groove as a transport system for bacterial products is widespread among molluscs and not just found in shipworms. Indeed, the presence of endosymbiotic bacteria in the gills of bivalves and other molluscs is a common phenomenon (Roeselers and Newton, 2012) and the utilisation of the bacterial products by transport via the food groove could allow these species to live in unusual environments, such as deep marine hydrothermal vent ecosystems (Distel *et al.*, 2000). Furthermore, the observation of intact bacteriocytes moving along the food groove towards the digestive system suggests that the shipworms could be using the food groove not only for the transport of the bacterial CAZymes, but also as a way of utilising the bacteria as a source of nutrition, in particular as a supply of nitrogen to complement the scarcity of this element in wood. Indeed, marine bivalves have been observed to utilise bacteria as food source, particularly when phytoplankton is scarce (McHenery *et al.*, 1979; Langdon and Newell, 1990). In shipworms nitrogen is fixed by the bacteria in the gills and it has been shown to be transferred from its source in the endosymbionts directly to the host shipworm gill's tissue (Carpenter and Culliney, 1975; Lechene *et al.*, 2007); however the transport of bacteria to the digestive system via the food groove could provide an additional nitrogen source directly to the digestive system, where it could be recycled and absorbed during the digestion process.

The compositional analysis, comparing the wood where the shipworms were grown to the frass they produced, confirmed that *L. pedicellatus* is able to extract nutrients from the timber, mainly attacking the cellulose but partially also the hemicellulose fraction of the wood, with a preference for the sugars mannose and xylose of this latter component. The ability to utilise cellulose as a source of nutrition is quite common among xylophagus animals, therefore these results do not come as a surprise. Indeed, many organisms such as bacteria, fungi, protozoa, marine and land invertebrates are well known to have cellulose and hemicellulose degrading abilities, which they perform by utilising a cocktail of carbohydrate-active enzymes that include glycosyl hydrolases, carbohydrate esterases, polysaccharide lyases and lytic polysaccharide monooxygenases (Coleman, 1978; Vaaje-Kolstad *et al.*, 2010; van den Brink and de Vries, 2011; Payne *et al.*, 2015), as explained in

the introduction (section 1.2). Termites, for example, are able to dissimilate 74-99% of the cellulose and 65-87% of the hemicellulose that they ingest (Bignell and Eggleton, 2000) and white-rot fungi slightly less (Sun *et al.*, 2013); recently, the marine crustacean *Limnoria quadripunctata* was discovered to able to turn over half of the cellulose it finds in wood, though the hemicellulose and lignin fractions are left mainly unconsumed (Besser *et al.*, 2018). Hemicelluloses can also be degraded with high efficiency by bacteria (Sun *et al.*, 2013).

Lignin remains largely undigested by *L. pedicellatus* and it accumulates in the frass. The ability of modifying or degrading lignin has only been discovered in a handful of organisms, since this heterogeneous polymer is highly recalcitrant to digestion. White-rot fungi are the only organisms in nature that can completely mineralise lignin, though this is done without energy value for the fungus (Shah *et al.*, 1992); recently some ligninolytic bacteria have been discovered that secrete peroxidase and laccase able to degrade lignin (Bugg *et al.*, 2011). Brown-rot fungi do not have such enzymes, and only modify the lining in order to be able to access hemicellulose and cellulose; for this function they utilise small reactive molecules (Arantes and Goodell, 2014). Termites were thought not to be able to breakdown lignin, however only recently some studies have suggested that they produce laccases that could have a role in lignin digestion (Tartar *et al.*, 2009; Coy *et al.*, 2010). It is possible that the shipworm has the ability of modifying the lignin in order to access the sugars found in the wood. Indeed, the experiments only aimed at quantifying the amount of lignin in wood and frass, while no qualitative analysis has been performed. In order to identify modification of the structure of lignin (such as acetylation or oxidation, for example), other analytical methods need to be utilised, such as Fourier-transform infrared (FTIR) spectroscopy or solid-state ^{13}C nuclear magnetic resonance (ssNMR).

The shipworm's ability to feed on wood is dependent on the enzymes found in the caecum, an organ that is characteristic of wood boring bivalves. The caecum is known to be related to wood digestion and absorption due to the presence of a typhlosole to increase the absorptive area, of a microvillar border that absorbs the nutrients and of an extensive vascular network to transport them to the rest of the body (Betcher *et al.*, 2012). The TEM images of the caecum have shown for the first time wood fragments in the caecum lumen, and a section of the typhlosole with a brush border on the caecum lumen side, confirming the importance of the caecum in wood digestion and absorption.

Work in this chapter focused on presenting some biochemical evidence of the wood-degrading abilities of the caecum contents, which were lacking in the literature. SDS-gel electrophoresis and protein identification by trypsinolysis and MALDI-TOF/TOF tandem mass spectroscopy showed that the caecum contains numerous bacterial and shipworms CAZymes, most of which are multimodular proteins containing GH or AA modules linked to CBMs, a characteristic that enhances catalytic efficiency (Hashimoto, 2006). Only one LPMO, of bacterial origin, was discovered in the caecum. This was annotated by a Blastx search against the NCBI non redundant database as belonging to the Auxiliary Activity class number 10 (AA10) and containing also a CBM of the family 10 (CBM10), and is the only recognised oxidative CAZyme in the caecum, while the other CAZymes are glycosyl hydrolase. No endogenous LPMOs were identified with this method. A DNS reducing sugar assay was used to investigate the substrate preference of the enzymes found in the caecum contents. The assay has shown that the caecum enzymes can degrade various substrates normally associated with woody biomass. In particular, substrates like galactan, mannan, glucomannan, LBG and xylan are those that produce the highest amount of reducing sugars when incubated with the caecum fluids, which reflect the results of the compositional analysis showing that mannan and xylan are the hemicelluloses mostly utilised by *L. pedicellatus*. No activity was recorded on Avicel, suggesting that the caecum does not contain enzymes able to degrade crystalline cellulose. However, other cellulosic substrates that are easier to digest, such as PASC, were degraded by the caecum contents, indicating how the enzymes here contained are responsible for the digestion of cellulose reported by the compositional analysis of wood versus frass; CMC was only partially degraded, but we cannot exclude the possibility that enzymes tightly bound to the wood have not been extracted during the sample preparation for the DNS reducing sugar assay.

The digestive glands have been reported to be the main organs where the endogenous wood-degrading enzymes are produced (Shipway, 2013). RNA sequencing of its tissues has shown the presence of numerous transcripts encoding CAZymes with a putative signal peptide for secretion (Shipway, 2013), while little evidence has so far been available on the reported ability of the “digestive” portion of the gland to perform intracellular wood digestion (Potts, 1923). By using TEM and light microscopy to visualise the digestive gland tissues, amoebic cells called phagocytes have been identified that could be responsible for intracellular wood digestion. These cells are contained in the “digestive” portion epithelium

but are also free to move in the gland lumen, and could engulf wood by the projection of pseudopodia. Phagocytes have been reported in the digestive system of various molluscs, in particular lamellibranchs (Yonge, 1937; George, 1952; Tripp, 1963), therefore intracellular food digestion is not a novelty. However, not only have phagocytes not been documented with micrographic evidence before for a shipworm, but intracellular digestion of wood – rather than other type of food – is uncommon, since it has been reported only for commensal ciliates found in the digestive system of termites (Kiuchi *et al.*, 2004). Despite being extremely interesting, these putative wood-degrading phagocytes are not fully understood, as it is not yet clear what contribution they give to the overall nutrient needs of the animals and how products resulting from the digestion of wood would be redistributed to the body.

The wood-degrading abilities of the digestive glands have also been studied by using a DNS reducing sugars assay, which showed that the glands are able to digest the same substrates as the caecum, though producing generally lower amount of reducing sugars. This indicates that the enzymes produced by the gland are possibly transported to the caecum, where they then accumulate, in conjunction with the bacterial ones, to digest of wood biomass.

The crystalline style is a digestive organ characteristic of bivalve molluscs, nonetheless it has not previously been studied in shipworm. It is known to help with extracellular digestion in various ways: it sorts the food coming from the mouth and distributes it appropriately to the various digestive organs on the basis of its size; it creates a cord of mucus-bound food that is drawn from the oesophagus to the stomach; it releases digestive enzymes by scraping against the gastric shield; it grinds food such as algae to break them open and make them more digestible; it lowers the pH of the stomach, making the food less viscous and optimising digestion (Purchon, 1971). The style is contained in an evagination of the stomach, the crystalline style sac, which was thought to produce the proteins the style is made of (Judd, 1979), though a cellulase produced by the digestive gland of the brackish water clam *Corbicula japonica* was recently found to be present in the style (Sakamoto *et al.*, 2008). In order to understand the style's protein composition, some styles were extracted from *L. pedicellatus* and visualised by SDS-PAGE electrophoresis, and the four main resulting bands were analysed by trypsinolysis and MALDI-TOF/TOF tandem mass spectroscopy. The results produced only a few matches, probably due to the limitation of the techniques that only identifies the most abundant proteins present in a

gel band. The bigger proteins, more than 300 kDa in size, were annotated as collagen, while cartilage was found in medium sized bands (100-150 kDa). These proteins appeared to be highly glycosylated when visualised with the Schiff's fuchsin-sulfite glycoprotein stain. No digestive enzymes were identified in the gel bands, possibly due to the low concentration of the protein. Other fainter bands are visible in the gel but their protein amount was below the limit of detection; it is possible that the digestive enzymes were contained in these bands, as their molecular weight is in the right range. Indeed, the enzymatic assays showed that the crystalline style of *L. pedicellatus* has the ability to degrade numerous polysaccharides, mainly mannans, glucans and xylans. An interesting protein was also isolated in the band with a molecular weight of 70 kDa, and therefore must be relatively abundant in the style. This was annotated as a membrane-attack complex/perforin-like protein, which are antimicrobial pore-forming proteins part of the defence system in vertebrates (Peitsch and Tschopp, 1991; Rosado *et al.*, 2008). A Blastx analysis revealed that this protein also contains an apextrin domain, which has been shown to allow pathogen recognition in other marine bivalves (Gerdol and Venier, 2015). It therefore is possible to hypothesise that the style might have a role in the recognition and disruption of bacterial cells in shipworms, however this subject will be discussed further in the next chapter, which will present the results of the full proteomic analysis of the crystalline style.

Chapter 4: Transcriptomic and proteomic analysis of the digestive organs of *Lyrodus pedicellatus*

4.1 Introduction

The Teredinids are well suited for the deconstruction of lignocellulose, which they use for nutritional purposes. They are important in the marine ecosystem because of their ability to recycle the carbon that would otherwise be locked in wood (Distel, 2003), and because they cause considerable damage to man-made wooden structures, resulting in substantial financial loss in some communities (Turner, 1966; Distel *et al.*, 2011). Furthermore, their array of lignocellulose degrading enzyme has the potential to be valuable in the bio-refinery industry for the mobilisation of sugars from lignocellulosic biomass (O'Connor *et al.*, 2014). Despite this, relatively little is known at present about the source and range of activities of the wood-degrading enzymes that shipworms are equipped with. As seen in previous chapters, shipworms harbour endosymbiotic bacteria in their gills, which are known to provide their hosts with lignocellulolytic enzymes to complement the ones that are endogenously produced. However, it remains unclear the degree to which shipworms depend on these bacterial enzymes. In the first instance, it is not clear the extent to which the bacteria participate to the production of digestive enzymes, and whether their presence is essential to the degradation of wood. Furthermore, it is unclear how the enzymes are distributed throughout the shipworm's digestive system and how the digestive task is shared between host and symbionts, since they both produce lignocellulolytic enzymes.

Honein and colleagues (2012) presented a short study on the transcriptomics of *Teredo navalis*, a species closely related to *L. pedicellatus*. Despite the publication only containing preliminary findings and never having been backed up with a more detailed analysis, the authors showed that the *T. navalis* transcriptome includes genes encoding for cellulose-degrading enzymes. In particular, they highlighted the presence of GHs from the families 1, 3, 9 and 45, which were annotated as endoglucanases or β-glucosidases. However, the RNA extraction was performed on the shipworm's whole body, therefore it was not possible to locate the origin and sites of enzymes production or to determine which are the major organs involved in wood digestion.

A study by O'Connor and colleagues (2014) made an important contribution by showing that the bacterial CAZymes found in the caecum of the shipworm *Bankia setacea* are produced by bacteria resident in the gills. Despite not being able to demonstrate the mechanism behind it, they also showed that while the gills bacterial proteome only contains 11% of CAZymes, the caecum bacterial proteome is almost entirely made up of these enzymes, which therefore must be selectively translocated from the gills to the caecum. According to the authors, the glycoside hydrolases produced by the endosymbiotic bacteria dominate the shipworm digestive system. However, the search of the proteomic data in the O'Connor paper was performed exclusively against the bacterial genomic DNA sequences derived from the gill tissues, with the consequent exclusion of endogenous proteins, therefore not giving a real indication of the relative abundance of endogenous versus bacterial enzymes.

A more detailed analysis of the transcriptome of a shipworm, with separate analysis of gills, digestive glands, caecum and intestine, was described in the PhD thesis of John Shipway (2013). This work aimed at identifying endogenous lignocellulolytic enzymes and their site of production in *L. pedicellatus*. The results were surprising, as the caecum appeared to have a very minor role in the transcription of endogenous GHs (only 0.7% of the total transcriptome), while the digestive glands were described as the major site of GHs production, with families 1, 9 and 10 being the most transcribed, followed by 2, 5, 13, 16, 18 and 45. The authors suggested that the lignocellulolytic enzymes are produced in the glands and secreted into the caecum, where wood digestion and absorption occur. However, this study only focused on the analysis of the endogenous enzymes, therefore not providing a full picture of the symbiosis between the shipworm and its endosymbionts. Furthermore, it relied only on one replica for the transcriptomic analysis, and the RNA extraction was performed with different techniques for different organs, with rRNA depletion carried out for the caecum and the gills, while poly(A) mRNA selection for the digestive gland, therefore excluding bacterial sequences from the analysis.

4.2 Aims of the chapter

The work described in this chapter aims to address the gap in the knowledge of the symbiosis between the shipworm *L. pedicellatus* and the endosymbionts resident in its gills, in relation to the production of wood-degrading enzymes. A detailed study of the shipworm

digestive process based around transcriptomic and proteomic data is described, with the aim of identifying the major genes and enzymes involved in wood digestion in different regions of the digestive system. The sources of the enzymes, both endogenous and bacterial, as well as their final destination are investigated. The studies will incorporate analysis performed in replicas of the transcriptome and proteome of gills, digestive glands, caecum and crystalline style and its sac and will encompass genes and enzymes produced by the animal as well as the symbiotic bacteria.

This detailed analysis will shed some light into this multifaceted symbiosis and on the mechanisms behind the combined system (eukaryote-prokaryotes) involved in the deconstruction of the recalcitrant lignocellulose. This analysis might also allow us to identify some novel lignocellulolytic enzymes with the potential to be utilised in the enzyme cocktail used in industries for the production of second-generation biofuels.

4.3 Transcriptomic analysis of digestive glands, caecum, gills and crystalline style sac

4.3.1 Sequencing of RNA from the digestive glands, caecum and gills

RNA was extracted from gills, digestive glands and caecum of three adult specimens of *L. pedicellatus*. Figure 4.1 and Appendix B show the results for the quality analysis performed with the Agilent 2200 Tape Station, which confirmed the RNA was of optimal quality (the RNA integrity number (RIN) ranged from 8.4 to 9.4), with no degradation and in sufficient amounts to proceed with RNA sequencing.

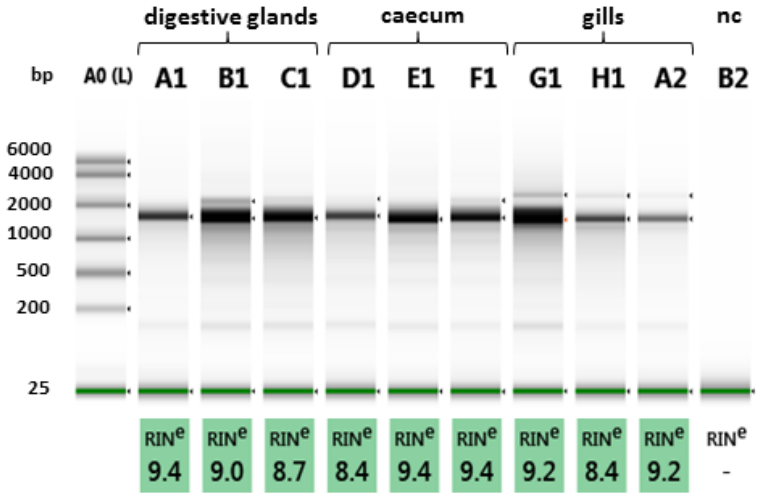


Figure 4.1. Agilent 2200 Tape station gel image showing the RNA extracted from the digestive glands, caecum and gills of three adult *L. pedicellatus*. RNA integrity numbers (RIN) are given at the bottom in the green rectangles. The first well on the left (A0 (L)) represents the ladder in base pairs (bp), while the last one on the right (B2) is the negative control (nc), which was distilled water.

The RNA was subjected to rRNA depletion, with consequent cleaning and concentration. Tape Station analysis (not shown) confirmed the depletion of the ribosomal RNA, a good distribution of mRNA sizes and gave RIN scores ranging from 1.5 to 3.2. RNA-sequencing libraries were then prepared for each mRNA sample and sequencing was performed using an IonTorrent sequencing platform. The results from the RNA sequencing are presented in Appendix A, which shows that 7,388,190 reads were generated for the gills, 7,526,229 for the digestive glands and 7,672,346 for the caecum. The mean read length was 187 base pairs for the gills, 183 for the digestive glands and 161 for the caecum.

4.3.2 Sequencing of RNA from crystalline style sac

The sequencing of RNA from the crystalline style was performed in a separate experiment and with a different technology than for the other digestive organs. The crystalline sac (without the style) was dissected from 38 adult specimens of *L. pedicellatus* and the RNA was isolated from all the samples pooled together. The Agilent 2200 Tape Station was used to quality check the RNA, indicating a RIN of 8.8, no degradation and sufficient concentration, therefore of excellent quality for sequencing (Figure 4.2). RNA sequencing was performed on an Illumina platform, after rRNA depletion and a final quality check (not shown) that confirmed the depletion of ribosomal RNA and the presence of the required amount of mRNA .

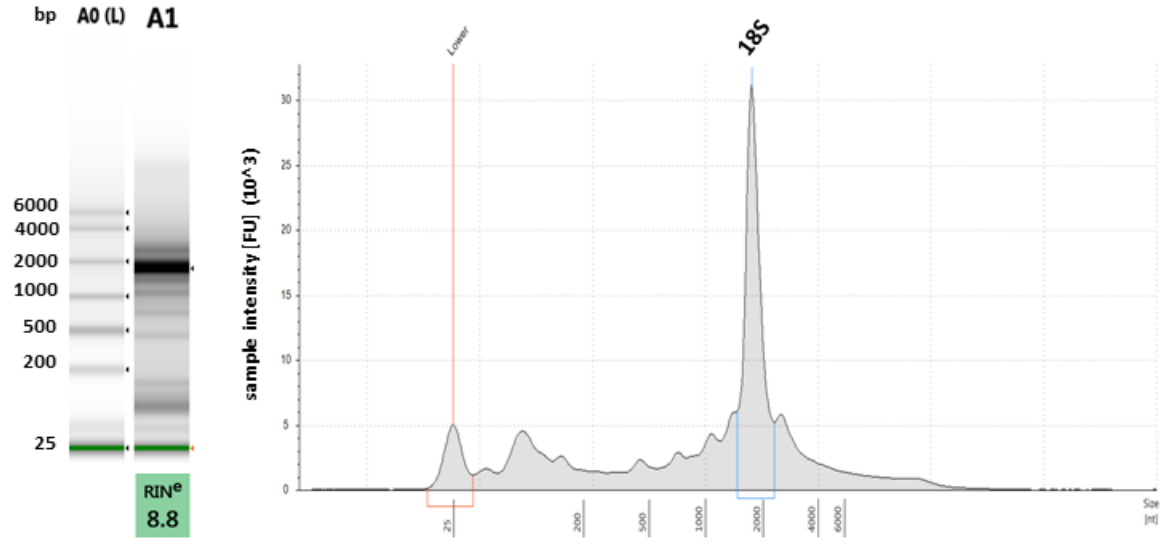


Figure 4.2. Agilent 2200 Tape station gel image (left) and electropherograms (right) showing the RNA extracted from the crystalline styles of 38 adult *L. pedicellatus*. The RNA integrity number (RIN) is given at the bottom of the gel image in the green rectangles. The first well on the left (A0 (L)) represents the ladder in base pairs (bp), while A1 is the 38 crystalline style samples pulled together.

The sequencing produced 308,957,623 reads for both ends, each read was 150 bases in length due to the Illumina technology sequencing technique. Quality control (see section 2.3.3 for details) was performed on the raw reads and the most meaningful results are shown in Appendix C. No sequences were flagged as being of low quality and both ends were almost free of rRNA (0.08% in end 1 and 0.17% in end 2). Excellent or good scores were reported for per base sequence quality, per sequences quality score, per sequence GC content and per base N content. Since per base sequence content and sequence duplication levels were of under optimal quality, trimming (see section 2.3.3 for details) was performed and only the trimmed reads were used for downstream analysis.

4.3.3. Analysis of the transcriptome of digestive glands, caecum, gills and crystalline style sac

The transcriptome of each organ was searched for CAZymes using the web server dbCAN and by sequence annotation against the NCBI non-redundant database (see material and methods for details, chapter 2.3.3). Only the most highly expressed 1,000 transcripts (calculated as transcripts per kilobase million, TPM) were analysed, representing 66.0% of the total transcriptome for the digestive glands, 66.4% for the caecum, 58.9% for the gills and 21.6% for the crystalline style sac. The CAZymes represent 0.9% of the most highly expressed 1,000 transcripts for the digestive glands, 0.5% for the caecum, 3.7% for the gills and 0.1% for the crystalline style sac, respectively. The different classes of CAZymes identified in each organ are represented in Figure 4.3, where bacterial enzymes are marked with an asterisk. Among the tissues analysed, only the digestive glands and caecum contain eukaryotic CAZyme transcripts in the most highly expressed 1,000 annotated genes, while the gills' transcriptome is comprised almost entirely of bacterial enzymes amongst the identified genes encoding for CAzymes (98.85%); the crystalline style sac includes only two CAZyme transcripts, both of bacterial origin.

The digestive glands CAZyme transcripts are dominated by glycosyl hydrolases, with just one carbohydrate esterase found, representing just 3.7% of the total CAZyme transcripts and belonging to the family 3. Among the GHs, the family 9 is the most abundant (35.3% of the total CAZymes), followed by family 45 (23.1%), family 1 (19.2%) and family 2 (11.3%). The other GHs are found in minor quantities and include families 18, 10, 13, 84 and 71 (together representing 7.5% of the total CAZymes).

Glycosyl hydrolases from family 30 instead represent the greater part (63.2%) of the CAZyme transcripts of the caecum, and are exclusive to this organ. All the other GHs are also found in the digestive glands and belong to the classes 1 (19.2%), 13 (5.1%), 84 (3.0%) and 18 (2.5%). As in the digestive glands, only one CE from family 3 was identified (7.0%).

The CAZyme transcripts range and composition is different in the gills' transcriptome. In addition to the various GHs (66.9% of the total), other enzymes groups were also found, namely CEs of the families 15 (1.9%) and 3 (0.5%) and AAs from family 10 (30.8%). The most abundant CAZyme transcripts are GH5s (41.5%), followed by AA10s (30.8%), GH6s (17.5%) and GH11s (6.7%), all of which are exclusive to the gills. Other CAZymes found in minor quantities are GH84, GH13, GH16 and GH10, which represent only 1.1% of the total.

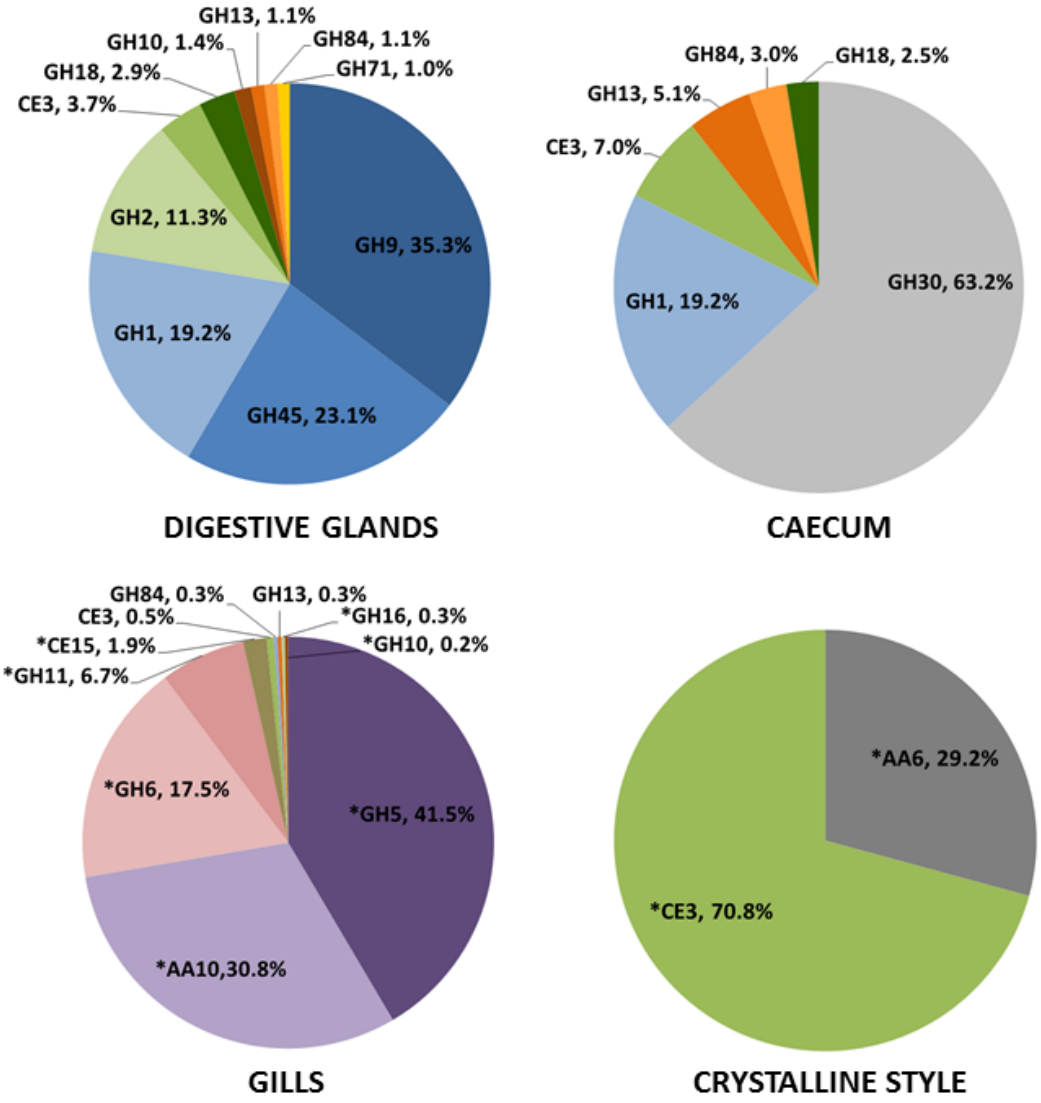


Figure 4.3. CAZymes classes identified in the transcripts of the digestive glands, caecum, gills and crystalline style sac of *L. pedicellatus*. The pie charts were obtained by searching each organ best 1,000 transcripts (calculated as TPM values) for CAZymes using dbCAN and then calculating their relative abundance. The charts do not include glycosyl transferases or transcripts containing a CBM but no other identified CAZymes. Transcripts encoding bacterial enzymes are marked with an asterisk (*).

Tables 4.1 to 4.4 list all the sequences predicted to encode for a CAZyme found in the most highly expressed 1,000 transcripts in each of the four organs analysed. For each enzyme, the transcript abundance is shown, as well as the percentage of the overall transcriptome that it represents. The tables also report the CAZy families and subfamilies and their relative e-values, as well as the transcript annotation and the species and higher taxonomical group it belongs to.

Table 4.1. Digestive gland transcripts encoding for CAZymes among the 1,000 most expressed transcripts. The table lists the name of the transcript (contig name), its abundance (TPM), its relative abundance in the overall transcriptome (TPM%), the CAZy modules (with subfamily after the underscore) identified in the transcript (CAZy modules) and their e-values (CAZy e-value), the annotation (annotation) and species (species) obtained by blasting against the NCBI non-redundant database, and whether the transcript is of eukaryotic or prokaryotic origin (domain).

Contig name	TPM	TPM %	CAZy modules	CAZy e-value	Annotation	Species	Domain
c151620_g1_i1	1820.4950	0.1820	GH9	2.7E-31	endoglucanase 13-like	*Mizuhopecten yessoensis*	eukaryotes
c86197_g1_i1	1247.1570	0.1247	GH45	5.00E-43	endoglucanase-like	*Crassostrea gigas*	eukaryotes
c178040_g2_i1	926.0908	0.0926	GH1	9.9E-161	lactase-phlorizin hydrolase-like isoform X1	*Lingula anatina*	eukaryotes
c166630_g2_i1	429.1136	0.0429	GH2	1.3E-10	glucosidase 2 subunit beta	*Exaiptasia pallida*	eukaryotes
c167653_g1_i1	242.5884	0.0243	GH2	4.00E-58	beta-mannosidase-like	*Crassostrea virginica*	eukaryotes
c181797_g2_i1	217.3767	0.0217	CE3	2.6E-11	leucine-rich repeat-containing protein DDB_G0290503-like	*Saccoglossus kowalevskii*	eukaryotes
c171242_g3_i1	174.9765	0.0175	GH18	2.7E-83	chitinase 3	*Crassostrea gigas*	eukaryotes
c175047_g4_i1	161.3171	0.0161	GH1	1.6E-25	lactase-like protein isoform X4	*Cynoglossus semilaevis*	eukaryotes
c177973_g3_i1	126.1040	0.0126	GH45	1.00E-47	endo-1,4-beta-glucanase	*Corbicula japonica*	eukaryotes
c156925_g1_i1	99.5883	0.0100	GH9	1.1E-27	cellulase	*Corbicula japonica*	eukaryotes
c176344_g2_i3	90.1000	0.0090	GH9	2.2E-42	endoglucanase 13-like	*Mizuhopecten yessoensis*	eukaryotes
c176344_g1_i3	89.9701	0.0090	GH9	3.4E-75	cellulase	*Corbicula japonica*	eukaryotes
c180291_g3_i2	80.3690	0.0080	GH10	2.00E-35	endo-1,4-beta-xylanase	*Corbicula japonica*	eukaryotes
c179335_g1_i1	63.7116	0.0064	GH13_17	4E-76	probable maltase	*Aplysia californica*	eukaryotes
c180935_g1_i1	62.9453	0.0063	GH84	6E-69	protein O-GlcNAcase	*Crassostrea gigas*	eukaryotes
c175836_g1_i3	57.8701	0.0058	GH71	1.1E-10	glucan endo-1,3-alpha-glucosidase agn1	*Aspergillus udagawae*	eukaryotes
c178040_g1_i2	53.0328	0.0053	GH1+GH1	9.4E-139	lactase-phlorizin hydrolase-like, partial	*Mizuhopecten yessoensis*	eukaryotes

Table 4.2. Caecum transcripts encodeng for CAZymes among the 1,000 most expressed ones. The table fields are the same as in Table 4.1.

Contig name	TPM	TPM %	CAZy modules	CAZy e-value	Annotation	Species	Domain
c153913_g1_i1	1501.1919	0.2008	GH30	2.00E-31	endo-1,6-beta-D-glucanase BGN16.3	*Crassostrea gigas*	eukaryotes
c181813_g3_i4	143.9217	0.0192	GH1	2.1E-127	lactase-phlorizin hydrolase-like, partial	*Mizuhopecten yessoensis*	eukaryotes
c177035_g4_i2	135.2853	0.0181	GH1	6.2E-19	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c181797_g2_i1	132.1192	0.0177	CE3	2.6E-11	leucine-rich repeat-containing protein DDB_G0290503-like	*Saccoglossus kowalevskii*	eukaryotes
c179335_g1_i1	131.0636	0.0175	GH13_17	4E-76	probable maltase	*Mizuhopecten yessoensis*	eukaryotes
c166743_g1_i1	125.9803	0.0168	GH30	6.6E-41	endo-1,6-beta-D-glucanase BGN16.3-like isoform X1	*Crassostrea virginica*	eukaryotes
c181813_g4_i1	114.2455	0.0153	GH1	2.1E-127	lactase-phlorizin hydrolase-like, partial	*Mizuhopecten yessoensis*	eukaryotes
c182208_g1_i1	102.1336	0.0137	GH1+GH1	1.4E-157	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c180935_g1_i1	77.2311	0.0103	GH84	6E-69	protein O-GlcNAcase	*Crassostrea gigas*	eukaryotes
c171242_g3_i1	65.1673	0.0087	GH18	2.7E-83	chitinase 3	*Crassostrea gigas*	eukaryotes
c182129_g2_i2	47.8331	0.0064	CE3	1.3E-10	putative RNA-directed DNA polymerase from transposon BS	*Exaiptasia pallida*	eukaryotes

Table 4.3. Crystalline style transcripts encoding for CAZymes among the 1,000 most expressed ones. The table fields are the same as in Table 4.1.

Contig name	TPM	TPM %	CAZy modules	CAZy e-value	Annotation	Species	Domain
DN361801_c10_g1_i3	142.1543	0.0142	CE3	1.40E-12	lipase	*Fortiea contorta*	prokaryotes
DN359415_c2_g1_i3	58.7367	0.0059	AA6	1.10E-29	flavodoxin family protein, NADPH-dependent FMN reductase	*Xanthomonas translucens*	prokaryotes

Table 4.4. Gills transcripts encoding for CAZymes among the 1,000 most expressed ones. The table fields are the same as in Table 4.1.

Contig name	TPM	TPM %	CAZy modules	CAZy e-value	Annotation	Species	Domain
c163430_g1_i1	2208.0614	0.4892	AA10+CBM10	8.00E-10	auxiliary activity family 10 domain-containing protein	*Alteromonadaceae bacterium Bs02*	prokaryotes
c177477_g5_i1	1885.8114	0.4178	GH5_53	1.00E-10	glycosyl hydrolase family 5_53 domain-containing protein	*Alteromonadaceae bacterium Bs02*	prokaryotes
c176007_g1_i1	1241.6663	0.2751	GH6	1.4E-41	cellobiohydrolase	*Alteromonadales bacterium BS08*	prokaryotes
c173837_g2_i2	740.5930	0.1641	AA10+CBM10	6.1E-37	chitin-binding protein	*Alteromonadaceae bacterium Bs12*	prokaryotes
c176917_g2_i1	672.9781	0.1491	GH5_53	3.1E-161	endoglucanase	*Cellvibrio sp. PSBB006*	prokaryotes
c121433_g1_i1	518.7870	0.1149	GH5+CBM2	7.00E-27	beta-mannosidase	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c160576_g1_i1	451.3260	0.1000	GH5+CBM10	8.00E-19	glycoside hydrolase family 5	*Teredinibacter turnerae*	prokaryotes
c176007_g2_i1	385.1313	0.0853	GH6	1.4E-41	cellobiohydrolase	*Microbulbifer thermotolerans*	prokaryotes
c169869_g1_i2	347.8859	0.0771	GH11	3.60E-36	1,4-beta-xylanase	*Micromonospora rifamycinica*	prokaryotes
c147988_g1_i2	237.4276	0.0526	GH11	3.00E-10	xylan 1,4-beta-xylosidase	*Streptomyces castelarensis*	prokaryotes
c177001_g2_i2	188.0658	0.0417	GH5_8+CBM10	2.1E-34	beta-mannosidase	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c165047_g1_i1	185.1705	0.0410	CE15	2.50E-27	carbohydrate esterase family 15 protein	*Piromyces sp. E2*	prokaryotes
c176379_g1_i2	101.2869	0.0224	GH5_2	1.7E-27	cellulase	*Pseudomonas sp. ND137*	prokaryotes
c178777_g2_i2	90.3214	0.0200	GH5_53	8.00E-43	glycosyl hydrolase family 5_53 domain-containing protein	*Alteromonadaceae bacterium Bs02*	prokaryotes
c178964_g4_i6	76.2346	0.0169	GH5+CBM10	6.00E-12	glycoside hydrolase family 5	*Teredinibacter sp. 991H.S.0a.06*	prokaryotes
c167683_g1_i2	76.0678	0.0169	GH6	9.00E-44	cellobiohydrolase	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c174327_g1_i1	68.5166	0.0152	AA10	3.1E-16	chitin-binding protein	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c177001_g3_i1	65.4455	0.0145	GH5+CBM10	6.4E-16	beta-mannosidase	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c181797_g2_i1	53.5174	0.0119	CE3	3.2E-11	leucine-rich repeat-containing protein DDB_G0290503-like	*Saccoglossus kowalevskii*	eukaryotes
c180189_g2_i1	47.8078	0.1059	AA10	1.2E-16	chitin-binding protein	*Teredinibacter turnerae*	prokaryotes
c169869_g1_i1	45.4726	0.0101	GH11+CBM10+CBM5	1.70E-14	1,4-beta-xylanase	*Cellvibrio sp. OA-2007*	prokaryotes
c164553_g1_i4	43.3619	0.0096	GH5_53	6.00E-52	endoglucanase	*Cellvibrio sp. PSBB006*	prokaryotes
c165315_g1_i2	36.7061	0.0081	GH5_8+CBM10+CBM2	3.00E-20	beta-mannosidase	*Alteromonadaceae bacterium Bs12*	prokaryotes
c180348_g2_i1	35.5616	0.0079	GH11	9.00E-40	1,4-beta-xylanase	*Cellvibrio sp. OA-2007*	prokaryotes
c180241_g2_i2	33.7378	0.0075	GH6	9E-10	glycoside hydrolase family 53	*Teredinibacter turnerae*	prokaryotes
c180935_g1_i1	31.2354	0.0069	GH84	6E-69	protein O-GlcNAcase	*Crassostrea gigas*	eukaryotes
c179335_g1_i1	29.8305	0.0066	GH13_17	4.5E-76	probable maltase	*Aplysia californica*	eukaryotes
c177241_g2_i1	27.1357	0.0060	GH16+CBM2	2.1E-35	endoglucanase	*Teredinibacter sp. 1162T.S.0a.05*	prokaryotes
c154562_g1_i1	24.9515	0.0055	GH10	1.5E-20	acetylxylan esterase	*Teredinibacter turnerae*	prokaryotes

The digestive glands most expressed 1,000 transcripts include a total of 17 sequences encoding for CAZymes. GH9, GH45, GH1 and GH2 modules are found in more than one transcript, and indeed these families are the most abundant CAZymes in the transcriptome. The other CAZy modules are only found in one transcript each. All the GH1s are annotated as "lactase-phlorizin hydrolase-like", proteins produced by the digestive glands of some bivalve molluscs, which have been shown to act as β-glucosidases, hydrolysing cellobiose into free glucose units (Sakamoto *et al.*, 2009). Most of the GH transcripts are annotated as possible endoglucanases, β-glucosidases or β-mannosidases, with only four contigs annotated differently as either endo-xylanases, chitinases or putative maltases.

Eleven different sequences encoding for CAZymes were identified in the caecum most expressed 1,000 transcripts. Some of the sequences are identical to those identified in the digestive glands (CE3 c181797_g2_i1, GH13_17 c179335_g1_i1, GH84 c180935_g1_i1 and GH18 c171242_g3_i1), and, therefore, there is the possibility that they could be contaminants, given that one of the digestive glands lies in close contact with the caecum. On the contrary, all the transcripts encoding for GH30s and GH1s (and one for the CE3 c182129_g2_i2) are only found in the caecum and in high amounts, suggesting that they are truly expressed in this organ. As in the digestive glands, most of the contigs are annotated as endoglucanases or β-glucosidases. Contig c182129_g2_i2 is annotated as a "putative RNA-directed DNA polymerase from transposon BS", however dbCAN finds similarity to a CE3 domain with a reasonable e-value (1.3e-10); it is possible though that this transcript is not actually a CAZyme.

The 1,000 most expressed transcripts in the gills include the highest number of CAZymes sequences, with a total of 29. All the bacterial transcripts encode for CAZymes that are unique to the gills, while the three eukaryotic ones (the CE3 c181797_g2_i1, the GH13_17 c179335_g1_i1 and the GH84 c180935_g1_i1) are produced in low amounts and are also found in the digestive glands and caecum, suggesting they may have a cellular function rather than a digestive one. Numerous contigs transcribe for either GH5s (11 transcripts), AA10s, GH6s or GH11s (four different transcripts each), which corresponds to the most highly abundant CAZymes. The annotations of the gill transcripts are different and more varied compared to the digestive glands and caecum. Indeed, while some endoglucanases and β-mannosidases are present, the majority of the contigs encode putative cellobiohydrolases, β-xylanases, cellodextrinases, lytic polysaccharide monooxygenases

and carbohydrate esterases. A putative acetylxylan esterase is also present, as well as a possible maltase; no β-glucosidases were identified.
Only two CAZymes transcripts were found in the crystalline style sac best 1,000 transcripts. These are both bacterial and represents only 0.0142% (DN361801_c10_g1_i3, a CE3) and 0.0059% (DN359415_c2_g1_i3, an AA6) of the total transcriptome.

Figure 4.4 shows the presence and abundance of each single CAZyme transcript in the four organs whose transcriptome was analysed. It clearly illustrates how the gills are specialised for the transcription of enzymes mainly from the families AA10, GH5, GH6 and GH11, while the digestive glands produces GH9s, GH45s, GH1s, GH2s and some GH10s. The caecum mainly transcribes GH30s and GH1s, this last family containing different transcripts from those found in the digestive glands. The crystalline style sac generally lacks CAZyme transcripts except for the AA6 and a CE3, which likely come from contamination with bacterial RNA. Only four transcripts are transcribed in all the organs (excluding the style sac), and include a possible maltase (GH13_17), a putative chitinase (GH18), a N-acetyl β-glucosaminidase (GH84) and a presumed carbohydrate esterase (CE3). The figure also shows the range of different subfamilies of GH5 that are produced in the gills. Subfamilies 53 and 8 are found in different transcripts, and subfamily 2 is also present, as well as some contigs for which a specific subfamily was not possible to determine. GH5s belonging to the subfamily 2 are known to typically have an endo-β-1,4-glucanase activity, while those from subfamily 8 are endo-β-1,4-mannanase and from subfamily 53 are normally cellodextrinases (Aspeborg *et al.*, 2012).

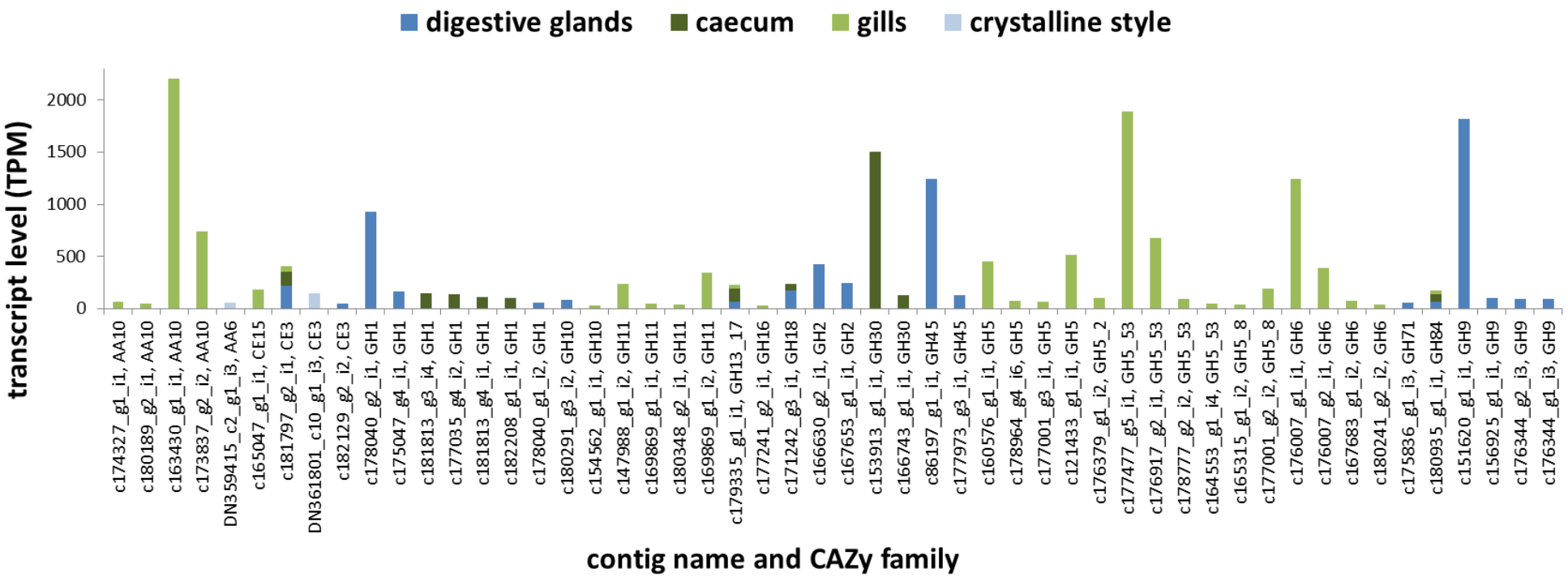


Figure 4.4. Transcript levels (in TPM) of the contigs with a CAZy domain identified in the 1,000 most expressed transcripts of the digestive glands, caecum, gills and crystalline style sac.

4.4 The role of the caecum in sugar transport

The caecum is an organ that among Teredinids is exclusively found in those that bore into wood. It is known to store wood particles, and various observations suggest that is may also be responsible for wood digestion and the absorption of the breakdown products (Bazylinski and Rosenberg, 1983; Betcher *et al.*, 2012). The evidence mainly comes from anatomical structures (extensive vascularisation, presence of a ciliated epithelium and of the typhlosole) together with some biochemical characterisation that shows the presence of wood-degrading enzymes, as discussed in detail in chapter 3.4.1. In this section, further evidence from the analysis of the shipworm's transcriptome is presented. The transport of glucose through the cytoplasmic membrane can be passive by diffusion, or requiring energy; in both cases carrier proteins are normally involved (Hanquet *et al.*, 2011). In eukaryotes, including invertebrates, they are classed into two different groups, those that are sodium-dependent (Wright, 2001) and those that are sodium-independent (Joost and Thorens, 2001). The *L. pedicellatus* most highly expressed 1,000 transcripts from the four organs were analysed for sequences annotated as sugar transporters. The search resulted in four different transcripts (listed in Table 4.5), but other less expressed ones were also identified in the most highly expressed 10,000 transcripts (data not presented). They all are annotated as glucose transporters, with two of them being sodium dependent and two sodium-independent, with small e-values and high similarity to transcripts found in other marine or freshwater bivalves or gastropods.

Table 4.5. Transcripts annotated as putative sugar transporters identified in the transcriptome of *L. pedicellatus.* The table lists the name of the transcript (contig name), its length (seq length), the annotation (annotation) with its e-value (annot e-value) and species (species) obtained by blasting against the NCBI non-redundant database, and whether the transcript is of eukaryotic or prokaryotic origin (domain).

Contig name	Contig length	Annotation	Annot e-value	Species	Domain
c177599_g1_i1	4235	solute carrier family 2, facilitated glucose transporter member 1-like isoform X1	2.00E-176	*Biomphalaria glabrata*	Eukaryotes
c168760_g1_i2	1928	sodium-dependent glucose transporter 1B-like	2.00E-81	*Mizuhopecten yessoensis*	Eukaryotes
c174944_g1_i2	2496	solute carrier family 2, facilitated glucose transporter member 1-like	0.0	*Aplysia californica*	Eukaryotes
c180106_g2_i1	1915	sodium-dependent glucose transporter 1B-like	6.00E-66	*Crassostrea virginica*	Eukaryotes

The expression levels of the transcripts listed in Table 4.5 are shown in Figure 4.5, which shows that the caecum is the main organ involved in the transcription of sugar transporters.

Three different sugar transporters are expressed in its tissues, one of which (c177599_g1_i1) is also found in the digestive gland transcriptome, where it could be linked to the presence of the wood-degrading phagocytes found in the white portion of the gland. A fourth transcript is instead found in the gills transcriptome and could be involved in the uptake of sugar from sea water (Welborn and Manahan, 1990). The protein encoded by c177599_g1_i1 is also found in the proteome of the caecum (and in the digestive glands proteome, but only among those proteins with just one significant match).

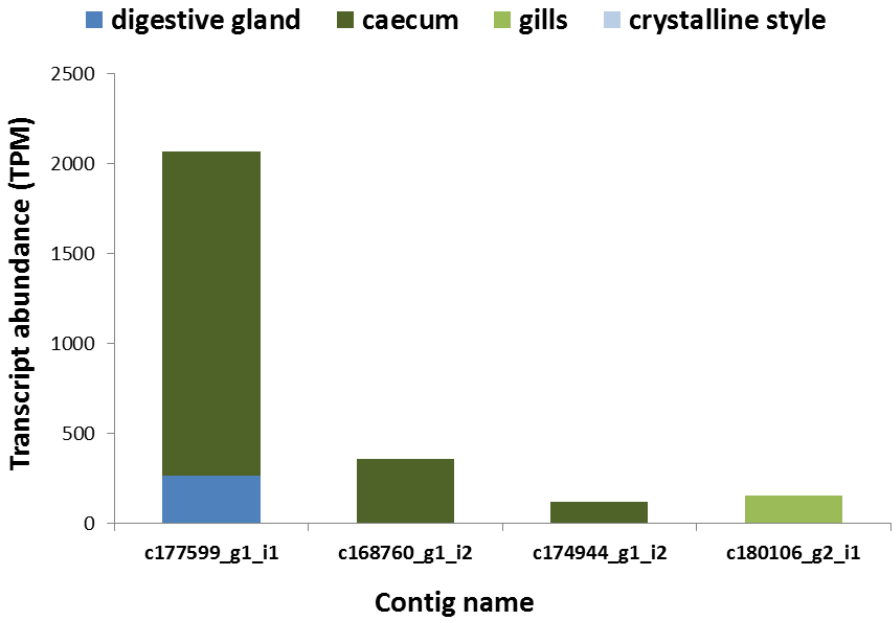


Figure 4.5. Expression levels of the putative glucose transporters identified in the transcriptome of *L. pedicellatus*. The graph shows the transcript abundance of the four contigs annotated as putative glucose transporters (listed in table 4.5) found in the best 1,000 transcripts of the digestive gland, caecum, gills and crystalline style.

4.5 Analysis of the proteome of digestive glands, caecum, gills and crystalline style

The protein content of gills, digestive glands, caecum fluids and crystalline style was analysed using mass spectrometry-based shotgun proteomics (details in materials and methods, section 2.4). Five adult *L. pedicellatus* were dissected for this purpose, and proteins from their organs were pooled together by type before trypsinolysis, to avoid individual variability that could skew the results. For the crystalline style, 21 different ones were collected from the same number of specimens and pooled together, due to the small size of the organ. Mascot searches were preformed against the *L. pedicellatus* transcriptome of digestive glands, gills, caecum and crystalline style sac. The resulting

matches were searched for CAZymes using the web server dbCAN and by sequence annotation against the NCBI non-redundant database. The analysis resulted in 1,455 proteins with more than one significant match identified for the digestive glands, 761 for the caecum, 687 for the gills and 69 for the crystalline style. The CAZymes identified among these proteins are shown, grouped by families, in Figure 4.6, where the presence of a bacterial enzyme is marked with an asterisk (*), while a carat (^) represents mixed bacterial and eukaryotic ones.

The digestive glands contain almost entirely endogenous proteins, with only 0.4% being bacterial; among the CAZymes (which represent 2.9% of the proteome), the percentage of bacterial CAZymes rises to 4.1%. The most abundant CAZy class is the GH1 (18.8%), closely followed by GH9 (15.3%) and GH5 (11.7%). This last class is the only one, together with the GH10, to contain some bacterial proteins. GH59 (6.8%), GH29 (5.6%) and GH18 (5.4%) are the following classes present in significant amounts, while all the other classes represent less than 4% each of the total of CAZymes. This only partially reflects the results of the transcriptomics, where GH1 and GH9 are also very abundant; however, the proteomics contains very small amount of GH45 and GH2, which were found to be abundant in the transcriptome. GH13, GH84 and GH71 families are also not identified in the proteomics data, while they are present in the transcriptome.

The proteomics of the caecum fluids contains 4.8% of prokaryotic proteins, and the CAZymes are 18.2% of the total. These CAZymes are a mixture of endogenous and prokaryotic proteins, with the former being the most abundant, representing 87.6% of the total CAZymes, compared to 12.4% for the bacterial ones. The GH1 family dominates the CAZy caecum proteome, with 31.1%, followed by the GH9 (21.3%) and GH5 (10.0%); these are also the most abundant families found in the digestive glands. The caecum proteome also contains the families GH2 (8.3%), GH134 (4.2%) and GH13 (3.9%). All the other families are found to be less than 4% each. Some CEs (from the families 1, 3, 6, 10, 12 and 15) and AAs (families 4, 7 and 10) have been identified in the proteome, but their amount is so small that they are not possible to visualise in the pie chart and they are grouped among those proteins representing less than 1% of the CAZymes. The CAZy abundance of the caecum does not closely reflect the transcriptome, where the GH30s were the most abundant class; GH84 and GH18 are also absent, while the only classes that are found both in the transcriptomics and the proteomics are the GH1, the GH13 and the CE3.

The majority of the gill proteome comprises eukaryotic proteins (90.3%, versus 9.7% prokaryotic). The CAZymes represent 1.23% of the total proteins, with 73.2% of them being from the shipworm (26.8% prokaryotic). The GH5 class, including only proteins of bacterial origin, dominates the proteome (26.8%), together with AA4s (24.1%), which on the contrary are endogenous and are annotated as vanillyl-alcohol oxidases. The rest of the CAZymes are also eukaryotic, including the families GH59 (12.3%), GH29 (8.0%), GH9 (6.5%), GH38 (6.1%), GH3 (5.4%), CE10 (5.0%), GH20 (3.4%) and GH31 (2.3%). The low amounts of prokaryotic CAZymes is surprising, but this result is probably due to procedural problems in the preparation of the samples (inefficient breakage of the bacterial cell walls) rather than being representative of the real protein composition. This probably reflects on the different classes found in the gills' transcriptome, given that the GH5 is the only family found in both proteome and transcriptome.

The crystalline style proteome is comprised mostly of endogenous proteins (92.4% versus 7.6% prokaryotic), while the CAZymes originate almost equally from the bacteria (54.2%) and the shipworm (45.8%). The style proteome contains seven different CAZymes families, among which the GH5 is by far the most abundant GH class (41%) and contains both eukaryotic and prokaryotic enzymes, although the former are three times more abundant. Endogenous GH9 constitute 25.0% of the CAZymes and prokaryotic GH134 mannanases 19.0%. Other classes found are GH16 (5.4%), CE3 (5.1%), GH35 (3.6%) and GH1 (1.2%), of which only the CE3 is bacterial. The crystalline style sac produces only one of the CAZyme class found in the style proteome (the CE3), which is found only in very small quantities among these enzymes (5.1% of the CAZymes in the proteome in contrast to 70.8% in the transcriptome) and could be due to bacterial contamination.

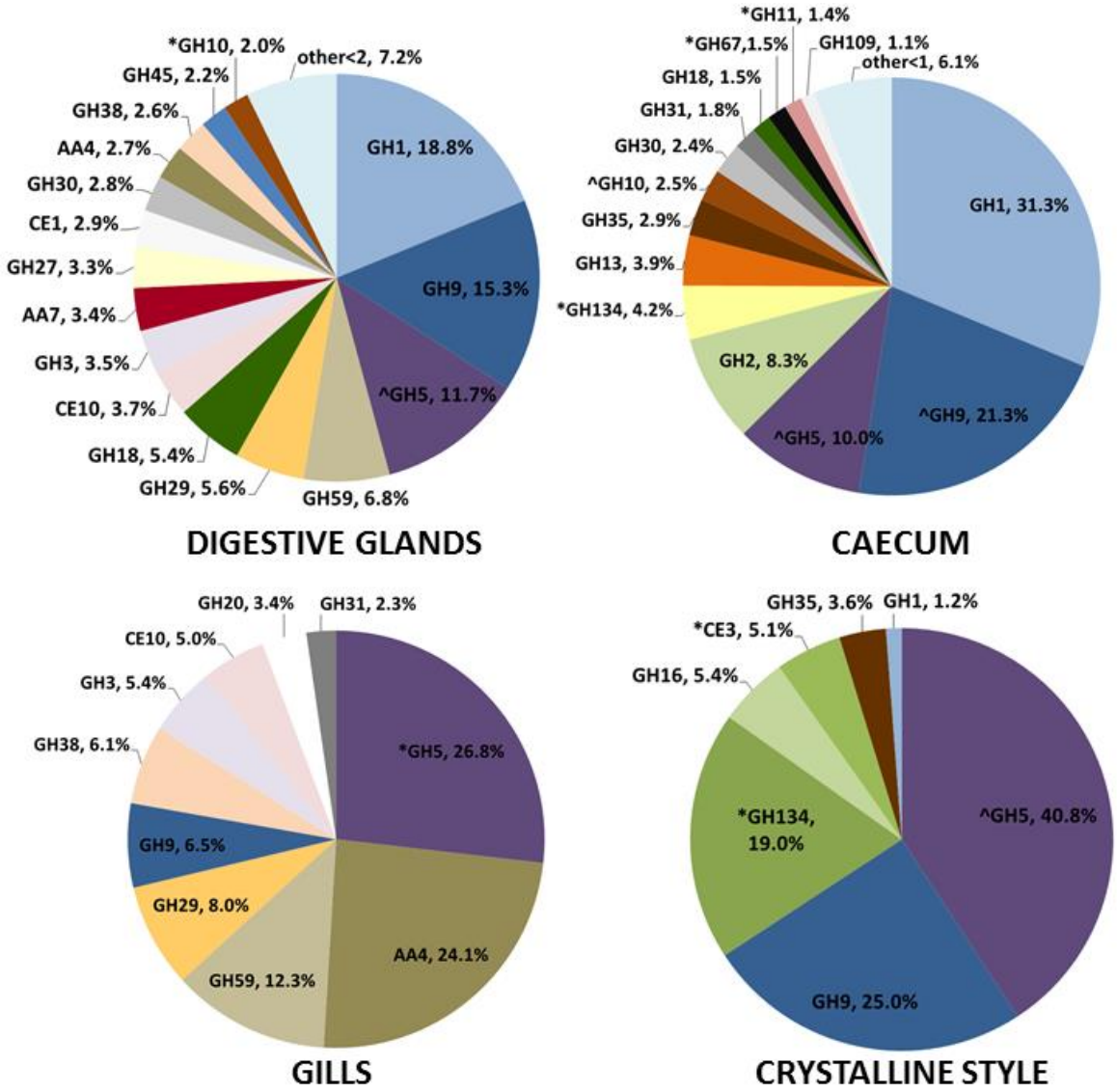


Figure 4.6. CAZyme classes identified in the proteomic analysis of the digestive glands, caecum, gills and crystalline style sac of *L. pedicellatus*. The pie charts were obtained by searching for CAZymes (using dbCAN) each organ's proteins with 2 or more significant matches against the transcriptome of gills, digestive glands, caecum and crystalline style sac, and then calculating their relative abundance. The charts do not include glycosyl transferases or proteins containing a CBM but no other identified CAZymes. Protein classes from bacterial enzymes are marked with an asterisk (*) and those that contain both bacterial and eukaryotic proteins with a carat (^).

Tables 4.6 to 4.9 list all the proteins containing at least one CAZy domain identified in the proteomics of the four organs analysed. For each protein, abundance is shown calculated as emPAI score (exponentially modified Protein Abundance Index (Ishihama *et al.*, 2005)), as well as the percentage of the overall proteome that it represents. The tables also report the CAZy families and subfamilies included in the protein and their relative e-values, as well as the protein annotation and the species and higher taxonomical group it belongs to.

Table 4.6. CAZymes of the digestive glands. The table shows the results for the Mascot search of the LC-MS/MS analysis performed on *L. pedicellatus* digestive glands against the transcriptome of gills, digestive glands, caecum and crystalline style sac. Only proteins with two or more significant matches are included. The table lists the name of the protein (contig name), it abundance calculated as emPAI score (emPAI score), its relative abundance among those with two or more significant matches (emPAI score %), the CAZy modules (with subfamily after the underscore) identified in the transcript (CAZy modules) and their e-values (CAZy e-value), the annotation (annotation) and species (species) obtained by blasting against the NCBI non-redundant database, and whether the transcript is of eukaryotic or prokaryotic origin (domain).

Contig name	emPAI score	emPAi score %	CAZy modules	CAZy e-value	Annotation	Species	Domain
c165315_g1_i2	1.22	0.22	GH5_8+CBM10+CBM2	3E-20	GH family 5_8 domain-containing protein	*Alteromonadaceae bacterium* Bs08	prokaryotes
c165524_g1_i1	0.74	0.13	GH9	2.2E-32	cytoplasmic aconitate hydratase-like isoform X1	*Crassostrea virginica*	eukaryotes
c176681_g1_i13	0.66	0.12	GH59	6.6E-136	galactocerebrosidase-like isoform X2	*Mizuhopecten yessoensis*	eukaryotes
c180475_g2_i5	0.56	0.10	GH29	1.5E-47	alpha-L-fucosidase	*Crassostrea gigas*	eukaryotes
c173596_g2_i1	0.55	0.10	AA7	6E-13	alkyldihydroxyacetonephosphate synthase	*Crassostrea gigas*	eukaryotes
c169024_g2_i1	0.55	0.10	GH5_10	2.4E-101	mannan endo-1,4-beta-mannosidase	*Mizuhopecten yessoensis*	eukaryotes
c175047_g4_i1	0.52	0.09	GH1	1.7E-25	lactase-like protein isoform X4	*Cynoglossus semilaevis*	eukaryotes
c181452_g1_i2	0.47	0.08	CE1	1.5E-52	S-formylglutathione hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c178040_g2_i1	0.47	0.08	GH1	1.1E-160	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c174752_g6_i1	0.44	0.08	AA4	2.6E-23	glyoxylate reductase/hydroxypyruvate reductase	*Ceratitis capitata*	eukaryotes
c175967_g4_i7	0.44	0.08	GH59	4.2E-112	galactocerebrosidase-like isoform X2	*Mizuhopecten yessoensis*	eukaryotes
c180131_g1_i2	0.43	0.08	GH1	9.7E-99	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c176241_g1_i1	0.42	0.08	GH3	2.5E-30	beta-D-xylosidase 5	*Mizuhopecten yessoensis*	eukaryotes
c176477_g1_i1	0.42	0.08	GH38+GH38	2.4E-46	lysosomal alpha-mannosidase	*Crassostrea gigas*	eukaryotes
c178040_g3_i1	0.38	0.07	GH1+GH1+GH1	6.1E-160	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c156925_g1_i1	0.38	0.07	GH9	2.4E-38	beta-1,4-glucanase 3	*Polyphaga aegyptiaca*	eukaryotes
c86197_g1_i1	0.35	0.06	GH45	5.00E-43	endoglucanase-like	*Crassostrea gigas*	eukaryotes
c177483_g5_i1	0.34	0.06	GH29	9.4E-59	alpha-L-fucosidase	*Strongylocentrotus purpuratus*	eukaryotes
c180385_g4_i3	0.33	0.06	GH1	1.1E-65	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c170561_g7_i1	0.33	0.06	GH10+CBM60	1.4E-28	GH family 10 domain-containing protein	*Alteromonadaceae bacterium* Bs02	prokaryotes
c171242_g3_i1	0.31	0.06	GH18	2.8E-83	chitinase 3	*Crassostrea gigas*	eukaryotes
c177931_g2_i1	0.31	0.06	GH18	3.7E-35	acidic mammalian chitinase	*Crassostrea gigas*	eukaryotes
c181365_g1_i1	0.30	0.05	GH27+CBM13	1.1E-129	N-acetylgalactosaminyltransferase 5-like isof. X1	*Crassostrea virginica*	eukaryotes
c168511_g1_i1	0.30	0.05	GH9	6.7E-109	cellulase	*Corbicula japonica*	eukaryotes
c178993_g2_i3	0.28	0.05	CE10	3.4E-22	S9 family peptidase	*Moorea producens*	eukaryotes
c176344_g1_i4	0.28	0.05	GH9	2E-123	cellulase	*Corbicula japonica*	eukaryotes
c178089_g2_i13	0.27	0.05	GH2	2.5E-79	beta-mannosidase-like	*Crassostrea virginica*	eukaryotes
c153913_g1_i1	0.27	0.05	GH30	7.7E-41	endo-1,6-beta-D-glucanase BGN16.3	*Crassostrea gigas*	eukaryotes
c178040_g1_i1	0.25	0.04	GH1+GH1	1E-138	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c176294_g3_i1	0.25	0.04	GH18	4.4E-37	di-N-acetylchitobiase-like	*Crassostrea virginica*	eukaryotes
c151620_g1_i1	0.24	0.04	GH9	7.5E-34	endoglucanase 13-like	*Mizuhopecten yessoensis*	eukaryotes

c179852_g6_i1	0.23	0.04	CE10	3.4E-36	arylacetamide deacetylase-like	*Crassostrea virginica*	eukaryotes
c180568_g1_i1	0.23	0.04	GH27	1.1E-22	alpha-N-acetylgalactosaminidase	*Crassostrea gigas*	eukaryotes
c173097_g2_i4	0.22	0.04	GH9+CBM2	1.6E-133	endo-1,6-beta-D-glucanase BGN16.3	*Crassostrea gigas*	eukaryotes
c177559_g1_i4	0.21	0.04	GH1	2E-41	lactase-phlorizin hydrolase-like	*Crassostrea virginica*	eukaryotes
c169869_g1_i2	0.21	0.04	GH11	3.6E-36	1,4-beta-xylanase	*Micromonospora rifamycinica*	prokaryotes
c166743_g1_i1	0.19	0.03	GH30	9.8E-41	endo-1,6-beta-D-glucanase BGN16.3	*Crassostrea gigas*	eukaryotes
c178224_g1_i3	0.18	0.03	GH1+GH1	2.6E-78	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c177408_g1_i3	0.17	0.03	GH31	2.5E-46	neutral alpha-glucosidase AB-like isoform X2	*Crassostrea virginica*	eukaryotes
c176563_g3_i3	0.17	0.03	GH9	2.2E-74	cellulase	*Corbicula japonica*	eukaryotes
c176983_g1_i3	0.15	0.03	GH1+GH1	2.3E-90	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c181959_g1_i5	0.15	0.03	GH20	1E-94	beta-hexosaminidase subunit beta-like isof. X3	*Crassostrea virginica*	eukaryotes
c179342_g2_i11	0.15	0.03	GH3	4.6E-51	beta-D-xylosidase 5 isoform X1	*Mizuhopecten yessoensis*	eukaryotes
c180449_g1_i10	0.14	0.03	GH9+CBM2	3.1E-114	cellulase	*Corbicula japonica*	eukaryotes
c177884_g1_i11	0.13	0.02	GH35	4E-56	beta-galactosidase-1-like protein 2	*Mizuhopecten yessoensis*	eukaryotes
c179397_g1_i1	0.12	0.02	GH31	1.2E-63	lysosomal alpha-glucosidase	*Crassostrea gigas*	eukaryotes
c180241_g3_i2	0.12	0.02	GH5_53	4.4E-186	GH family 5_53 domain-containing protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c177062_g4_i1	0.09	0.02	AA1_3	6E-12	laccase-2 isoform X1	*Crassostrea gigas*	eukaryotes
c177364_g2_i5	0.09	0.02	CE10	1.4E-21	acylamino-acid-releasing enzyme-like	*Octopus bimaculoides*	eukaryotes
c178525_g2_i1	0.09	0.02	GH1+GH1+GH1	1.9E-100	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c182208_g1_i1	0.03	0.01	GH1+GH1	1.9E-157	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c182171_g1_i4	0.03	0.01	GH20+GH20	1.6E-86	beta-hexosaminidase	*Crassostrea gigas*	eukaryotes

Table 4.7. CAZymes of the caecum. The table shows the results for the Mascot search of the LC-MS/MS analysis performed on *L. pedicellatus* caecum against the transcriptome of gills, digestive glands, caecum and crystalline style sac. Only proteins with two or more significant matches are included. The table fields are the same as in Table 4.6.

Contig name	emPAI score	emPAi score %	CAZy family	CAZy e-value	Annotation	Species	Domain
c88022_g1_i2	2.43	0.82	GH5	2.00E-29	endo-beta 1-4 mannanase	*Haliotis discus discus*	eukaryotes
c122374_g1_i1	2.24	0.76	GH134+3xCBM10	1.1E-36	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c175047_g4_i1	1.99	0.67	GH1	1.7E-25	beta-glucosidase	*Thermobacillus composti* KWC4	eukaryotes
c178040_g3_i1	1.61	0.54	GH1+GH1+GH1	6.1E-160	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c178040_g3_i2	1.59	0.54	GH1+GH1+GH1	3.2E-160	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c173097_g2_i4	1.45	0.49	GH9	1.6E-133	endoglucanase E-4	*Crassostrea gigas*	eukaryotes
c23661_g1_i1	1.43	0.48	GH9	4.00E-18	cellulase	*Haliotis gigantea*	eukaryotes
c4960_g1_i1	1.41	0.48	GH13	0.074	endoglucanase 13-like	*Aplysia californica*	eukaryotes
c178089_g2_i1	1.27	0.43	GH2	1.00E-114	beta-mannosidase	*Crassostrea gigas*	eukaryotes
c178040_g2_i1	1.17	0.40	GH1	1.1E-160	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c151620_g1_i1	1.14	0.39	GH9	7.5E-34	endoglucanase	*Mizuhopecten yessoensis*	eukaryotes
c177884_g1_i15	1.04	0.35	GH35	7E-56	beta-galactosidase-1-like protein 2, partial	*Crassostrea gigas*	eukaryotes
c177535_g3_i3	0.99	0.34	GH5_10	5.1E-39	beta-mannanase from blue mussel Mytilus edulis	*Mytilus Eedulis*	eukaryotes

c169024_g2_i1	0.94	0.32	GH5_10	2.4E-101	mannan endo-1,4-beta-mannosidase	*Mytilus edulis*	eukaryotes
c178326_g2_i4	0.93	0.31	GH9	8.00E-26	endoglucanase E-4	*Crassostrea gigas*	eukaryotes
c179345_g2_i6	0.91	0.31	GH2	2.00E-94	beta-mannosidase-like	*Crassostrea gigas*	eukaryotes
c165524_g1_i1	0.89	0.30	GH9	2.2E-32	cytoplasmic aconitate hydratase-like isoform X1	*Crassostrea gigas*	eukaryotes
c178040_g1_i1	0.87	0.29	GH1+GH1	1E-138	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c178089_g2_i13	0.81	0.27	GH2	2.5E-79	beta-mannosidase-like	*Limulus polyphemus*	eukaryotes
c119776_g1_i1	0.80	0.27	GH67	7.00E-47	alpha-glucuronidase	*Teredinibacter turnerae*	prokaryotes
c178065_g1_i3	0.79	0.27	GH1	6.00E-24	beta-glucosidase	*Corbicula japonica*	eukaryotes
c181524_g3_i2	0.79	0.27	GH2	3.00E-152	beta-glucuronidase	*Chrysochloris asiatica*	eukaryotes
c180385_g4_i3	0.76	0.26	GH1	1.1E-65	lactase-phlorizin hydrolase	*Callorhinchus milii*	eukaryotes
c180131_g1_i2	0.72	0.24	GH1	9.7E-99	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c165315_g1_i2	0.70	0.24	GH5_8+CBM10+CBM2	3E-20	beta-mannosidase	*Alteromonadaceae bacterium* Bs12	prokaryotes
c176983_g1_i4	0.70	0.24	GH1+GH1	1.3E-111	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c166743_g1_i1	0.69	0.23	GH30	9.8E-41	endo-1,6-beta-D-glucanase BGN16.3-like	*Crassostrea gigas*	eukaryotes
c180449_g1_i10	0.69	0.23	GH9+CBM2	3.1E-114	cellulase	*Corbicula japonica*	eukaryotes
c179345_g2_i11	0.69	0.23	GH2	9.00E-107	beta-mannosidase-like	*Crassostrea gigas*	eukaryotes
c172990_g3_i3	0.69	0.23	GH9	8.00E-10	cellulase	*Haliotis kamtschatkana*	eukaryotes
c181813_g5_i3	0.65	0.22	GH1	2.1E-65	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c176344_g1_i3	0.64	0.22	GH9	1.3E-77	cellulase	*Corbicula japonica*	eukaryotes
c156925_g1_i1	0.63	0.21	GH9	2.4E-38	cellulase	*Corbicula japonica*	eukaryotes
c178525_g2_i4	0.62	0.21	GH1+GH1+GH1	1.9E-100	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c153913_g1_i1	0.61	0.21	GH30	7.7E-41	endo-1,6-beta-D-glucanase BGN16.3-like	*Crassostrea gigas*	eukaryotes
c168511_g1_i1	0.61	0.21	GH9	6.7E-109	cellulase	*Corbicula japonica*	eukaryotes
c177148_g1_i2	0.60	0.20	GH1	2.00E-46	lactase-like protein	*Crassostrea gigas*	eukaryotes
c172023_g1_i1	0.59	0.20	GH10	2E-15	endo-1,4-beta-xylanase	*Corbicula japonica*	eukaryotes
c176344_g1_i4	0.55	0.19	GH9	2E-123	cellulase	*Corbicula japonica*	eukaryotes
c178525_g2_i1	0.54	0.18	GH1+GH1+GH1	1.9E-100	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c170561_g7_i1	0.52	0.18	GH10	1.4E-28	GH family 10 domain-containing protein	*Alteromonadaceae bacterium* Bs02	prokaryotes
c177035_g4_i2	0.50	0.17	GH1	4.6E-20	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c171242_g3_i1	0.50	0.17	GH18	2.8E-83	putative chitinase 3	*Crassostrea gigas*	eukaryotes
c176344_g2_i3	0.48	0.16	GH9	1.9E-42	endoglucanase A-like isoform X2	*Biomphalaria glabrata*	eukaryotes
c167147_g1_i2	0.47	0.16	GH1	7.8E-50	lactase-phlorizin hydrolase-like	*Saccoglossus kowalevskii*	eukaryotes
c176563_g3_i3	0.47	0.16	GH9	2.2E-74	cellulase	*Corbicula japonica*	eukaryotes
c176117_g2_i7	0.45	0.15	CE6	1E-10	sialate O-acetylesterase-like isoform X1	*Octopus bimaculoides*	eukaryotes
c167807_g1_i1	0.44	0.15	GH1	3.2E-11	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c172204_g1_i1	0.43	0.15	GH31	3.2E-82	glyoxylate reductase/hydroxypyruvate reductase	*Mizuhopecten yessoensis*	eukaryotes
c169869_g1_i1	0.41	0.14	GH11+CBM10+CBM5	1.7E-14	1,4-beta-xylanase	*Cellvibrio sp.* OA-2007	prokaryotes
c177348_g1_i10	0.41	0.14	GH35	2.90E-129	glycogen phosphorylase, muscle form-like	*Mizuhopecten yessoensis*	eukaryotes
c178326_g1_i1	0.40	0.14	GH9	7.3E-60	endoglucanase 14-like	*Aplysia californica*	eukaryotes
c175019_g2_i1	0.38	0.13	GH13_24+GH13_15	2.4E-44	alpha-amylase	*Corbicula fluminea*	eukaryotes
c176756_g1_i11	0.36	0.12	GH16	3.5E-38	endo-1,3-beta-D-glucanase	*Tapes literata*	eukaryotes
c182208_g1_i1	0.36	0.12	GH1+GH1	1.9E-157	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes

c167653_g1_i1	0.36	0.12	GH2	1.00E-58	beta-mannosidase-like	*Crassostrea gigas*	eukaryotes
c177997_g2_i1	0.35	0.12	GH1+GH1	4.9E-74	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c86197_g1_i1	0.35	0.12	GH45	5.00E-43	endoglucanase-like	*Crassostrea gigas*	eukaryotes
c175142_g1_i6	0.34	0.12	GH109	3.6E-12	oxidoreductase YrbE-like isoform X2	*Lingula anatina*	eukaryotes
c178224_g1_i5	0.34	0.12	GH1+GH1	4.1E-89	lactase-phlorizin hydrolase-like	*Lingula anatina*	eukaryotes
c169869_g1_i2	0.33	0.11	GH11	3.6E-36	1,4-beta-xylanase	Micromonospora rifamycinica]	prokaryotes
c177559_g1_i4	0.33	0.11	GH1	2E-41	lactase-phlorizin hydrolase-like	*Crassostrea virginica*	eukaryotes
c180241_g3_i2	0.33	0.11	GH5_53	4.4E-186	GH family 5_53 domain-containing protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c177931_g2_i1	0.31	0.10	GH18	3.7E-35	acidic mammalian chitinase	*Crassostrea gigas*	eukaryotes
c179397_g1_i1	0.31	0.10	GH31	1.2E-63	lysosomal alpha-glucosidase	*Crassostrea gigas*	eukaryotes
c180385_g2_i2	0.30	0.10	GH1	6.5E-77	lactase-phlorizin hydrolase-like isoform X2	*Lingula anatina*	eukaryotes
c174752_g6_i1	0.28	0.09	AA4	2.6E-23	putative D-lactate dehydrogenase, mitochondrial	*Crassostrea gigas*	eukaryotes
c179423_g1_i1	0.27	0.09	CE3+CE3+CBM10	1.3E-38	carbohydrate esterase, family 3	*Teredinibacter turnerae* T7901	prokaryotes
c132896_g1_i1	0.27	0.09	GH9	3.00E-54	glycoside hydrolase family 9	*Teredinibacter turnerae*	prokaryotes
c181452_g1_i2	0.25	0.08	CE1	1.5E-52	S-formylglutathione hydrolase-like	*Limulus polyphemus*	eukaryotes
c176461_g2_i2	0.24	0.08	GH109	4.5E-11	trans-1,2-dihydrobenzene-1,2-diol dehydrogenase	*Lingula anatina*	eukaryotes
c177919_g1_i10	0.24	0.08	GH1	1.8E-137	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	eukaryotes
c180385_g3_i4	0.22	0.07	GH1	2.9E-39	lactase-phlorizin hydrolase-like	*Mizuhopecten yessoensis*	eukaryotes
c170561_g5_i1	0.22	0.07	GH10+CBM60	2E-53	CE family 6&GH family 10 domain-cont. protein	*Alteromonadaceae bacterium* Bs08	prokaryotes
c178993_g2_i3	0.20	0.07	CE10	3.4E-22	prolyl tripeptidyl peptidase-like	*Lingula anatina*	eukaryotes
c176118_g2_i4	0.20	0.07	GH31	1.1E-80	sucrase-isomaltase, intestinal-like	*Lingula anatina*	eukaryotes
c182171_g1_i4	0.20	0.07	GH20+GH20	1.6E-86	putative beta-hexosaminidase	*Crassostrea gigas*	eukaryotes
c171811_g1_i2	0.18	0.06	GH13_24	3E-144	alpha-amylase	*Hyriopsis cumingii*	eukaryotes
c177241_g1_i3	0.17	0.06	GH45+GH45+CBM10	7.4E-29	GH family 45 domain-containing protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c158412_g1_i2	0.16	0.05	CE12	1.3E-10	isoamyl acetate-hydrolyzing esterase 1 homolog	*Crassostrea gigas*	eukaryotes
c173596_g2_i1	0.16	0.05	AA7	6E-13	alkyldihydroxyacetonephosphate synthase	*Crassostrea gigas*	eukaryotes
c173837_g2_i2	0.16	0.05	AA10+CBM10	6.1E-37	chitin-binding protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c177712_g2_i7	0.16	0.05	GH1+GH1	8.9E-60	beta-glucosidase	*Corbicula japonica*	eukaryotes
c177442_g1_i2	0.14	0.05	GH9+CBM3	7.6E-71	glycoside hydrolase family 9	*Alteromonadales bacterium* BS08	prokaryotes
c177250_g1_i1	0.12	0.04	GH13_15+GH13_15	1.5E-79	alpha-amylase, partial	*Cerastoderma edule*	eukaryotes
c174824_g1_i1	0.09	0.03	CE15+CE15+CBM57	7.4E-35	carbohydrate esterase family 15 domain protein	*Teredinibacter turnerae* T7901	prokaryotes
c181959_g1_i5	0.09	0.03	GH20	1E-94	beta-hexosaminidase subunit alpha-like isoform X3	*Lingula anatina*	eukaryotes
c181813_g3_i4	0.09	0.03	GH1	8.9E-130	lactase-phlorizin hydrolase-like isoform X2	*Lingula anatina*	eukaryotes
c181032_g1_i1	0.08	0.03	GH35	3.2E-110	beta-galactosidase-1-like protein 2	*Mizuhopecten yessoensis*	eukaryotes
c176477_g1_i1	0.07	0.02	GH38+GH38	2.4E-46	lysosomal alpha-mannosidase	*Crassostrea gigas*	eukaryotes

Table 4.8. CAZymes of the gills. The table shows the results for the Mascot search of the LC-MS/MS analysis performed on *L. pedicellatus* gills against the transcriptome of gills, digestive glands, caecum and crystalline style sac. Proteins with two or more significant matches are in black, while those with just one match are in green, and therefore the relative abundance among those with more than two matched could not be calculated. The table fields are the same as in Table 4.6.

Contig name	emPAI	emPAi	CAZy family	CAZy e-value	Annotation	Species	Domain
c165315_g1_i	0.7	0.33	GH5_8+CBM10+CBM	3E-20	beta-mannosidase	*Alteromonadaceae bacterium* Bs12	prokaryotes
c174752_g6_i	0.63	0.30	AA4	2.6E-23	D-lactate dehydrogenase, mitochondrial	*Crassostrea gigas*	eukaryotes
c175967_g4_i	0.32	0.15	GH59	4.2E-112	galactocerebrosidase-like isoform X4	*Lingula anatina*	eukaryotes
c122374_g1_i	0.26	N/A	GH134+3xCBM10	6.70E-31	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium* Bs12	prokaryotes
c177483_g5_i	0.21	0.10	GH29	9.4E-59	alpha-L-fucosidase	*Strongylocentrotus purpuratus*	eukaryotes
c181452_g1_i	0.18	N/A	CE1	8.50E-25	S-formylglutathione hydrolase-like	*Acanthaster planci*	eukaryotes
c165524_g1_i	0.17	0.08	GH9	2.2E-32	cytoplasmic aconitate hydratase-like isoform X1	*Crassostrea gigas*	eukaryotes
c176477_g1_i	0.16	0.08	GH38+GH38	2.4E-46	lysosomal alpha-mannosidase	*Crassostrea gigas*	eukaryotes
c179342_g2_i	0.14	0.07	GH3	3.1E-51	beta-xylosidase/alpha-L-arabinofuranosidase 1-like isoform	*Mizuhopecten yessoensis*	eukaryotes
c178993_g2_i	0.13	0.06	CE10	3.4E-22	prolyl tripeptidyl peptidase-like	*Lingula anatina*	eukaryotes
c177001_g3_i	0.11	N/A	GH5+CBM10	2.40E-13	beta-mannosidase	*Teredinibacter sp.* 1162T.S.0a.05	prokaryotes
c181959_g1_i	0.09	0.04	GH20	1E-94	beta-hexosaminidase subunit alpha-like	*Lingula anatina*	eukaryotes
c165327_g1_i	0.09	N/A	GH15	1.20E-10	phosphorylase b kinase regulatory subunit alpha	*Mizuhopecten yessoensis*	eukaryotes
c168055_g1_i	0.08	N/A	GH103	5.10E-55	lytic murein transglycosylase B	*Teredinibacter sp.* 1162T.S.0a.05	prokaryotes
c173596_g2_i	0.08	N/A	AA7	5.00E-13	alkyldihydroxyacetonephosphate synthase	*Crassostrea gigas*	eukaryotes
c176294_g3_i	0.08	N/A	GH18	1.80E-36	di-N-acetylchitobiase-like	*Crassostrea virginica*	eukaryotes
c173140_g2_i	0.07	N/A	GH10	5.40E-51	1,4-beta-xylanase	*Luteimonas sp.* 100111	prokaryotes
c175079_g3_i	0.06	0.03	GH31	7.2E-149	lysosomal alpha-glucosidase-like	*Sturnus vulgaris*	eukaryotes
c171616_g1_i	0.06	N/A	GH5_8	2.60E-39	beta-mannosidase	*Saccharophagus degradans*	prokaryotes
c179423_g1_i	0.06	N/A	CE3+CE3+CBM10	6.10E-16	carbohydrate esterase	*Teredinibacter turnerae*	prokaryotes
c174824_g1_i	0.05	N/A	CE15+CE15+CBM57	1.70E-28	carbohydrate esterase	*Alteromonadales bacterium* BS08	prokaryotes
c177408_g1_i	0.05	N/A	GH31	2.50E-46	neutral alpha-glucosidase AB-like isoform X2	*Crassostrea virginica*	eukaryotes
c179310_g1_i	0.05	N/A	CE6	0.00016	sialate O-acetylesterase-like isoform X3	*Mizuhopecten yessoensis*	eukaryotes
c178689_g1_i	0.04	N/A	GH18	5.00E-11	di-N-acetylchitobiase isoform X1	*Crassostrea gigas*	eukaryotes
c179335_g1_i	0.04	N/A	GH13_17	4.00E-76	probable maltase	*Lingula anatina*	eukaryotes
c181406_g2_i	0.04	N/A	CE1	3.60E-06	protein FAM172A-like isoform X1	*Crassostrea virginica*	eukaryotes
c182171_g1_i	0.02	N/A	GH20+GH20	1.40E-50	beta-hexosaminidase	*Crassostrea gigas*	eukaryotes

Table 4.9. CAZymes of the crystalline style. The table shows the results for the Mascot search of the LC-MS/MS analysis performed on *L. pedicellatus* crystalline style against the transcriptome of gills, digestive glands, caecum and crystalline style sac. Only proteins with two or more significant matches are included. The table fields are the same as in Table 4.6.

Contig name	emPAI	emPAi	CAZy family	CAZy e-value	Annotation	Species	Domain
c165315_g1_i2	0.60	2.51	GH5_8+CBM10+CBM2	7.3E-16	beta-mannosidase GH family 5_8 domain-containing protein	*Alteromonadaceae bacterium* Bs12	Prokaryotes
c151620_g1_i1	0.56	2.35	GH9	1.1E-85	cellulase	*Corbicula japonica*	Eukaryotes
c122374_g1_i1	0.52	2.18	GH134+3xCBM10	7E-37	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium* Bs12	Prokaryotes
c176917_g2_i1	0.41	1.72	GH5_53+CBM10+CBM2	3.40E-55	GH family 5_53 domain-containing protein	*Alteromonadaceae bacterium* Bs08	Prokaryotes
c169024_g2_i1	0.30	1.26	GH5_10	1E-31	mannan endo-1,4-beta-mannosidase	*Mytilus edulis*	Eukaryotes
c165524_g1_i1	0.19	0.80	GH9	1.3E-32	cytoplasmic aconitate hydratase	*Crassostrea gigas*	Eukaryotes
c176756_g1_i11	0.18	0.75	GH16	2.1E-38	endo-1,3-beta-D-glucanase	*Tapes literata*	Eukaryotes
c179423_g1_i1	0.17	0.71	CE3+CE3+CBM10	8.1E-39	carbohydrate esterase	*Teredinibacter turnerae*	Prokaryotes
c180176_g1_i1	0.12	0.50	GH134+3XCBM10	5.8E-59	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium* Bs12	Prokaryotes
c177884_g1_i11	0.12	0.50	GH35	2.5E-56	beta-galactosidase-1-like protein 2, partial	*Crassostrea gigas*	Eukaryotes
c173097_g2_i4	0.09	0.38	GH9+CBM2	9.9E-134	endoglucanase E-4	*Crassostrea gigas*	Eukaryotes
c177535_g3_i3	0.06	0.25	GH5_10	8.1E-29	beta-mannanase	*Mytilus edulis*	Eukaryotes
c178040_g1_i2	0.04	0.17	GH1+GH1+GH1+GH1+GH1	1.8E-159	lactase-phlorizin hydrolase-like	*Biomphalaria glabrata*	Eukaryotes

The digestive glands include a total of 52 proteins containing CAZy domains. Surprisingly, the most abundant of them is a GH5_8 of bacterial origin, which is transcribed in the gills. Together with a GH5_53, a GH10 and a GH11, also bacterial and transcribed in the gills (some of them however are not in the most highly expressed 1,000 transcripts), they represent the only prokaryotic proteins of the gills. The endogenous proteins contain quite a wide array of CAZy families (21 different ones), with GH1 and GH9 being the most represented (11 and 8 proteins respectively), and indeed these families are the most abundant CAZymes in the proteome. CE10 and GH18 domains are found in three proteins, while all the other families are represented by one or two proteins. There are three different proteins belonging to the family of Auxiliary Activities. One is an AA4 and is annotated as a glyoxylate reductase/hydroxypyruvate reductase, while the CAZy database records activity of vanillyl-alcohol oxidase for this family. The other is an AA7, annotated as a perizosomal alkyldihydroxyacetonephosphate synthase and classed as having glucooligosaccharide oxidase or chitooligosaccharide oxidase activity in the CAZy database. Interestingly, the last one is an AA1_3, a putative laccase or ferroxidase or p-diphenol:oxygen oxidoreductase, as it contains at least two cupredoxin domains, which are copper-containing domain involved in inter-molecular electron transfer reactions (Adman, 1991). It is transcribed in the gland but in very low amount, and is found in the caecum proteomics but among the proteins that only have one significant match.

Ninety different proteins containing a CAZy domain were identified in the caecum proteome, of which 14 are prokaryotic. As in the digestive glands, the CAZy families are plentiful (27 different ones), but the most represented families are GH1 (25 different eukaryotic proteins) and GH9 (16 proteins of which only two are prokaryotic), many of which are transcribed in the digestive glands. AA10s, GH11s, GH134s, CE15s and CE3s are exclusively of bacterial origin. The GH5s include three different subfamilies (10, 53 and 8) and they are of both prokaryotic and eukaryotic origin, with the subfamilies 53 and 8 though being only produced by the bacteria in the gills. Three proteins have a GH10 module, and they are of mixed origin and present in similar amounts. The two proteins with an AA4 or AA7 module are the same that we described in the glands' proteome, suggesting that they come from that organ, despite not being identified in its most highly expressed 1,000 transcripts.

The gill proteome only comprises 10 proteins containing a CAZy module, of which just one is bacterial in origin (a GH5_8). This result is surprising, since the gills are the organ of

residence of the endosymbiotic bacteria, which are thought to produce the bacterial lignocellulolytic enzymes that are then found in the caecum. For this reason, we decided to include in the table presented here also those proteins that have only one significant match, hence adding ten more eukaryotic and seven prokaryotic CAZymes to the list (two GH5s, of which one of subfamily 8, one multimodular GH134 annotated as a mannanase, a putative xylanase GH10, a GH103, a CE3 and a CE15). Most of the latter are found both in the gill transcriptome and caecum proteome, giving a further indication that the bacterial lignocellulolytic enzymes are produced by gill bacteria and then delivered to the caecum to contribute to the digestion of wood; they probably were not found in our initial analysis just because of the difficulty of breaking the bacterial cell walls to release the enzymes that can then be detected by shotgun proteomics.

Thirteen CAZymes belonging to seven different families were identified in the crystalline style proteome, whose origin is both bacterial and endogenous. The most abundant CAZy families are GH5 and GH9, with four and three proteins respectively. Other families include CE3, GH134 (2 proteins), GH1, GH16 and GH35. All the bacterial proteins found in the crystalline style contain CBM modules of the family 10 and in some instances of family 2, and they all are transcribed in the gills, while the shipworm proteins rarely include CBMs and are produced in the digestive glands.

4.5.1 The non-CAZy proteins of the crystalline style

The CAZymes identified in the crystalline style of *L. pedicellatus* have been presented in the previous paragraph. Since the protein composition of the crystalline style of a shipworm has never been studied to date, we analysed the whole proteome of the style beyond the CAzymes, and the results are shown in Figure 4.7. The major proteins found in the style of eukaryotic origin, with the exceptions of the prokaryotic ones that are found among the CAZymes. The most abundant proteins identified are structural ones, mainly mucins (22.5%), tubulin (15.0%) and collagen (11.4%). Another relatively abundant (6.8%) group identified is that containing proteins annotated as “membrane attack complex/perforin domain-containing proteins”. These are proteins named after a domain that is found in members of the membrane attack complex (MAC) and in perforin (PF) (Rosado *et al.*, 2007). These proteins typically have an antimicrobial function, forming part of the defence system in many organisms, including animals, plants, protozoa, fungi and bacteria, known to oligomerise on the cell membrane and create a pore that causes cell lysis and therefore

death (Peitsch and Tschopp, 1991; Rosado *et al.*, 2008). The styles also contains other proteins annotated as "complement component C3" (2.6%) and involved in the the innate and adaptive immune response against invading pathogens (Sahu and Lambris, 2001), as well as "hemocytin-like proteins" (4.1%), which are thought to be homologues of the mammalis von Willebrand factor and therefore also invovled in immunity (Yamakawa and Tanaka, 1999). A fructose-1, 6-bisphosphate aldolase constitutes 2.1% of the style, while the rest (21.4%) is made up of proteins that are found in small percentages (less than 2%) and that cannot be grouped into bigger categories; these are listed in table 4.9.

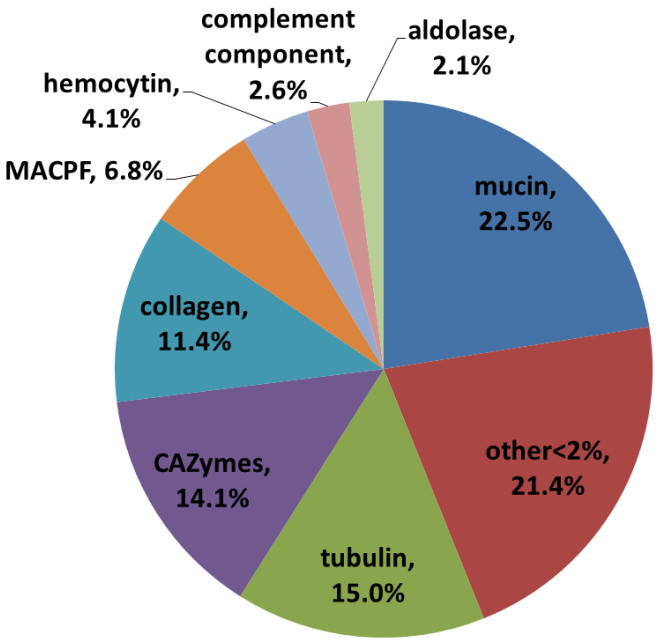


Figure 4.7. The crystalline style protein composition of *L. pedicellatus*. Pie chart showing the crystalline style relative protein abundance (calculated from the emPAI score). MACPF=membrane attack complex/perforin (MACPF) domain-containing proteins; aldolase=fructose-1, 6-bisphosphate aldolase, hemocytin=hemocytin-like protein.

Table 4.10. Crystalline style proteomics results. The table shows the results for the Mascot search of the LC-MS/MS analysis performed on *L. pedicellatus* crystalline style against the transcriptome of gills, digestive glands, caecum and crystalline style sac, excluding the CAZymes. Only proteins with two or more significant matches are included. The table fields are the same as in Table 4.6.

Contig name	emPAI score	emPAi score %	Annotation	Species	Domain
DN340064_c0_g4_i1	1.47	6.16	beta-tubulin, partial	*Anthopleura elegantissima*	Eukaryotes
DN355708_c3_g2_i4	1.27	5.32	collagen alpha-4(VI) chain	*Crassostrea gigas*	Eukaryotes
DN350151_c5_g1_i1	1.2	5.03	MACPF domain-containing protein 4	*Crassostrea gigas*	Eukaryotes
c169268_g2_i5	0.89	3.73	mucin-5AC-like	*Crassostrea gigas*	Eukaryotes
DN344200_c2_g1_i4	0.88	3.69	mucin-like	*Mizuhopecten yessoensis*	Eukaryotes
c174353_g2_i2	0.86	3.60	tubulin A, partial	*Hyriopsis cumingii*	Eukaryotes
DN352516_c1_g8_i6	0.85	3.56	mucin-5AC-like	*Crassostrea gigas*	Eukaryotes
DN356805_c10_g2_i3	0.82	3.43	mucin-2-like	*Crassostrea gigas*	Eukaryotes
DN356805_c10_g2_i18	0.78	3.27	mucin-2-like	*Crassostrea gigas*	Eukaryotes
DN353760_c3_g2_i4	0.74	3.10	hemocytin-like	*Biomphalaria glabrata*	Eukaryotes
DN344200_c4_g92_i1	0.7	2.93	mucin-5AC-like	*Crassostrea gigas*	Eukaryotes
c180086_g1_i5	0.68	2.85	tubulin beta chain isoform X2	*Strongylocentrotus purpuratus*	Eukaryotes
c180696_g1_i3	0.61	2.55	complement component C3	*Ruditapes decussatus*	Eukaryotes
DN357388_c1_g3_i10	0.58	2.43	tubulin alpha-1 chain	*Melipona quadrifasciata*	Eukaryotes
DN358989_c6_g6_i4	0.57	2.39	collagen alpha-1(XII) chain-like	*Octopus bimaculoides*	Eukaryotes
DN359103_c3_g1_i8	0.55	2.30	collagen alpha-4(VI) chain	*Crassostrea gigas*	Eukaryotes
c165516_g1_i1	0.49	2.05	fructose-1, 6-bisphosphate aldolase	*Meretrix meretrix*	Eukaryotes
c157004_g1_i1	0.47	1.97	uncharacterised	*Crassostrea gigas*	Eukaryotes
DN350151_c5_g1_i5	0.43	1.80	apextrin-MACPF domain-containing protein 4	*Saccoglossus kowalevskii*	Eukaryotes
DN343231_c6_g1_i5	0.39	1.63	mucin-2-like	*Lingula anatina*	Eukaryotes
c159122_g1_i1	0.34	1.42	collagen alpha-4(VI) chain	*Crassostrea gigas*	Eukaryotes
c165399_g1_i3	0.29	1.21	atrial natriuretic peptide-converting enzyme-like	*Tetranychus urticae*	Eukaryotes
c178325_g2_i1	0.29	1.21	arginine kinase	*Pholas orientalis*	Eukaryotes
c44602_g1_i1	0.28	1.17	putative perlucin 5	*Haliotis discus discus*	Eukaryotes
c178787_g7_i1	0.26	1.09	dermatopontin-like	*Aplysia californica*	Eukaryotes
DN344200_c4_g90_i2	0.24	1.01	hemocytin-like	*Aplysia californica*	Eukaryotes
DN355231_c4_g3_i4	0.23	0.96	glyceraldehyde-3-phosphate dehydrogenase	*Lottia gigantea*	Eukaryotes
DN355708_c4_g4_i1	0.23	0.96	matrilin-2-like	*Lingula anatina*	Eukaryotes
DN295848_c0_g1_i2	0.22	0.92	oncoprotein-induced transcript 3 protein	*Crassostrea gigas*	Eukaryotes
DN349989_c4_g2_i15	0.21	0.88	sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1-like	*Aplysia californica*	Eukaryotes
DN320266_c1_g5_i2	0.19	0.80	triosephosphate isomerase-like	*Crassostrea gigas*	Eukaryotes
DN357771_c5_g7_i5	0.19	0.80	actin, non-muscle 6.2	*Hydra vulgaris*	Eukaryotes
c167054_g1_i1	0.18	0.75	Niemann-Pick C1 protein-like isoform X2	*Biomphalaria glabrata*	Eukaryotes
c177078_g6_i1	0.17	0.71	transaldolase-like isoform X3	*Crassostrea gigas*	Eukaryotes

c166928_g1_i1	0.15	0.63	beta-Ig-H3/fasciclin	*Lottia gigantea*	Eukaryotes
c180197_g2_i1	0.15	0.63	malate dehydrogenase, cytoplasmic-like	*Crassostrea gigas*	Eukaryotes
c179411_g3_i3	0.14	0.59	aspartate aminotransferase, cytoplasmic	*Lottia gigantea*	Eukaryotes
DN354309_c3_g3_i2	0.14	0.59	tomoregulin-2-like	*Crassostrea gigas*	Eukaryotes
DN342855_c2_g2_i2	0.13	0.54	periostin-like	*Crassostrea gigas*	Eukaryotes
DN335277_c2_g9_i1	0.11	0.46	nerve hemoglobin	*Spisula solidissima*	Eukaryotes
c166397_g1_i1	0.1	0.42	leucine-rich repeats and immunoglobulin-like domains protein 3	*Crassostrea gigas*	Eukaryotes
c173766_g10_i1	0.1	0.42	leucine-rich repeat and death domain-containing protein 1-like	*Crassostrea gigas*	Eukaryotes
DN352090_c3_g1_i14	0.1	0.42	phosphoenolpyruvate carboxykinase [GTP]-like	*Crassostrea gigas*	Eukaryotes
DN342595_c1_g6_i3	0.09	0.38	mitochondrial H+ ATPase a subunit	*Pinctada fucata*	Eukaryotes
DN348602_c1_g3_i1	0.09	0.38	arginine kinase	*Crassostrea gigas*	Eukaryotes
DN348842_c6_g6_i4	0.09	0.38	signal peptide, CUB and EGF-like domain-containing protein 1 isoform X2	*Crassostrea gigas*	Eukaryotes
c179370_g1_i11	0.07	0.29	von Willebrand factor type egf and pentraxin domain-containing protein 1	*Crassostrea gigas*	Eukaryotes
DN349615_c7_g4_i3	0.07	0.29	heat shock protein 70	*Meretrix meretrix*	Eukaryotes
DN339117_c0_g1_i5	0.06	0.25	vascular endothelial growth factor A	*Meretrix meretrix*	Eukaryotes
DN353238_c3_g2_i1	0.06	0.25	tektin-4-like	*Crassostrea gigas*	Eukaryotes
DN354219_c2_g4_i8	0.06	0.25	mucin-19-like	*Aplysia californica*	Eukaryotes
DN354229_c2_g5_i2	0.06	0.25	deleted in malignant brain tumors 1 protein-like	*Branchiostoma floridae*	Eukaryotes
DN357645_c6_g5_i1	0.06	0.25	ovochymase-1	*Crassostrea gigas*	Eukaryotes
c178087_g1_i2	0.05	0.21	elongation factor 2	*Crassostrea gigas*	Eukaryotes
DN352474_c0_g1_i7	0.05	0.21	regucalcin-like	*Crassostrea gigas*	Eukaryotes
c182101_g1_i2	0.04	0.17	fibrocystin-L-like	*Crassostrea gigas*	Eukaryotes

4.6 The role of the digestive glands and caecum in immunity

Xylophagus animals, both vertebrates and invertebrates, normally rely on the presence of symbiotic microorganisms (bacteria, archaea, fungi and protozoa) that reside in their digestive system for the degradation of wood (Xie *et al.*, 2014). As described in section 1.3.3, shipworms are different because their symbiotic CAZymes-producing bacteria are instead resident in the gills, while the caecum has been found to be almost devoid of bacteria (Betcher *et al.*, 2012). Well-developed and morphologically varied bacterial population have instead been described in the intestine, which are different from those found in the gills (Betcher *et al.*, 2012), and are not thought to be important in wood degradation. The mechanism by which the caecum is kept almost bacteria-free is at present unknown. Bivalve immune systems are relatively well-known because of their importance in the human food industry and the fact that by being filter feeders they sieve and accumulate in their tissues pathogenic microorganisms that are found in sea water (Zannella *et al.*, 2017). Shipworms normally alternate suspension feeding with wood boring - when they are provided with suspended phytoplankton – therefore they have to face the same immunological challenge; this is further complicated by the presence of endosymbiotic bacteria in the gills, which have to be kept under control and distant from the caecum, while still being able to deliver the CAZymes they produce. The major immune response in marine bivalve is phagocytosis performed by haemocytes, but they also possess a wide range of humoral (extracellular) tools to fight microbes which include apoptosis, chemotaxis, lectins, galectins, complement-like molecules, β 1,3-glucan-binding proteins, lipopolysaccharides, antimicrobial peptides, fibrinogen-related proteins, lysozymes, cathepsins, peptoglycan-recognition proteins, thioester bearing proteins, bactericidal permeability increasing proteins, leucine-rich repeat domains, ferroxidases and phenoloxidases (Bachere *et al.*, 1995; Song *et al.*, 2010; Vasta, 2012; Allam and Raftos, 2015; Zannella *et al.*, 2017; Wang *et al.*, 2018).

In order to find out which tissues transcribe proteins involved in immunity, and in which organs these proteins are consequently found, we examined the most highly expressed 1,000 transcripts of the four organs of *L. pedicellatus*, as well as proteome of the organs, including only proteins with more than one significant match. The search resulted in 32 different transcripts, which are listed in Table 4.11, where their protein abundance in the various organs is also shown. The most abundant transcripts are those annotated as complement component C1 or C3, as well as lectins and the closely related galectins-like

proteins. Proteins forming pores on the cell membrane of pathogens (MACPF) have four transcripts, as well as hemicentin-like proteins, and three were found annotated as macrophage receptors. Beta-1,3-glucan-binding proteins, lysosymes, apoptosis inducing factors and hemocytin-like proteins are found only in one or two transcripts. The caecum seems to be the organ where the highest number of proteins transcribed by these genes is found, as well as the highest protein amount, followed by the digestive glands. The gills contain only few proteins possibly involved in immunity, mainly galectins, a lysozyme, an apoptosis inducing factor and a member of the complement component. The crystalline style contains only three proteins involved in immunity: a MACPF protein, a hemocytin-like protein and the complement component C3.

Table 4.11. Transcripts annotated as putative proteins involved in immunity identified in the transcriptome of *L. pedicellatus* digestive glands, caecum, gills and crystalline style sac. The table lists the number (#) and name of the transcript (contig name), its length (seq length), the annotation (annotation) with its e-value (annot e-value) and species (species) obtained by blasting against the NCBI non-redundant database, whether the transcript is of eukaryotic or prokaryotic origin (origin), as well as the protein abundance calculated as emPAI score (emPAI score) obtained from shotgun proteomics performed on the four organs.

#	Contig name	Contig length	Annotation	Annot e-value	Species	Domain	Gland emPAI	Caecum emPAI	Gills emPAI	Style emPAI
1	c179848_g1_i1	1802	apoptosis-inducing factor 3 isoform X2	4.92E-174	*Crassostrea gigas*	Eukaryotes	0.09	0.66	0.11	0.00
2	c176988_g3_i1	5148	beta-1,3-glucan-binding protein	6.14E-60	*Crassostrea gigas*	Eukaryotes	0.00	0.00	0.00	0.00
3	c178292_g2_i1	1732	beta-1,3-glucan-binding protein	1.86E-16	*Crassostrea virginica*	Eukaryotes	0.00	0.00	0.00	0.00
4	c151810_g1_i1	840	complement C1q tumor necrosis factor-related	9.00E-07	*Larimichthys crocea*	Eukaryotes	0.1	0.38	0.00	0.00
5	c141058_g1_i1	703	complement C1q tumor necrosis factor-related	6.25E-54	*Crassostrea gigas*	Eukaryotes	0.00	0.29	0.00	0.00
6	c165349_g1_i4	678	complement C1q tumor necrosis factor-related	3.82E-24	*Mizuhopecten yessoensis*	Eukaryotes	0.00	0.00	0.00	0.00
7	c158018_g1_i1	902	complement C1Q-like protein	8.07E-17	*Hyriopsis cumingii*	Eukaryotes	0.00	0.11	0.00	0.00
8	c161031_g1_i1	845	complement C1q-like protein 3	3.48E-12	*Crassostrea virginica*	Eukaryotes	0.00	0.12	0.00	0.00
9	c165402_g1_i1	537	complement C1q-like protein 4	1.58E-06	*Crassostrea virginica*	Eukaryotes	0.00	0.66	0.00	0.00
10	c164050_g1_i1	716	complement C1q-like protein 4 isoform X1	7.06E-67	*Crassostrea gigas*	Eukaryotes	0.12	0.14	0.00	0.00
11	c156799_g1_i1	763	complement C1q-like protein 4	4.00E-07	*Mizuhopecten yessoensis*	Eukaryotes	0.00	0.00	0.00	0.00
12	c164314_g1_i1	687	complement C1q-like protein 4, partial	8.00E-06	*Biomphalaria glabrata*	Eukaryotes	0.00	0.14	0.00	0.00
13	c180696_g1_i3	5337	complement component C3	4.00E-144	*Ruditapes decussatus*	Eukaryotes	0.03	0.00	0.11	0.67
14	c179022_g4_i4	668	C-type lectin	1.49E-19	*Littorina littorea*	Eukaryotes	0.00	0.00	0.00	0.00
15	c181228_g2_i1	488	C-type lectin	3.56E-30	*Meretrix meretrix*	Eukaryotes	0.00	0.68	0.00	0.00
16	c164426_g1_i2	634	C-type lectin	1.36E-30	*Meretrix meretrix*	Eukaryotes	0.00	0.00	0.00	0.00
17	c156875_g1_i1	841	C-type lectin 2	3.6	*Anguilla japonica*	Eukaryotes	0.00	0.00	0.00	0.00
18	c164057_g1_i1	1350	galectin	1.00E-19	*Hyriopsis cumingii*	Eukaryotes	0.43	0.39	0.07	0.00
19	c178415_g3_i2	1386	galectin	2.00E-83	*Hyriopsis cumingii*	Eukaryotes	0.19	0.14	0.07	0.00
20	c162467_g1_i2	1284	lysozyme	4.02E-60	*Ruditapes philippinarum*	Eukaryotes	0.28	0.00	0.23	0.00
21	DN350151_c5_g1_i6	3190	MACPF domain-containing protein 4	4.05E-140	*Mytilus galloprovincialis*	Eukaryotes	0.34	1.18	0.00	0.00
22	DN350151_c5_g1_i3	2581	MACPF domain-containing protein 4	6.42E-112	*Mytilus galloprovincialis*	Eukaryotes	0.33	1.42	0.00	0.00
23	DN350151_c5_g1_i1	2839	MACPF domain-containing protein 4	2.64E-141	*Mytilus galloprovincialis*	Eukaryotes	0.00	0.00	0.00	1.2
24	DN350151_c5_g1_i4	3078	MACPF domain-containing protein 4	2.06E-110	*Mytilus galloprovincialis*	Eukaryotes	0.00	0.00	0.00	0.00
25	DN344200_c4_g90_i2	1144	hemocytin-like	1.58E-60	*Aplysia californica*	Eukaryotes	0.00	0.00	0.00	0.24
26	c174145_g1_i10	581	hemicentin-1-like	0.016	*Crassostrea virginica*	Eukaryotes	0.00	0.36	0.00	0.00
27	c177653_g3_i1	691	hemicentin-1-like	1.29E-22	*Crassostrea virginica*	Eukaryotes	0.00	0.15	0.00	0.00
28	c178872_g1_i4	710	hemicentin-2-like	7.10E-27	*Mizuhopecten yessoensis*	Eukaryotes	0.00	0.00	0.00	0.00
29	c174180_g1_i1	449	hemicentin-2-like	4.00E-11	*Mizuhopecten yessoensis*	Eukaryotes	0.00	0.22	0.00	0.00
30	c176985_g2_i1	552	macrophage mannose receptor 1	6.86E-29	*Saccoglossus kowalevskii*	Eukaryotes	0.15	0.00	0.00	0.00
31	c181960_g2_i4	4875	macrophage mannose receptor 1 isoform X4	0	*Crassostrea gigas*	Eukaryotes	0.03	0.00	0.00	0.00
32	c176985_g3_i1	2051	macrophage mannose receptor 1-like	1.71E-65	*Pundamilia nyererei*	Eukaryotes	0.00	0.00	0.00	0.00

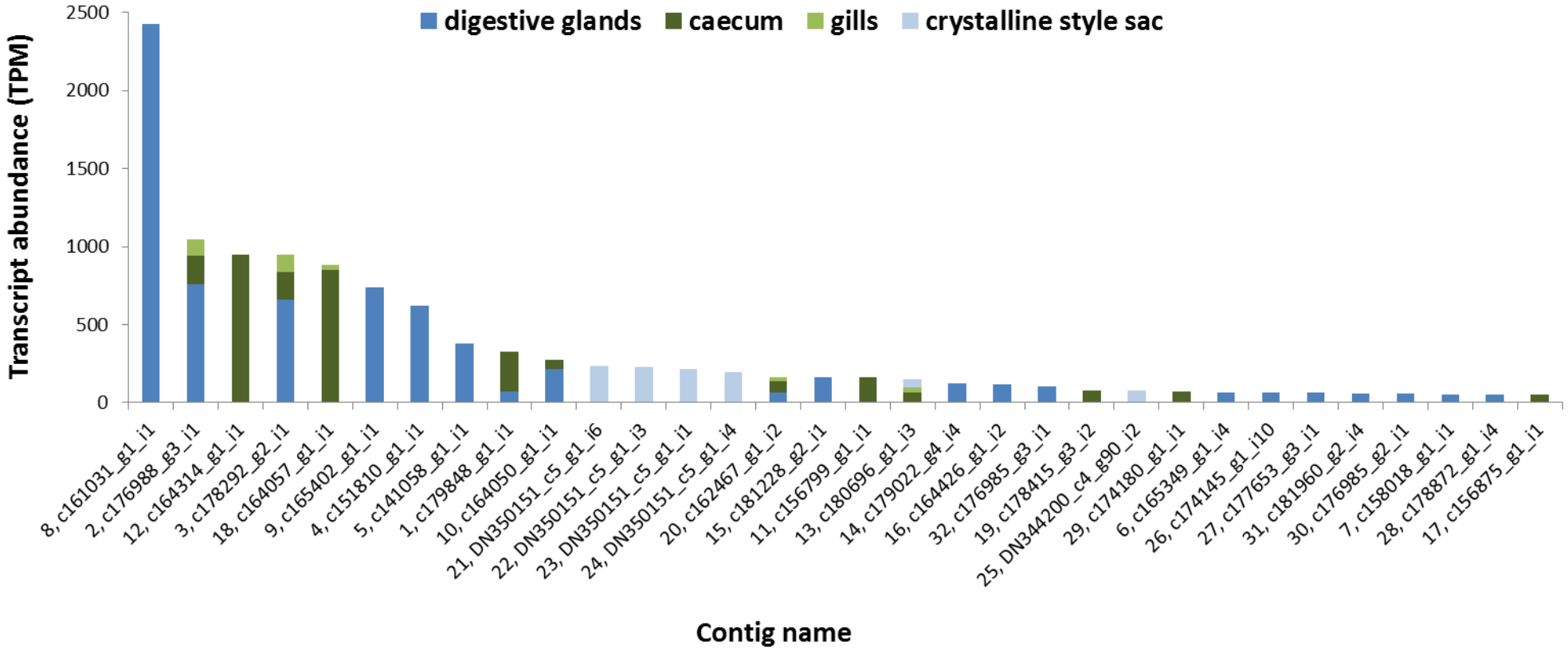


Figure 4.8. Expression levels of the putative protein involved in immunity identified in the transcriptome of *L. pedicellatus*. The graph shows the transcript abundance of the thirty-two contigs annotated as putative immunity proteins (listed in table 4.6) found in the best 1,000 transcripts of the digestive gland, caecum, gills and crystalline style. The number before the contig name on the X axis corresponds to the number on the first column of Table 4.6.

The expression levels of the transcripts listed in Table 4.6 are shown in Figure 4.8, which shows that the digestive glands are the main organs involved in the transcription of proteins related to immunity, with some important contributions from the caecum and minor ones from the crystalline style. The most expressed transcripts are those of the complement component, which are mostly expressed by the digestive glands, but also by the caecum. Beta-1,3-glucan-binding proteins also appear among the top four expressed contigs, and are transcribed by digestive glands, caecum and gills, though the former produce more transcripts than the other organs. Among the most expressed ten genes, the apoptosis inducing factor is produced mainly by the caecum, as well as galectin. The four contigs expressing for the MACPF proteins are transcribed exclusively by the crystalline style sac, which also expresses for the protein annotated as hemocytin-like and for a member of the complement component.

4.7 Discussion

A detailed analysis of the transcriptome and proteome of four shipworm tissues have been performed: the gills where the endosymbiotic bacteria reside, the digestive glands, where most of the endogenous CAZymes are thought to be produced, the caecum, which is the main site of wood digestion, and the crystalline style, which facilitates wood digestion. Our studies show that the caecum contains a cocktail of eukaryotic and prokaryotic CAZymes, but that few of these are transcribed in the caecum tissues. Indeed, the majority of the CAZymes in the caecum appear to be transcribed in the gills by the bacteria, or in the digestive gland by the shipworm. The crystalline style contains a large amount of CAZymes, but most of these originate in the gills and digestive gland, similarly to the caecum. Our results indicate a complex digestive system in which digestive proteins are transported over considerable distances to their site of action.

The bacteria in the gland of Deshayes (gills) transcribe lignocellulolytic enzymes, which later become part of the wood-degrading CAZyme cocktail found in the caecum fluids, as has already been described in previous papers (Waterbury *et al.*, 1983; Distel, 2003). Our transcriptomic data however show that the bacteria do not just produce enzymes of the same families transcribed by the shipworm; on the contrary it looks like that there is a degree of role separation, where the bacteria specialise in certain classes and the shipworm in others. Indeed, CAZymes of the family GH6 (cellobiohydrolases) are only produced by

the bacteria, which also take care almost entirely of the GH5s. This family includes different subfamilies, with GH5_53 (cellodextrinases), GH5_8 (endo-β-1,4-mannosidases), GH5_2 (endo-β-1,4-glucanases) and other non-specified subfamilies of GH5s found in many transcripts produced by the bacteria in the gills, while none in the digestive gland or caecum. Only two eukaryotic GH5_10s (endo-β-1,4-mannanases) and one unspecified GH5 are found in the proteome of the digestive glands, caecum or crystalline style, but their transcripts are not detectable even in the most highly expressed 10,000 transcripts of the glands, caecum or crystalline style sac. GH11s (endoxylanases) are also exclusively transcribed by the bacteria in the gills, and then found in the caecum; GH10s (also endoxylanases) are also prokaryotic, with one exceptions of a eukaryotic transcript in the digestive glands transcriptome and a protein encoded by a different transcript found in the proteome of the caecum. Furthermore, all the transcripts encoding for LPMOs of the family AA10, and consequently all the AA10s found in the caecum and gills proteomics, are entirely produced by the endosymbiotic bacteria. Finally, among the CE enzymes, which are overall not expressed in high amounts in the shipworm's digestive organs, CE15s are exclusively transcribed by the bacteria in the gills and then found in the caecum, while the other classes (CE1, CE3, CE10 and CE12) are almost entirely produced by the shipworm in both the digestive glands and caecum.

The shipworm instead seems to be taking care of producing all the GH1s and GH9s, which are annotated as endoglucanases or β-glucosidases, as well as all the GH2s (β-glucosidases, β-mannosidases and β-glucuronidases) and GH30s (endo-1,6-beta-D-glucanases) and almost all the GH45s (endoglucanases), with just the exception of one bacterial GH45 protein found in the caecum proteomics. Most of these transcripts are found in the digestive glands; however, the caecum also produces a certain number of GHs, which are mainly GH30s and GH1s, with some CE3s, GH13s, GH84s and GH18s. This result is in contrast with the findings of Shipway (2013), who suggested that the caecum is not a site of GH transcription and only plays a limited role in the production of wood-degrading enzymes. However, it is true that the range and amount of GHs produced in the caecum is limited compared to the digestive glands, and that only some GHs are related to wood degradation, such as GH1, GH30 and CE3; GH13s (amylases) instead could have a role in the degradation of phytoplankton and GH18s and GH84s (chitinases and N-acetyl β-glucosaminidase) in the defence against fungal pathogens (Kawabata *et al.*, 1996; Ekenler

and Tabatabai, 2004; Dahiya *et al.*, 2006), helping to keep the caecum free of pathogens and competitors for the sugars.

The transcriptomics results confirmed the role of the caecum in the digestion and absorption of wood, which was already suggested by the images obtained by TEM and the enzymatic assays presented in chapter 3. Indeed, the analysis of the transcripts involved in glucose transport have shown how these are highly transcribed in the caecum, compared to the digestive glands and the gills. This indicates that the sugar resulting from the degradation of the wood is taken up by the caecum tissues and used for metabolism. Finally, the transcriptomics show that the caecum is also involved in the production of some proteins involved in immunity, mainly those of the complement component and galectins, but also apoptosis inducing factor, β-1,3-glucan binding proteins and hemycentin-like proteins. Most of the other immunity related proteins are transcribed in the digestive glands and found in the caecum proteome, where they probably work in synergy with the ones produced by the caecum. This new finding is important because it indicates the mechanisms by which the shipworm manages to keep its endosymbiont population under control and its caecum almost completely bacteria-free (Betcher *et al.*, 2012).

When compared to other lignocellulose-eating animals, the organisation of the digestive system of the shipworms appears unique. Indeed, while it was once thought that only the symbiosis with cellulolytic microorganisms allowed animals to digest grass or wood, it is now well known that many invertebrates are able to produce endogenous cellulases, first discovered in plant-parasitic nematodes (Smant *et al.*, 1998) and termites (Watanabe, 1998), and later in molluscs (Sakamoto *et al.*, 2007; Tsuji *et al.*, 2013), in a total of 27 species of insects belonging to six different orders (Watanabe and Tokuda, 2010) and in crustaceans (King *et al.*, 2010; Milatovič and Štrus, 2010; Bui and Lee, 2015). However, the range of endogenous lignocellulose degrading enzymes so far discovered in these animals is limited, presenting mainly cellulase genes from the GH family 5, 9 and 45 (Cragg *et al.*, 2015). Glycoside hydrolases from family 7, which are cellobiohydrolases, are also known in filamentous fungi (Payne *et al.*, 2015), amoebas (Eichinger *et al.*, 2005) and marine isopods (King *et al.*, 2010), and GH10 (xylanases) in a freshwater bivalve (Sakamoto and Toyohara, 2009). The wood borer crustacean *Limnoria quadripunctata* instead revealed to possess a diverse range of CAZymes in its digestive transcriptome, which comprises GH7, GH9, GH5,

GH35, GH30 and others in minor amounts (GH2, GH13, GH16, GH18, GH20, GH31, and GH38) (King *et al.*, 2010). The shipworm's range of endogenous CAZymes appears even broader, including also GHs from the families 2, 71 and 84, some of which are found in high abundance in both the transcriptome and proteome of *L. pedicellatus*. This remarkable amount of endogenously produced CAZymes is reflected on the lower contribution to lignocellulose digestion performed by the endosymbionts, which are responsible for only 18.2% percent of the CAZymes found in the caecum proteome. However, as explained above, the symbiotic bacteria are mainly focused on the production of specific types of CAZymes, and for this reason are probably indispensable for a complete and efficient digestion of the lignocellulosic biomass. In particular, their contribution seems important when considering the oxidative breakdown on cellulose, which is performed by LPMO-type enzymes. Indeed, no endogenous LPMOs related to wood-degradation have been identified in the transcriptome or proteome of *L. pedicellatus*, while the only LPMOs found in the caecum are AA10s encoded by the symbionts. This is not the case for the hexapod *Thermobia domestica*, which has recently been discovered to produce endogenous LPMOs from the family AA15 (Sabbadin *et al.*, 2018a).

Lacking from the repertoire of wood-degrading enzymes in the digestive system of *L. pedicellatus* are enzymes for the deconstruction or modification of lignin. Indeed, as seen in the introduction (section 1.1.4), lignocellulosic material is composed by cellulose microfibrils and hemicellulose, coated and bound by the polyphenolic stress resistant lignin. Full access to the sugars contained in the cellulose and hemicellulose fractions is dependent on the removal or modification of lignin, otherwise part of the glucose remains inaccessible. As already discussed in section 1.2, bacteria and white-rot fungi possess ligninolytic enzymes such as peroxidases, laccases and superoxide dismutases (Bugg *et al.*, 2011; Pollegioni *et al.*, 2015), while brown-rot fungi use Fenton chemistry for lignin modification (Arantes and Goodell, 2014; Zhang *et al.*, 2016). Among terrestrial and marine wood-eating animals, so far only the marine wood borer *Limnoria* seems to be able to modify lignin to access the abundant glucose (Besser *et al.*, 2018). In the shipworms' digestive system, only one CAZyme classed as a laccase (family AA1_3) was identified in the digestive gland proteome, though it was in very low quantities and it was not found in the caecum fluids proteome or in the best 1,000 transcripts of any of the organs studied, suggesting that it might not be related to wood-degradation.

The analysis of the proteomics results of the four organs highlighted a good overall correspondence between the level of transcriptions of the CAZymes in the gills or in the digestive glands and the abundance of the relative proteins in the caecum fluids. Indeed, highly transcribed proteins like GH1s, GH9s, GH2s and GH5s are found in high amounts in the caecum. However, some CAZymes of the caecum seem to be very abundant while they are poorly transcribed in the gills or digestive glands, like the bacterial GH134s and eukaryotic GH35s and GH45s. Similarly, certain highly expressed transcripts (e.g. the bacterial AA10s and GH6s) are only found in small quantities in the caecum. These findings suggest that the level of translation efficiency could be poor for certain proteins, or indicate the presence of post-transcriptional and post-translational modifications (de Sousa Abreu *et al.*, 2009); protein stability and proteolysis may be important factors, particularly given the presence in the caecum of abundant proteases, as highlighted by the proteomics analysis (data not shown). Furthermore, the results indicate the presence of variability in the proteins produced by the bacteria in the gills, with similar but non-identical proteins identified in different shipworms specimens, possibly due to the presence of different strains of bacteria; this resulted in a more challenging matching of peptides during the shotgun proteomics procedure, possibly leading to the lack of identification of certain proteins. Finally, measurement errors, as well instruments sensitivity thresholds could have also contributed to the lack of correlation between certain proteins and their mRNA concentrations (Szallasi, 1999; Baldi and Long, 2001; Greenbaum *et al.*, 2003; de Sousa Abreu *et al.*, 2009). In *L. quadripunctata*, the GHs were found to make up only a small portion (1.74%) of the gut proteome, the organ where wood digestion takes place (Eborall, 2013). The classes that dominate the proteome are GH9, GH30 and GH35, but GH5, GH7 and GH16 were also identified. The shipworms' caecum proteome includes similar CAZymes classes but lacks any GH7 (cellobiohydrolases), which are possibly substituted by the symbiont-encoded GH6; GH1 GH2, GH134, GH13 and GH10 are also abundant CAZymes found in the shipworm proteome but not in *Limnoria*.

The analysis of the crystalline style proteome and the style sac transcriptome has provided us with first ever data on the crystalline style protein composition and origin for a bivalve mollusc, let alone a shipworm. The style is composed mainly of structural proteins that make up its scaffold and consistency, such as mucin, tubulin and collagen. According to the transcriptomics and microscopic results, these proteins are produced and secreted by the

secretory cells found in the tissues of the crystalline style sac, and then integrated in the forming style by the action of the cilia that line the sac. The style's carbohydrate degrading abilities are a result of the presence of CAZy proteins of the families GH5_8, GH5_53, GH5_10, GH9, GH1, GH134, GH16, GH35 and CE3. These CAZymes contribute to the overall lignocellulose degradation by performing extracellular digestion of the ingested wood before it reaches the caecum. The most surprising results involve the origin of the styles' CAZymes. Indeed, the style sac best 1,000 transcripts included only two wood-degrading enzymes, both of which of bacterial origin and that can therefore be attributed to contamination. The CAZymes found in the style proteomics are both prokaryotic and eukaryotic, the former produced by the endosymbiotic bacteria in the gills and the latter by the shipworm in the digestive glands, even though most of them are not highly expressed. While the presence in the crystalline style of an endogenous cellulase has been reported before for the freshwater bivalve *Corbicula japonica* (Sakamoto *et al.*, 2008), the amount and range of endogenous CAZymes found in the style of *L. pedicellatus* is unexpected and fascinating, suggesting that the crystalline style has a bigger than expected role in wood degradation in this - and possibly other - shipworm species. The presence of bacterial CAZymes in the style's proteins is also a new finding, which emphasises the level of cooperation between symbiont and host in the process of wood degradation.

The analysis of the non-CAZy proteins of the crystalline style has also produced new findings. Indeed the most abundant protein groups after structural proteins are those annotated as MACPF-proteins, hemocytin-like proteins and members of the complement component, all of which are involved in immunity (Peitsch and Tschopp, 1991; Yamakawa and Tanaka, 1999; Rosado *et al.*, 2008). Furthermore, the MACPF domain found in *L. pedicellatus* is C-terminally fused to an apextrin domain (ApeC), which has been shown to allow bacterial recognition and inactivation in other animals including marine bivalves (Gerdol and Venier, 2015) by its ability to bind to bacterial muramyl dipeptide (Huang *et al.*, 2014). In contrast, no lysozyme proteins have been recorded in the shipworm's style proteins, which are normally present in other bivalves and gastropods, where they are involved in the inactivation of bacteria for both protection and nutrition purposes (Nilsen *et al.*, 1999; Olsen *et al.*, 2003; Xue *et al.*, 2007; Takahashi and Itoh, 2011). In light of these findings, and the presence in the style of bacterial enzymes described above, we suggest that the crystalline style of *L. pedicellatus* could be involved in the degradation of the bacteriocytes that travel from the gills to the digestive system via the food groove. When

they bacteriocytes reach the mouth, they could be mixed with the food particles and reach the stomach, where they would be subject to mechanical disruption by the action of the style apex grinding on the gastric shield. The breaking up of the bacteriocytes could free single bacteria, which could then be attacked by the immunity-related proteins found in the crystalline style, in particular those that can recognise gram-negative bacteria, such as MACPF-containing proteins and those of the complement component. This mechanism, though not so far confirmed by unequivocal experimental evidence, could however explain how the bacteriocytes are degraded and how the CAZymes produced by the bacteria are released from the eukaryotic cells and later found as important components of the crystalline style and caecum fluids.

Chapter 5: Heterologous expression and characterisation of endogenous and bacterial lignocellulolytic enzymes from *Lyrodus pedicellatus*

5.1 Introduction

Chapter four presented the analysis of the transcriptome and proteome of the main digestive organs of *L. pedicellatus*. Examination of the resulting dataset led to the identification of putative enzymes responsible for the degradation of lignocellulose, as well as the tissues where these enzymes are produced and the organs where wood degradation mainly takes places. This investigation produced an unprecedented insight into the digestion of wood in shipworms. Nevertheless, the work performed so far would not be complete without the accomplishment of a further step necessary for a better knowledge of the CAZymes identified in the shipworm's organs. In order to perform the analysis presented in chapter four, the nucleotide sequences obtained from the RNA sequencing of the digestive organs were translated into open reading frames (ORFs) and annotated using bioinformatics tools. This allowed us to assign a function to the transcripts identified in the gills, digestive glands, caecum and crystalline style, and subsequently to the proteins found in the proteome of the same organs. This transcriptomic and proteomic approach allows relatively rapid and inexpensive insight into the digestive process, when compared to the alternative of determining gene and protein function by the expensive and time consuming experimental route (e.g. through functional assays, microarray analysis, RNA interference, etc.) (Gabaldón and Huynen, 2004; Jiang *et al.*, 2016). However, despite its undeniable value, the process holds some downsides, which have effects on the reliability of the annotation. Indeed, the methods used for protein function prediction are based mainly on sequence homology and protein domain identification, as well as evolutionary data, phenotypic profiles, protein-protein interaction and gene expression profiles. Although this is a powerful approach, the actual function of specific proteins needs to be experimentally verified, as these predictions are not completely accurate, particularly as in some cases sequence similarity can be low and minor differences in sequence can lead to significant difference in enzyme function (Bork and Bairoch, 1996; Doerks *et al.*, 1998; Rost *et al.*, 2003; Whisstock and Lesk, 2003; Manzoni *et al.*, 2016). It has been estimated that around 40-60% of sequences can be assigned to some aspect of function, but this can be not

inaccurate; there is also a considerable percentage of proteins that still remains uncharacterised (Andrade *et al.*, 1999). Furthermore, the protein function term is quite a broad one, since function can be assessed at different levels (such as biochemical, cellular, developmental, etc.) and proteins can have overlapping roles in more than one process or pathway (Rost *et al.*, 2003).

In light of this, there is a need to validate the annotation of the proteins identified in the shipworm's transcriptome and proteome. Since obtaining and testing high amounts of native proteins from an animal involves technical and ethical challenges, an alternative approach is the expression of the proteins in a host organism by the use of recombinant DNA technology, followed by purification and activity testing. This technique is called "heterologous protein expression" and is nowadays widely applied in many fields of science. Host organisms for protein production vary from single cells systems such as bacteria and yeast, to filamentous fungi and plants or cell cultures of more complex animals such as insects and mammals (Gomes *et al.*, 2016). The choice of the right expression system depends on the characteristic of the protein to be expressed (e.g. molecular weight, codon bias, presence of disulphide bridges or post-translational modifications, productivity) (Gustafsson *et al.*, 2004; Wurm, 2004; Gomes *et al.*, 2016), but generally a stepwise approach is preferable, with attempts first made in the easiest systems such as *Escherichia coli* (Gräslund *et al.*, 2008). In the case of lignocellulolytic enzymes, such as those identified in *L. pedicellatus*, there are many technical challenges to be considered, such as low protein production levels, poor secretion ability, decreased specific activity, glycosylation and high costs (Lambertz *et al.*, 2014). Furthermore, the marine environment from which these enzymes come from adds another complication related to the presence of multiple disulphide bridges (Hatahet *et al.*, 2014), which makes heterologous protein expression more challenging. Despite these technical challenges, heterologous protein expression was attempted for some of the most expressed or interesting *L. pedicellatus* lignocellulolytic enzymes, choosing the expression system predicted to be appropriate for the various targets.

5.2 Aims of the chapter

This chapter aims at testing the activity of some of the proteins identified in the proteome and/or transcriptome of *L. pedicellatus* digestive organs. This will be obtained by performing the following steps:

- selection of the targets from the transcriptomic and proteomic results presented in chapter four, on the base of protein abundance or transcript interest in the context of lignocellulose degradation;
- determination of the proteins' amino acid sequence and domain structure and of the main biochemical parameters that could affect protein expression and purification, such as molecular weight, theoretical pI, codon usage and presence of disulphide bridges. This will be performed using bioinformatics tools;
- utilisation of recombinant DNA techniques for the cloning of the gene sequence into an appropriate expression vector, transformation into competent cells and selection of the transfected cells;
- overexpression of the protein in the selected host and purification;
- thermal shift assay to determine the folding state of the protein;
- activity assays to determine whether the protein has the predicted activity (e.g. lignocellulose degradation or antibacterial activity). In the case of CAZymes or other lignocellulose degrading enzymes, different carbohydrate substrates will be tested to determine the enzyme's preference, as well as different environmental conditions such as pH, temperature and salinity.

5.3 Target enzymes selection and cloning

Due to time and budget constraints, only a limited number of targets were selected for cloning, expression and characterisation among the shipworm's transcriptome and proteome. The selection was performed on the basis of predicted CAZyme abundance in the caecum or crystalline style (the former because of its role in lignocellulose digestion, the latter because of its novel and unusual function related to wood degradation), or because the protein or transcript had an interesting annotation and included a signal peptide, hence it was a secreted protein and was not likely to have a function related to cellular metabolism. This was the case for some transcripts that were selected because

their annotation suggested they could be involved in lignocellulose degradation despite not being annotated as CAZymes, or that they could have antibacterial activity and therefore a role in controlling the proliferation of bacteria in the caecum or in the release of bacterial enzymes from bacterial cells transported from the gills via the food groove. Thirty targets were initially selected, but some were later abandoned due to new proteomic or transcriptomic results that gave different abundance estimates or annotations, or because of technical difficulties in the cloning of the transcripts. Cloning of some bacterial sequences was challenging, since their contigs were not full length and 3' and/or 5' RACE PCR approaches were needed to obtain the missing part of the sequence. Table 5.1 shows the remaining 18 sequences that were successfully amplified and for which heterologous expression was attempted in one or more host organisms. Six of them are shipworm proteins, five of which were annotated as highly similar to proteins from marine or freshwater bivalve molluscs. One target, annotated as "hephaestin-like", showed highest similarity to proteins from bird species, but was selected as it has a series of laccase-like domains and might be involved in lignin degradation. The other 12 proteins originate from the endosymbiotic bacteria, and most of them have annotations highly similar to marine gram-negative proteobacteria able to degrade complex carbohydrates such as cellulose, alginate and chitin.

Table 5.2 illustrates the details of the various expression attempts in different heterologous hosts. Some sequences were codon optimised for expression in *E.coli*, while the non-optimised sequences were used for expression in the other systems. The host organisms included *E.coli* (both for expression in the cell's cytoplasm and periplasm), the yeasts *Pichia pastoris* and *Saccharomyces cerevisiae*, insect cells from *Spodoptera frugiperda* and human embryonic kidney (HEK) cells. Different strains and vectors were used, depending on the expression hosts. The main purification tags used were strep and his, while other tags were used to improve protein solubility (Halo, GST, MBP, SUMO, TF, TRX and FLAG, table 5.3). *In vitro* translation was also attempted for four sequences, with no success. Among the 18 cloned targets, 12 were successfully expressed to produce recombinant proteins, but purification was not achieved for three proteins, and two more turned out to be incorrectly folded (as assessed by thermal shift analysis) and hence not active. A total of seven heterologously expressed active proteins were therefore obtained, of which five were of bacterial and two of animal origin. The next sections of this chapter will describe the results of the cloning, expression and characterisation of the five bacterial CAZymes.

Table 5.1. The targets of heterologous recombinant protein expression. The table shows the prokaryotic and eukaryotic proteins that were successfully amplified from the shipworm or symbionts cDNA and for which heterologous expression was attempted in various host organisms. Annotation was done against the NCBI non-redundant database. CAZy domains were searched using dbCAN. The table lists the name of the transcript from which the protein was cloned (contig name), whether the protein is eukaryotic (euk) or prokaryotic (prok) (domain), whether it has a signal peptide (signal peptide), its annotation, species and e-value and the CAZy modules identified in the protein sequence (CAZy domains).

Contig name	Domain	Signal peptide	Annotation	Species	E-value	CAZy domains
c174818_g2_i2	eukaryotes	yes	thioester-containing protein	*Azumapecten farreri*	0.0	none
c169024_g2_i2	eukaryotes	yes	mannan endo-1,4-β-mannosidase	*Mizuhopecten yessoensis*	8E-129	GH5_10
c178377_g1_i1	eukaryotes	yes	hephaestin-like protein	*Egretta garzetta*	0.0	none
c176563_g3_i3	eukaryotes	yes	cellulase	*Corbicula japonica*	2.2E-74	GH9
c176623_g3_i4	eukaryotes	yes	DBH-like monooxygenase protein 1	*Mizuhopecten yessoensis*	6E-146	none
c164680_g1_i4	eukaryotes	yes	MACPF domain-containing protein 4	*Mytilus galloprovincialis*	3E-112	none
c176007_g2_i1	prokaryotes	yes	cellobiohydrolase	*Microbulbifer thermotolerans*	1.4E-41	GH6
c177001_g3_i1	prokaryotes	yes	beta-mannosidase	*Teredinibacter sp.1162T.S.Oa.05*	6.4E-16	GH5_8+CBM10
c173140_g2_i3	prokaryotes	yes	hypothetical protein	*Alteromonadaceae bacterium Bs12*	1E-171	GH10
c169869_g1_i2	prokaryotes	yes	1,4-beta-xylanase	*Micromonospora rifamycinica*	2E-83	GH11+CBM10+CBM5
c122374_g1_i1	prokaryotes	yes	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium Bs12*	2E-127	GH134+ CBM10+CBM10+CBM10
c180176_g1_i1	prokaryotes	yes	plant cell wall polysaccharide active protein	*Alteromonadaceae bacterium Bs12*	7E-104	GH134+ CBM10+CBM10+CBM10
c173837_g2_i2	prokaryotes	yes	auxilliary activity family 10 domain-containing protein	*Alteromonadaceae bacterium Bs12*	6E-92	AA10+CBM10
c169443_g1_i1	prokaryotes	yes	glucose dehydrogenase	*Brevundimonas subvibrioides*	4E-114	none
c251108_g1_i1	prokaryotes	yes	glycoside hydrolase family 5 domain protein	*Teredinibacter turnerae T7901*	0.0	GH5_4+CBM10+CBM5
c91252_g1_i1	prokaryotes	yes	acetylxylan esterase / xylanase	*Teredinibacter turnerae*	0.0	GH10+CE6+CBM10+CBM5+CBM60
c9656_g1_i1	prokaryotes	yes	endo-1,4-beta-xylanase	*Teredinibacter turnerae*	0.0	GH11+CE4+CBM10+CBM60
c167805_g1_i5	prokaryotes	yes	glycoside hydrolase family 5	*Teredinibacter turnerae*	0.0	GH5_53+GH11+CBM5+CBM10

Table 5.2. The expression attempts for the target prokaryotic and eukaryotic proteins. The table lists the name of the transcript from which the protein was cloned (contig name), whether the protein sequence was codon optimised for *E. coli* (cod. opt.=codon optimisation), the different host organisms used for the heterologous expression (host) and the strain used (strain), the vectors used for cloning (vector), the purification or solubility tags added to the sequence (tag), whether the protein was successfully expressed and purified (expr/purif) and if it was properly folded, therefore an active protein (activity), which was assessed by a thermal shift assay (see section 2.7.3 for information on materials and methods).

Contig name	Cod. opt.	Host	Strain	Vector	Tag	Expr/ purif	Act.
c174818_g2_i2	no	*E.coli* (cytoplasmic), *P. pastoris,* mammalian cells	Origami 2(DE3), CF034 10a, HEK 293T	pET52b(+), pPICZαB, pOPING, pOPING3Flag, pOPINGG2Strep,	strep, his	no	n/a
c169024_g2_i2	yes	In vitro, *E.coli* (cytoplasmic and periplasmic), *P. pastoris,* mammalian cells	Origami 2(DE3), Rosetta 2(DE3), Lemo21(DE3), BL21(DE3), CF034 10a, HEK	PURExpress kit (NEB), pET52b(+) , pETFPP30, pPICZαB , pOPING, pOPING3Flag, pOPINGG2Strep, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C,	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/no	n/a
c178377_g1_i1	yes	In vitro, *E.coli* (cytoplasmic and periplasmic), *S. cerevisiae, P. pastoris,* mammalian cells	Origami 2(DE3), Rosetta 2(DE3), Lemo21(DE3), BL21(DE3), BJ2169, CF034 10a, HEK 293T	PURExpress kit (NEB), pET52b(+) , pETFPP30, pYES2, pPICZαB , pOPING, pOPING3Flag, pOPINGG2Strep, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C, pOPINTF, pOPINTRX	strep, his, Halo, GST, MBP, SUMO, TF, TRX, FLAG	yes/yes	yes
c176563_g3_i3	yes	In vitro, *E.coli* (cytoplasmic and periplasmic), *P. pastoris,* mammalian cells	Origami 2(DE3), Rosetta 2(DE3), Lemo21(DE3), BL21(DE3), BJ2169, CF034 10a,	PURExpress kit (NEB), pET52b(+) , pETFPP30, pPICZαB, pOPING, pOPING3Flag, pOPINGG2Strep, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C,	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/no	n/a
c176623_g3_i4	no	*E.coli* (cytoplasmic), *P. pastoris,* mammalian cells	Origami 2(DE3), CF034 10a, HEK 293T	pET52b(+), pPICZαB, pOPING, pOPING3Flag, pOPINGG2Strep	strep, his	yes/no	n/a
c164680_g1_i4		*E.coli* (cytoplasmic), *P. pastoris*	Origami 2(DE3), Rosetta 2(DE3), CF034 10a	pET52b(+), pPICZαB	strep, his	yes/yes	yes
c176007_g2_i1	yes	In vitro, *E.coli* (cytoplasmic and periplasmic), insect cells	Origami 2(DE3), Rosetta 2(DE3), BL21(DE3), Sf9	PURExpress kit (NEB), pET52b(+), pETFPP30, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C, pOPINTF, pOPINTRX pOPINF, pOPINE	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/yes	no
c177001_g3_i1	no	*E.coli* (cytoplasmic and periplasmic), insect cells	Origami 2(DE3), Rosetta 2(DE3), BL21(DE3), Rosetta-Gami2(DE3), Sf9	pET52b(+), pETFPP30, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C, pOPINTF, pOPINTRX pOPINF, pOPINE	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/yes	yes
c173140_g2_i3	no	*E.coli* (cytoplasmic and periplasmic), insect cells	Origami 2(DE3), Rosetta 2(DE3), BL21(DE3), Rosetta-Gami2(DE3), Sf9	pET52b(+), pETFPP30, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C, pOPINTF, pOPINTRX pOPINF, pOPINE	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/yes	no

c169869_g1_i2	no	*E.coli* (cytoplasmic and periplasmic), insect cells	Origami 2(DE3), Rosetta 2(DE3), BL21(DE3), Rosetta-Gami2(DE3), Sf9	pET52b(+), pETFPP30, pOPINHALO7, pOPINJ, pOPINM, pOPINS3C, pOPINTF, pOPINTRX pOPINF, pOPINE	strep, his, Halo, GST, MBP, SUMO, TF, TRX,	yes/yes	yes
c122374_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	yes/yes	yes
c180176_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	yes/yes	yes
c173837_g2_i2	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	yes/yes	yes
c169443_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	no	n/a
c251108_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	no	n/a
c91252_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	no	n/a
c9656_g1_i1	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	no	n/a
c167805_g1_i5	no	*E.coli* (cytoplasmic and periplasmic)	BL21(DE3), Rosetta-Gami2(DE3)	pET52b(+), pETFPP30	strep	no	n/a

Table 5.3. Solubility tags used for recombinant protein expression.

Short name	Long name	Source	Amino acids number	Molecular weight	Purpose
Halo	Haloalkane dehalogenase	Prokaryotic	314	35.3	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility; visualization of the subcellular localization of a protein of interest; immobilization of a protein of interest.
GST	Glutathione *S*-transferases	Eukaryotic/ prokaryotic	220	26.9	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility; purification tag.
MBP	Maltose binding protein	Prokaryotic	389	42.5	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility; purification tag.
SUMO	Small ubiquitin-like modifier	Eukaryotic	115	13.2	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility.
TRX	Thioredoxin	Prokaryotic	127	13.8	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility; improves protein crystallisation because it readily forms several crystals itself.
FLAG	Flag-tag	Artificial	8	-	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility; purification tag; isolation of protein
TF	Trigger factor	Prokaryotic	450	50.2	Enhances functional protein production in prokaryotic and eukaryotic expression systems, based upon significantly improved protein stability and solubility.

5.4 The bacterial *Lps*GH5_8

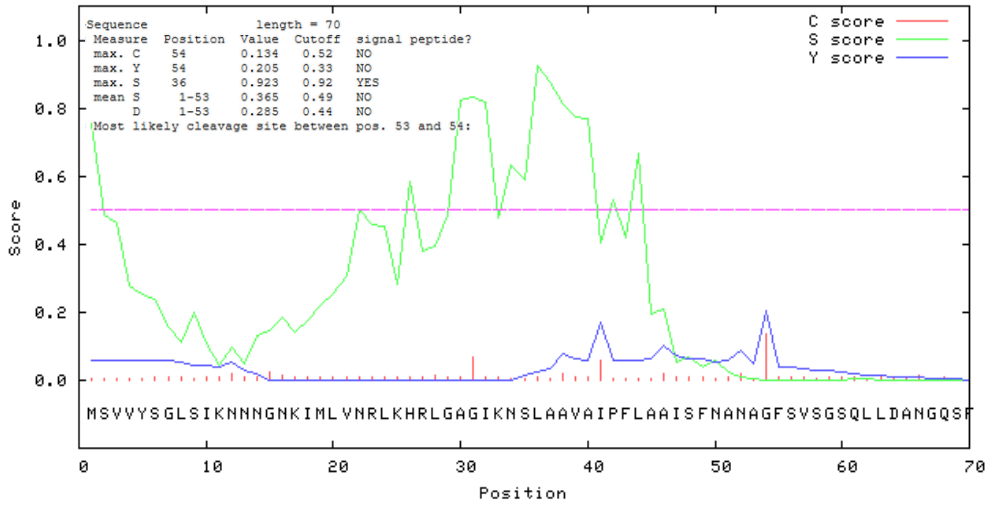
5.4.1 Sequence analysis and cloning

The nucleotide sequence encoding for the bacterial protein *Lps*GH5_8 was obtained from the contig c177001_g3_i1 of the *L. pedicellatus* transcriptome presented in chapter 4.3, representing the 18th most transcribed CAZyme of the gills. Table 5.1 and 5.2 show its annotation details and expression attempts. The main protein parameters were predicted using the tools described in materials and methods (section 2.4.3) and are shown in Figure 5.1. The protein comprises of 444 amino acid residues (Fig. 5.1A) and displays a signal peptide for secretion at the N-terminus, which includes the first 53 amino acids (Fig. 5.1B). A GH5 subfamily 8 and a CBM10 domain (which has cellulose binding function) follow the signal peptide and are separated by a poly-serine linker (Fig. 5.1C). The protein has a theoretical isoelectric point (pI) of 4.65 and an expected molecular weight (MW) of 46.5 kDa, and it was predicted to have five possible disulphide bonds.

A

***Lps*GH5_8**
MSVVYSGLSIKNNNGNKIMLVNRLKHRLGAGIKNSLAAVAIPFLAAISFNANAGFSVSGSQLLDANGQSFVMRGVNHP HAWYTGETGSFADIAATGANTVRVVLSNGHQWTRNSASDVANVISLCKQNQLICVLEVHDATGGGEEGAAGTIAGAA QYWVDIASALQGEEDYVIINIANEPVGNNQPSSMWINDHISAIQKLRNAGLTHTLMIDAPNWGQDWENVMRNNAP QVAAADSLNNTMFSVHMYQVYQDYSTINGYLTGFVSDHNLPIIVGEFGGDHQGDEVDEDSILQVAQQNGIGYLGWS WSGNGDCCTTLDIVNNFNGSSLTSWGNRLINGTNGIAQTSQRASVYGGGSTTSSSSSSSSSNNSSSSSSSSSSSSSSS SSSGGGACTECNWYGSTYDLCNNQNSGWGWENQQSCIGIDTCEGQNGNGGPICSGS-

B



C

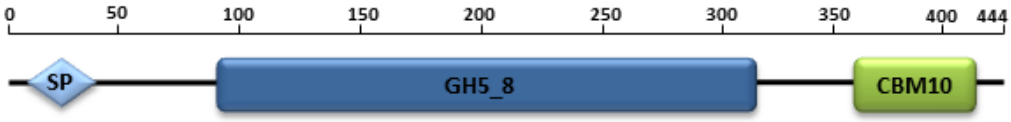


Figure 5.1 *Lps*GH5_8 predicted sequence and domains. **A.** The amino acid sequence obtained by translating the nucleotide sequence with the ExPASy translate tool. Red indicates the signal peptide and the symbol – the stop codon. **B.** Prediction of signal peptide cleavage site presence and location obtained with the SignalP 3.1 Server. **C.** The domain structure obtained with dbCAN. SP=signal peptide.

The *Lps*GH5_8 sequence (with its CBM) was amplified from *L. pedicellatus* gill's cDNA (Fig. 5.2A) and purified. The primers used for the PCR, the thermo-cycling conditions and the detailed methods for the cloning of the protein are described in section 2.5. Ligation cloning into the cloning vector pSC-B-amp/kan was followed by white-blue screening of the transformed cells and confirmation by colony PCR of the presence of the inserted gene for LpsGH5_8 (Fig. 5.2B). The plasmids were extracted from the cells by miniprep and the presence of the correct sequence was verified by Sanger sequencing; the validated amino acid sequence was deposited in the European Nucleotide Archive (ENA) (accession number PRJEB28739). PCR was used to amplify the sequence with primers carrying a 15 base pairs extensions complementary to the ends the vector pET52b(+) and to linearise the vector by reverse PCR at the KpnI restriction site. In-Fusion cloning was used to clone the gene into the vector pET52b(+), which has an N-terminal Strep-tag II and subsequent human rhinovirus (HRV) 3C cleavage site for protein purification, as well as 10X His-tag at the C-terminus. The obtained construct was transformed in *E. coli* Rosetta-Gami 2(DE3) competent cells and the positive colonies were screened by colony PCR (Fig. 5.2C).

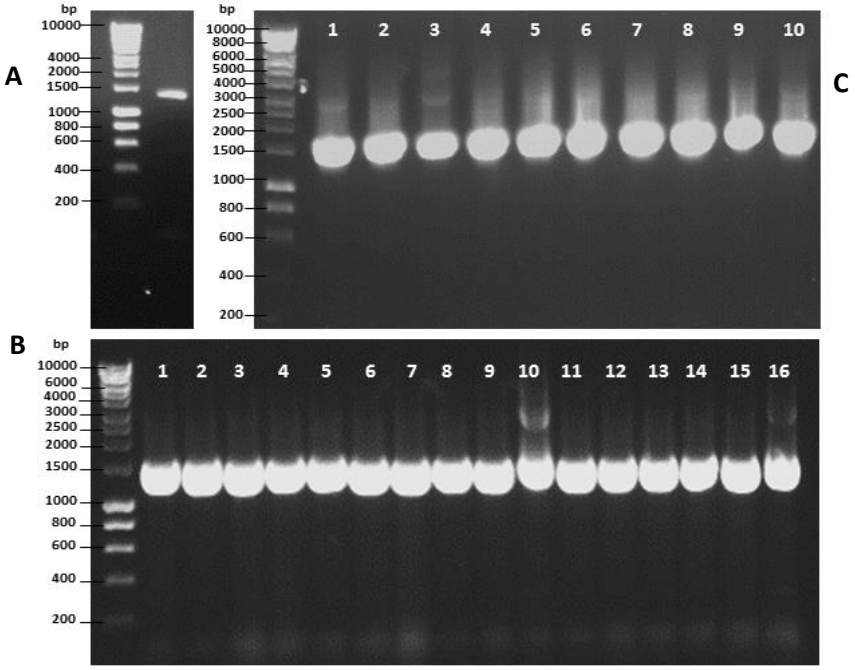


Figure 5.2. Gel electrophoresis of the cloning steps of *Lps*GH5_8. **A**. Initial amplification of the gene from *L. pedicellatus* gill's cDNA. The PCR product lies at the right height for the predicted gene size of 1.332 base pairs. **B.** Colony PCR performed to confirm the presence of the *Lps*GH5_8 insert in the cells transformed with the cloning vector pSC-B-amp/kan. All the 16 screened colonies contained an insert of the right length for *Lps*GH5_8. **C.** Second colony PCR performed to confirm the presence of the *Lps*GH5_8 insert in the cells transformed with the expression vector pET52b(+). All the ten screened colonies contain an insert of the correct size. The ladder used for the three gels was the HyperLadder 1kb from Bioline.

5.4.2 Expression

Rosetta-Gami 2(DE3) cells from colony two of Fig. 5.2C were grown in 100 ml of LB medium and induced with 0.5 mM IPTG, and were harvested after incubation at 20 °C, 30 °C and 37 °C for 20 hours (details of the expression and purification steps are in chapter 2.6.3). After cell lysis, the non-induced and induced soluble fractions were electrophoresed into a SDS-PAGE gel to confirm the presence of *Lp*sGH5_8 and compare expression levels at the different temperatures. Figure 5.3A-B shows that expression of recombinant soluble *Lp*sGH5-8 was successful at 20 °C, with small quantities obtained also at 30 °C, while none at 37°C.

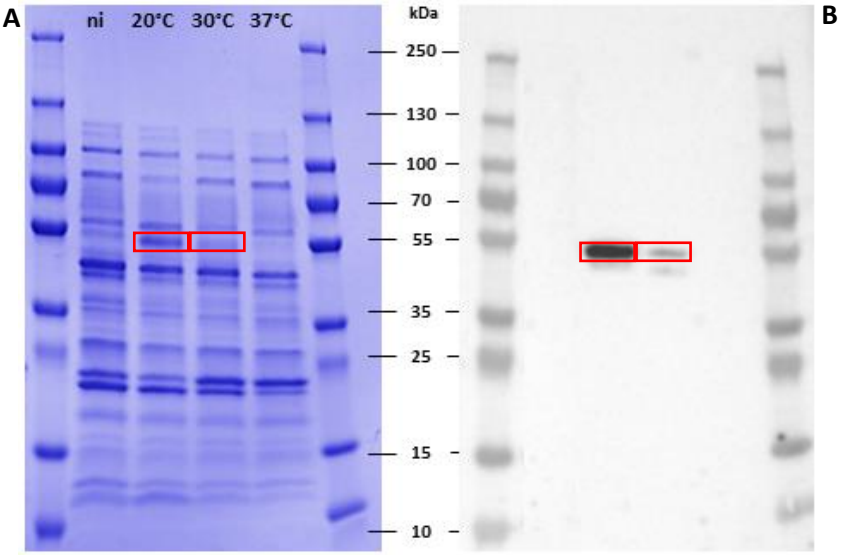


Figure 5.3. Analysis of the expression trial of *Lp*sGH5_8. **A.** SDS-PAGE of the soluble fraction obtained from 100 ml culture of non-inducted (ni), and induced cells at 20, 30 and 37 °C. **B.** Western blot anti-Strep analysis of the same samples as in A. The expected MW of *Lp*sGH5_8 is ~46.5 kDa, which is in line with the product seen in both in A and B (red rectangles). The loaded samples were 30 µl in volume. The protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific.

Expression was scaled up to one litre of LB medium, following the same protocol and incubation at 20 °C after induction. The resulting soluble cell fraction was filtered and subjected to affinity chromatography by running through a 5 ml StrepTrap column and eluting with 2.5 mM desthiobiotin in PBS. Figure 5.4A-B shows the purification chromatogram and the SDS-PAGE analysis of the eluted fractions containing absorbance peaks. Fraction 6 contained most of the purified protein, with minor quantities also seen in fractions 7 and 8. The SDS-PAGE gel shows the presence of two adjacent bands at a MW similar to that expected for *Lp*sGH5_8. To investigate the nature of the two bands, protein identification by MALDI-TOF mass spectrometry was performed, and both of them were

identified as *Lp*sGH5_8, but no differences in the amino acids composition could be identified, though it is possible that one band represents a shorter form of the protein caused by protease degradation.

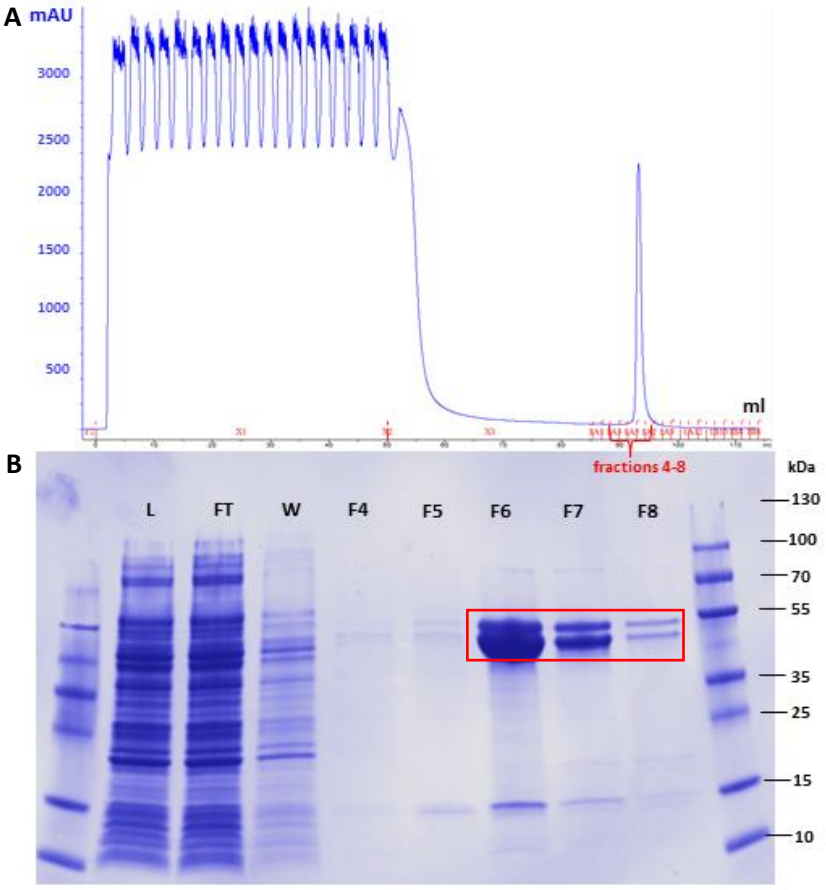


Figure 5.4. Affinity purification of *Lp*sGH5_8. **A.** Chromatogram of the purification performed with a 5 ml StrepTrap column. **B.** SDS-PAGE analysis of the purification presented in A. L=lysate, FT=flow through, W=wash, F4 to F8 are the fractions from A. The volume of the loaded samples was 30 µl and the protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific. The expected MW of *Lp*sGH5_8 is ~46.5 kDa, which is in line with the main band seen in fractions 6-8 in B (red rectangle); other minor bands are seen due to the purification not being completely efficient.

Fractions 6 and 7 from the affinity purification were combined and the Strep-tag was successfully cleaved using the Turbo HRV 3C protease, procedure that was confirmed by SDS-PAGE and western blot analysis (not shown). Gel filtration in PBS was performed to isolate the pure protein, and the relevant peaks were verified by SDS-PAGE analysis. Figure 5.5A-B shows the purification chromatograms and the SDS-PAGE analysis of the eluted fractions containing absorbance peaks. Peak number two contained *Lp*sGH5_8, which resulted at an adequate level of purity for downstream activity assays. The fractions from peak two were therefore combined together, concentrated to 2 ml at 0.5 mg/ml (verified with a Bradford assay), for a total of 1 mg of protein. A thermal shift assay (Fig. 5.5C)

confirmed the correct folding of the protein, which provided a melting temperature (Tm) of 43.0 °C (R^2=0.999).

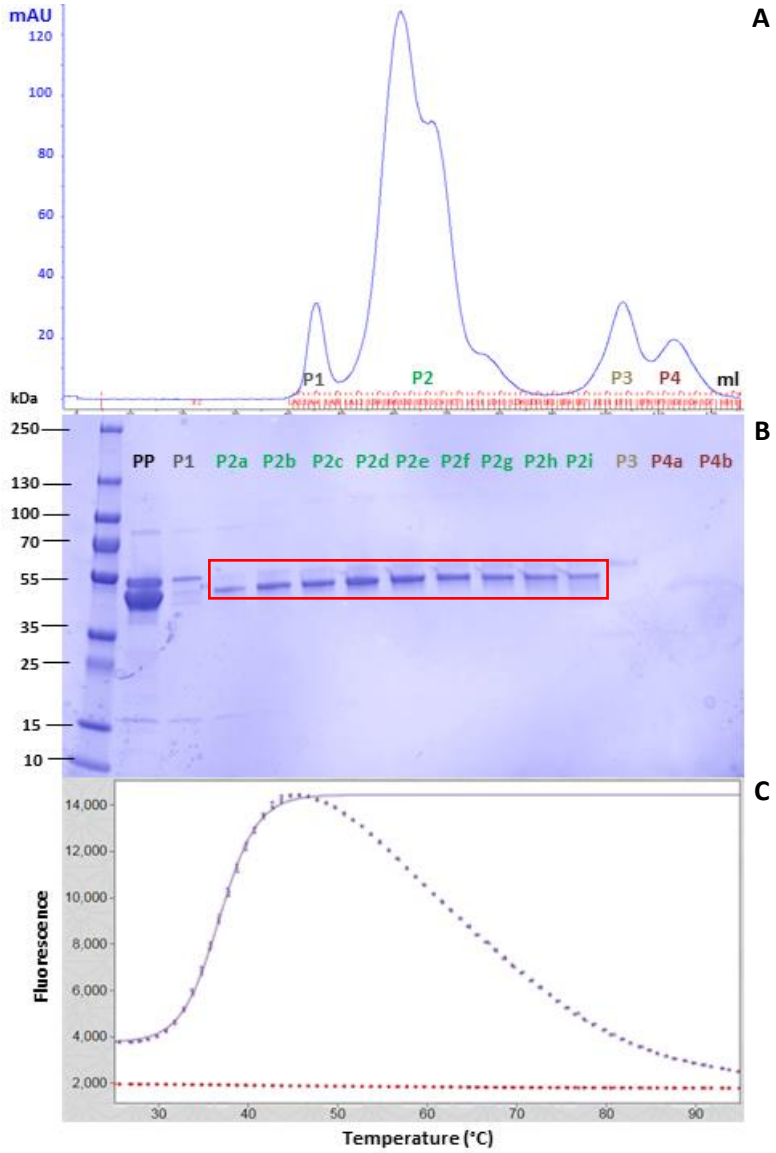


Figure 5.5. Gel filtration of *Lps*GH5_8. **A.** Chromatogram of the gel filtration performed with a HiLoad 16/600 Superdex 75 pg column. **B.** SDS-PAGE of the peaks from the purification presented in A. The volume of the loaded samples was 15 µl and the protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific. PP=pre-purification, P1=peak 1, P2=peak 2, P3=peak 3, P4=peak 4. The expected MW of *Lps*GH5_8 is ~46.5 kDa, which is in line with the main band seen in fractions from peak two (red rectangle). **C.** Thermal shift assay for the pulled and concentrated fraction from B (purple) and for the buffer used as a negative control (red).

A version of *Lps*GH5_8 without its CBM module was also cloned, expressed and purified with identical methods and similar results to those for the full-length sequence (data not shown). Figure 5.6 presents the chromatograms of the gel filtration and the SDS-PAGE analysis performed after the affinity purification and removal of the Strep-tag. Each eluted

fractions containing *Lps*GH5_8 without CBM was confirmed as correctly folded by the use of a thermal shift assay (Fig 5.6C) and it was then combined with the others and concentrated to 0.7 mg/ml in a volume of 2 ml, for a total of 1.4 mg of protein. The T_m of the fraction ranged from 45.0 to 45.6 °C (R^2=0.997-0.998).

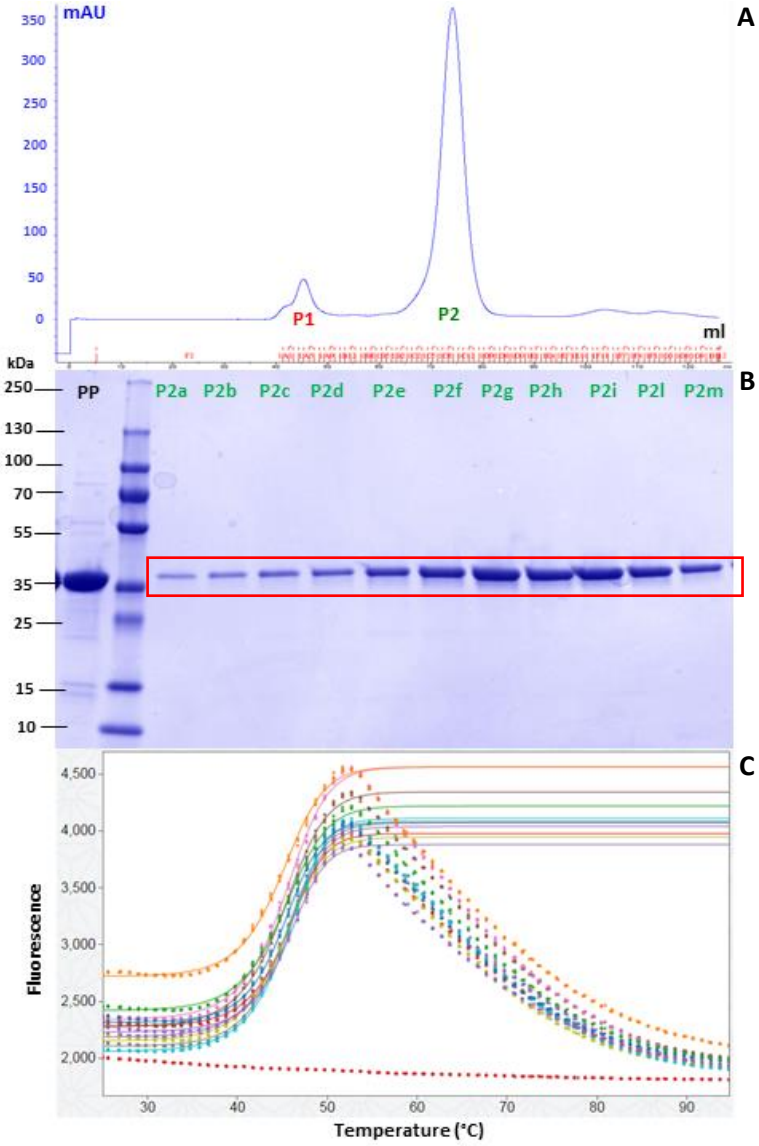


Figure 5.6. Gel filtration of *Lps*GH5_8 without CBM. **A.** Chromatogram of the gel filtration performed with a HiLoad 16/600 Superdex 75 pg column. **B.** SDS-PAGE of the peaks from the purification presented in A. The volume of the loaded samples was 15 µl and the protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific. PP=pre-purification, P2=peak 2. The expected MW of *Lps*GH5_8 without CBM is ~35.4 kDa, which is in line with the main band seen in fractions from peak two (red rectangle). **C.** Thermal shift assay for each of the fractions analysed in the SDS=PAGE presented in B; the buffer used as a negative control is in red.

5.4.3 Activity assays

5.4.3.1 Activity against polysaccharides

A search of *Lp*sGH5_8 amino acid sequence against the NCBI nrdb finds similarities to β-mannosidases of various gammaproteobacteria, and dbCAN recognises in the sequence a GH5 domain of the subfamily 8, which is thought to include only extracellular endo-β-1,4-mannosidases (Aspeborg *et al.*, 2012). However, the CAZy database reports for this subfamily activity on different substrates such as mannose, xylose, glucose, lichenin, galactose and others. To test which substrates *Lp*sGH5_8 is active on, polysaccharides were incubated with the enzyme and the reaction products analysed for reducing sugars using dinitrosalicylic acid (DNS) reducing sugar assays on a range of different polysaccharides. The enzyme was tested both as its full-length version and without the CBM10, which has been shown to have cellulose-binding activity (Raghothama *et al.*, 2000). Figure 5.7A presents the results of the activity assays to test substrate preference, which were performed by incubating the protein for two hours in sodium phosphate buffer pH 8.0 at 30 °C and 320 rpm. The assay shows that both version of the protein are able to degrade glucomannan and galactomannan, with some activity also recorded for LBG and arabinogalactan, but had little activity on β-galactan. The presence of the CBM does not seem to contribute significantly, as the activity of the two protein versions was found to have only minor and inconsistent differences. The protein does not seem to be able to degrade xylan substrates or cellulose (both in its water-soluble form CMC or PASC and surprisingly was also not active on mannan. The protein was also tested on milled scots pine, but no activity was noted with this assay.

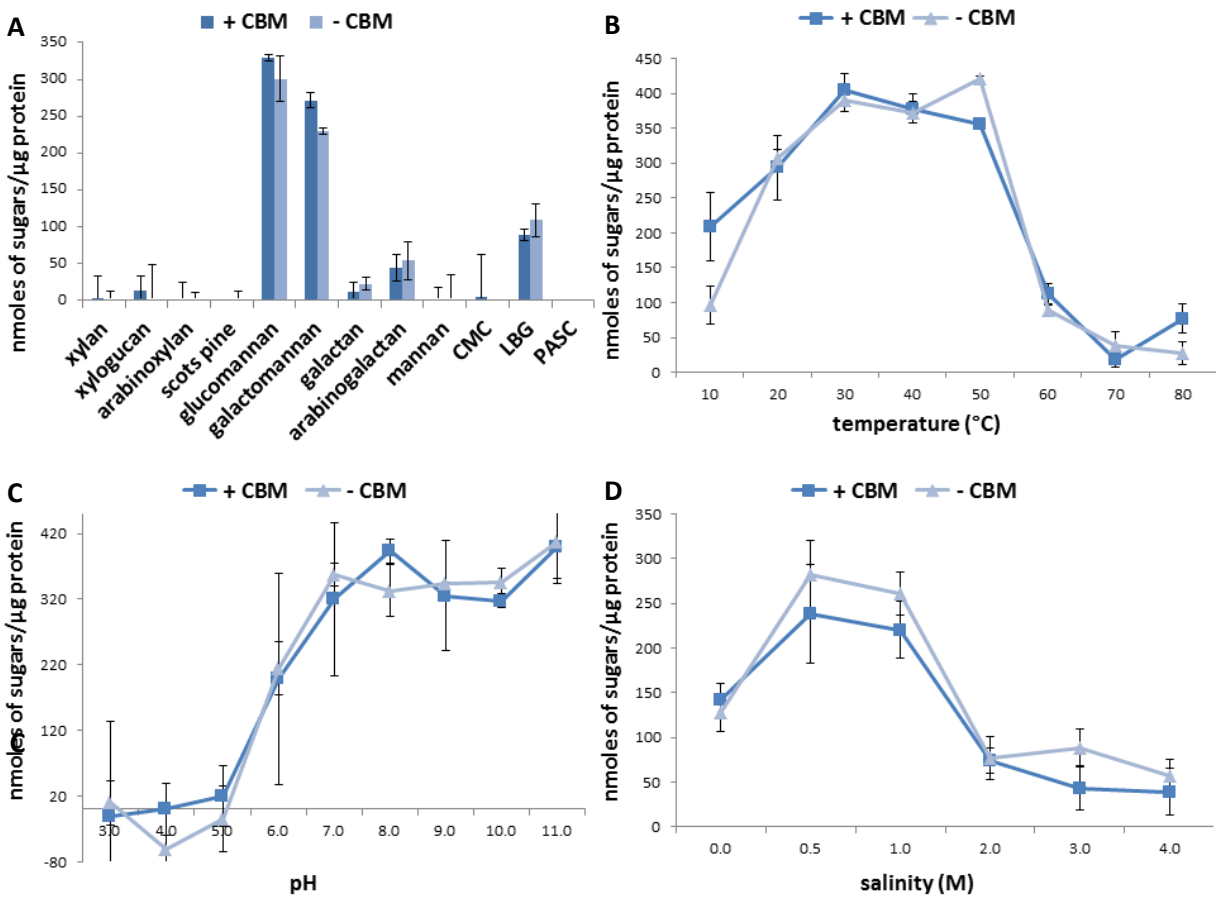


Figure 5.7. DNS reducing sugar assays for *Lps*GH5_8 with and without CBM. **A.** The protein was tested on a range of polysaccharides at pH 7.8, 30 °C and 320 rpm for two hours. CMC= carboxymethyl cellulose, LBC=locust bean gum, PASC= phosphoric acid swollen cellulose. **B**, **C** and **D** show the protein activity at different temperatures, pHs and salinities, respectively.

The protein preference for different temperatures, pHs and salinities was tested with the same methods, and the results are shown in Figure 5.7B-D. *Lps*GH5_8 preferred range of temperature seems to be between 30 and 50 °C; higher temperatures cause a sudden and dramatic drop in activity, while at lower temperatures the activity is substantially reduced. These results are in line with the protein melting temperature of 43-45 °C (with-without CBM) recorded by thermal shift assay. The protein is not active at acid pH, while activity starts at pH 6.0 and peaks at 7.0-8.0, with the higher range of alkaline pH (8.0-11.0) showing unaltered activity, indicating a tolerance to alkaline environments. The range of preferred salinity is between 0.5-1.0 M NaCl, with higher salt concentrations causing a sudden loss of activity. These results are very similar between the two versions of the protein with or without CBM, with only minor differences that are not enough to suggest any contribution of the CBM in the overall protein activity levels.

5.4.3.2 Heat maps

The results of the DNS reducing sugars assays show the enzymes preference for pH and temperature, but these parameters have been tested separately. In order to identify the enzymes optimal activity range at pH and temperature values combined together, a heat map for *Lp*sGH5_8 with its CBM was produced, following the protocol developed by Herlet and colleagues (2017) and described in section 2.8.2. Briefly, 96 well PCR plates were prepared containing *Lp*sGH5_8 together with its preferred substrate (glucomannan) and buffer (at a pH ranging from 4.5 to 8.0) and were heated at temperatures between 30 and 74 °C in a thermal cycler using a temperature gradient. The results are shown in Figure 5.8.

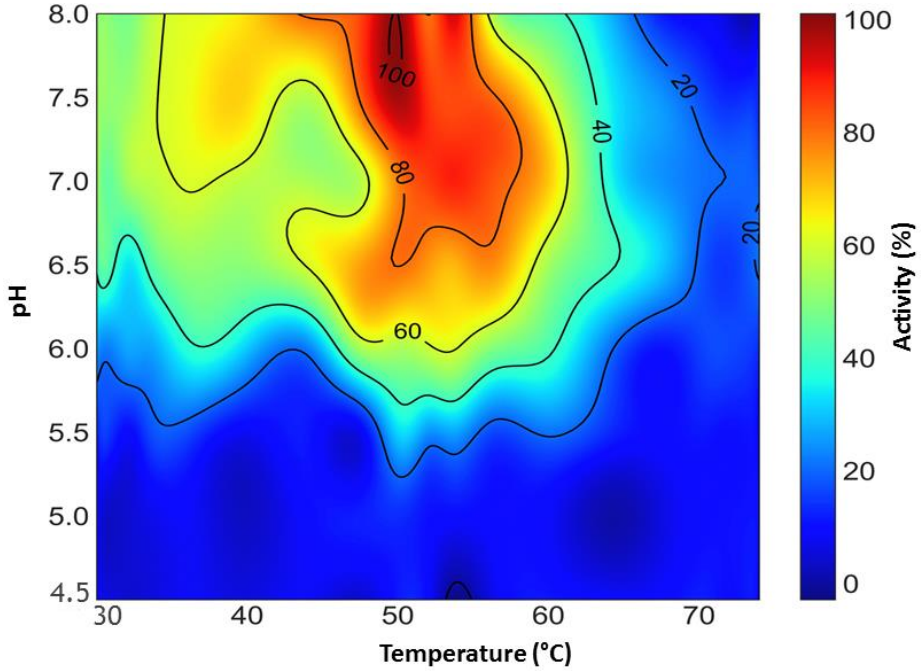


Figure 5.8. Heat map for *Lp*sGH5_8. The substrate was glucomannan (1%), the temperature range 30 to 74 °C, the pH range 4.5 to 8.0 (citrate buffer) and 10 µl of protein (at 0.5 mg/ml) were leaded in each well.

The heat maps provided an immediate visual summary of the protein activity at different pH and temperatures, and highlighted the protein hot spot. *Lp*sGH5_8 showed best activity for pHs ranging from 6.0 to 8.0 (and would have probably shown activity for a higher pH if they had been included in the experiment set up) and for temperatures between 35 to 60 °C; a hot spot was identified at the pH of around 7.6 and temperature of 51 °C. The results obtained with the heat maps are similar to those obtained with DNS sugar assays performed with separate measurements for the two parameters, but would benefit from testing the protein at higher pH in order to obtain a more complete map.

5.4.3.3 Polysaccharide analysis using carbohydrate gel electrophoresis

The DNS reducing sugar assays provided evidence on *Lps*GH5_8 substrate, temperature, pH and salinity preferences, allowing us to confirm the annotation of beta-mannanases and to determine the preferred environmental conditions; to obtain more detailed information on the mechanism of enzyme action, polysaccharide analysis using carbohydrate gel electrophoresis (PACE) was performed. This separation technique allows visualisation of the products of the digestion of a protein with the chosen substrates, therefore providing qualitative and quantitative characterisation of enzymatic activity, such as the enzyme site of hydrolysis along the carbohydrate chain and therefore the products formed during the digestion (Goubet *et al.*, 2002; Kosik *et al.*, 2012). The purified *Lps*GH5_8, with and without CBM, was mixed with the substrates, which were chosen on the base of availability and of the known protein substrate preference; scots pine was included to investigate the digestion products of the wood where *L. pedicellatus* was grown for the proteomic and transcriptomic studies presented in Chapter 4. The reaction was left at 30 °C overnight before the removal of the undigested polysaccharides with ethanol precipitation. The dried digestion products were than separated in a polyacrylamide gel (for details on the method see section 2.8.3) and they are shown in Figure 5.9. These assays were performed by Marta Busse-Wicher at the Department of Biochemistry, Cambridge University.

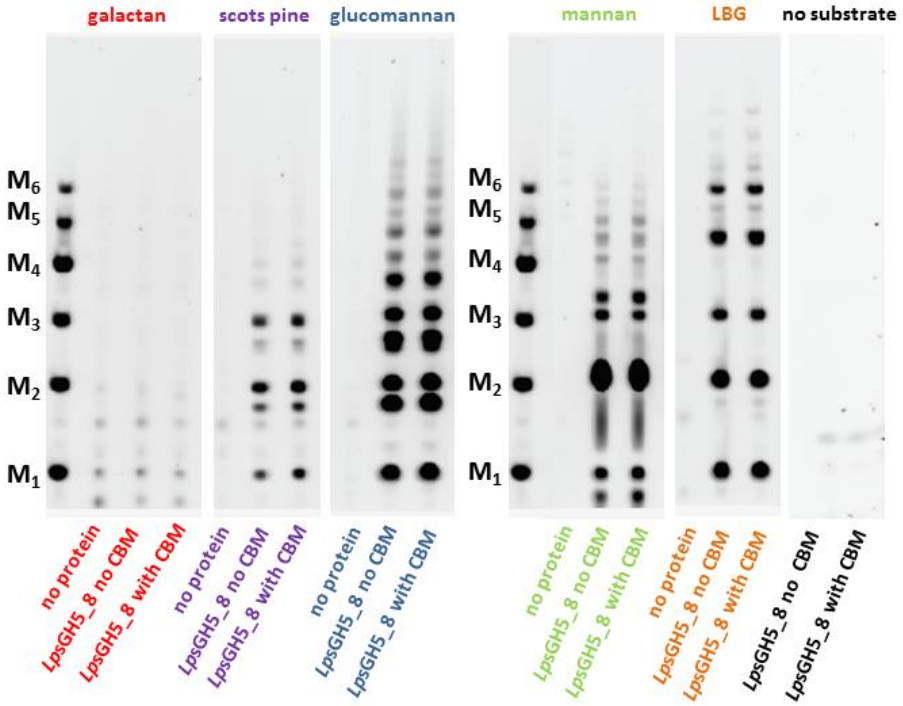


Figure 5.9. PACE analysis performed on selected substrates for *Lps*GH5_8 with and without CBM. 1 µg of protein was used for each reaction. M1 to M6 represent the mannose standards containing from one to six molecules of mannose. Negative controls are those with no protein (right panel) or no substrates (first lane of each panel). LBG=locust bean gum.

The assay suggests that *Lp*sGH5_8 is active on a range of mannans, releasing mannose, mannobiose and longer mannoolisaccharides from glucomannan, LBG and mannan. Contrary to the DNS reducing sugar assays, products were also seen on scots pine, indicating activity on this wood. Faint bands are visible on the gel presenting the digestion with galactan, denoting minor activity on this substrate, similar to what is seen for the DNS reducing sugar assay (Fig. 5.7A). Bands not corresponding to the mannose standards are also visible in the gel, suggesting that *Lp*sGH5_8 can cleave sugar decorations along the mannose chain; the nature of these sugars could be assessed by re-running the PACE assay using standards for other polysaccharides.

5.5 The bacterial *Lp*sGH134a and *Lp*sGH134b

5.5.1 Sequence analysis and cloning

The nucleotide sequences encoding for the bacterial proteins *Lp*sGH134a and *Lp*sGH134b were obtained from the contigs c122374_g1_i1 and c180176_g1_i1 of the *L. pedicellatus* transcriptome presented in chapter 4.3. The former is one of the most abundant bacterial CAZymes of the caecum, gills and crystalline style proteome, while the latter is only found in the crystalline style. Table 5.1 and 5.2 show their annotation details and expression attempts. The main protein parameters were predicted using the tools described in materials and methods (section 2.4.3), and are shown in Figure 5.10. *Lp*sGH134a comprises of 393 amino acid residues and *Lp*sGH134b of 433 (Fig. 5.10A), and they both display a signal peptide for secretion at the N-terminus, which includes the first 22 amino acids for the former and 53 for the latter (Fig. 5.10B). Three CBM10 domains are found after the signal peptide, followed by a GH134 domain, and the modules are separated by poly-serine linkers (Fig. 5.10C). The proteins have, respectively, a theoretical pI of 4.51 and 4.63, an expected MW of 40.4 and 44.9 kDa, and they both were predicted to have ten possible disulphide bonds.

A

***Lps*GH134a**
MKPIIKVVFMFVIYILSSQSFAQSQCNWYGTNYPVCQNQSTEWGWENNQSCIGPETCAANGGSSSGGSSSSSSSSSSSSTGGSGGAAGGQCDWYGSNYPLCTNQNSGWGWENNQSCIGPDTCTNPSGSSGGGSGNSSSSSSSSSSSSSSTGGSASGGTCNWYGTDHPMCVNTASGWGWENNQSCISQAECGTQGGSGGSSGGGNTGGGSNNGGSAGGSCGVGTCPDSLSCPGGASCGCYTVSGLGANKRAYQDAGADRRFLASAMMETELMDTNYTYGDGKSGDAFNAGATKQNWGMIRQCHSAWSGYGSGDYAVSDQMNSSRSLDVQVYGECRSYFGNNWFAGHRNGSSGLENPNTSDIQRFKEGYDWTYNMLNGHECDDVRFWVDIPAI-
***Lps*GH134b**
MGANLSCWLFRASTFLGGNTRYSHIHNRSFFMKPIIKVVFMFVIYLLSSQSFAQSQCDWYGTNYPVCQNQSSEWGWENNQSCIGPETCAANGGSSSGGSGGGSSSSSSSSSSSSSSSTGGSGGAAGGQCDWYGSNYPLCSNQNSGWGWENNQSCIGPDTCTSPSGSSGSGSGNSSSSSSSSSSSSSSSTGGSANGGTCNWYGTNHPMCVNTNSGWGYENNQSCISQAECDTQGGSGSSSSGGNTGGGSNNGGGSTGGGSCGVGTCPDSLNCPGGASCGCYTVSGLGANKRAYQDAGADRRFLASAMMETELMDTNYTYGDGKSGDAFNAGATKQNWGMIRQCHSAWSGYGADDYSVSDEMNSSRSLDVQVYGECRSYFGNNWFAGHRNGSSGLDNPNTDDIQRFREGYDWTYNMLNGHECDDVRFWVDIPAI-

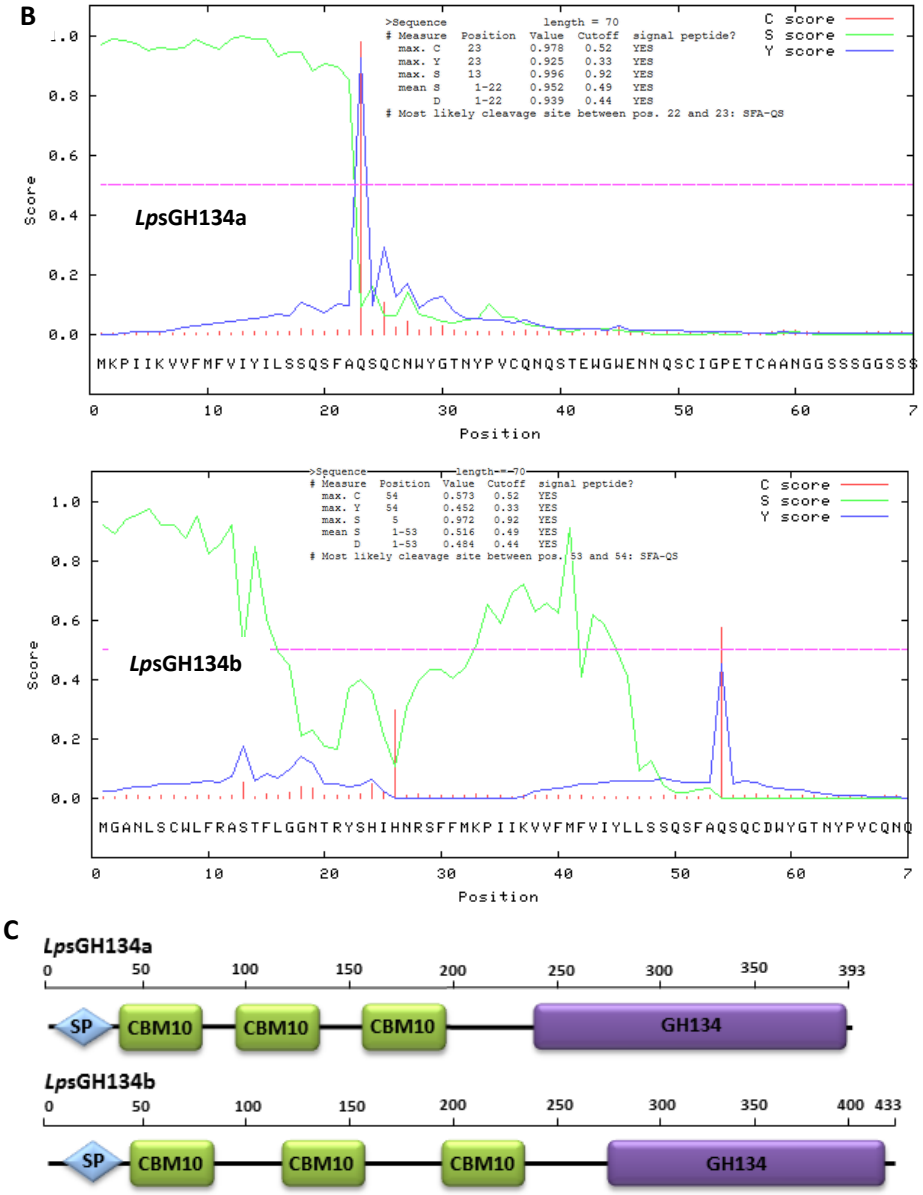


Figure 5.10. *Lps*GH134a and *Lps*GH134b predicted sequences and domains. **A.** The amino acid sequences obtained by translating the nucleotide sequences with the ExPASy translate tool. Red indicates the signal peptide and the symbol – the stop codon. **B.** Prediction of signal peptide cleavage site presence and location obtained with the SignalP 3.1 Server. **C.** The domain structure obtained with dbCAN. SP=signal peptide.

As for *Lps*GH5_8, the sequence for *Lps*GH134a and *Lps*GH134b (with their three CBMs) were amplified from *L. pedicellatus* gill's cDNA and purified (Fig. 5.11A1-A2). The primers used for the PCR, the thermo-cycling conditions and the detailed methods for the cloning of the protein are described in section 2.5 Ligation cloning into the cloning vector pSC-B-amp/kan was followed by white-blue screening of the transformed cells and confirmation by colony PCR of the presence of the inserted gene for *Lps*GH134a or *Lps*GH134b (Fig. 5.11B1-B2). The plasmids were extracted from the cells by miniprep and the presence of the correct sequences was verified by Sanger sequencing; the validated amino acid sequences were deposited in the ENA (accession number PRJEB28739).

Cloning into the expression vector (pET52b(+)) and transformation into the competent cells *E. coli* Rosetta-Gami 2(DE3) followed the same procedure as for LpsGH5_8 (Chapter 5.4.1). The screened colonies are shown in Fig. 5.11C1-C2.

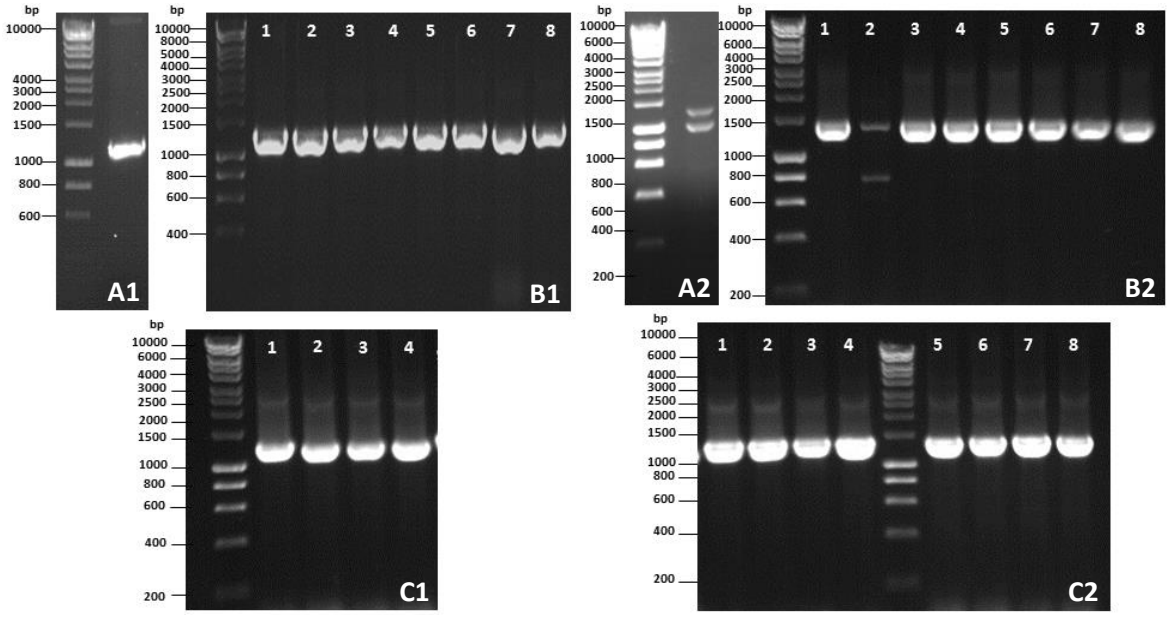


Figure 5.11. Gel electrophoresis of the cloning steps of *Lps*GH134a and *Lps*GH134b. **A**. Initial amplification of *Lps*GH134a (A1) and *Lps*GH134b (A2) from *L. pedicellatus* gill's cDNA. The PCR products lie at the right height for the predicted gene sizes of 1,179 and 1,302 base pairs respectively, though a double band is seen in A2. **B.** Colony PCR performed to confirm the presence of the *Lps*GH134a (B1) and *Lps*GH134b (B2) inserts in the cells transformed with the cloning vector pSC-B-amp/kan. All the eight screened colonies contained an insert of the right length for *Lps*GH134a, while seven of the eight colonies for *Lps*GH134b. **C.** Second colony PCR performed to confirm the presence of the *Lps*GH134a (C1) and *Lps*GH134b (C2) inserts in the cells transformed with the expression vector pET52b(+). All the four screened colonies contained an insert of the correct size for *Lps*GH134a, and all eight for *Lps*GH134b. The ladder used for the three gels was the HyperLadder 1kb from Bioline.

5.5.2 Expression

The Rosetta-Gami 2(DE3) cells from colony number one of the gel presented in Fig. 5.11C1 and colony number two of the gel of Fig. 5.10C2 were grown in 100 ml of LB and auto-induction (AI) medium, induced (only LB) with 1 mM IPTG, and were harvested after incubation at 16 °C and 20 °C for 20 hours (details of the expression and purification steps are in section 2.6). After cell lysis, the non-induced and induced soluble fractions were electrophoresed into a SDS-PAGE gel to confirm the presence of *Lps*GH134a and *Lps*GH134b and compare expression levels at the different temperatures with the two media. Figure 5.12A-B shows that expression of recombinant soluble *Lps*GH134a was successful only in LB at both temperatures, while for *Lps*GH134b at 20°C with both media and very little at 16 °C with LB.

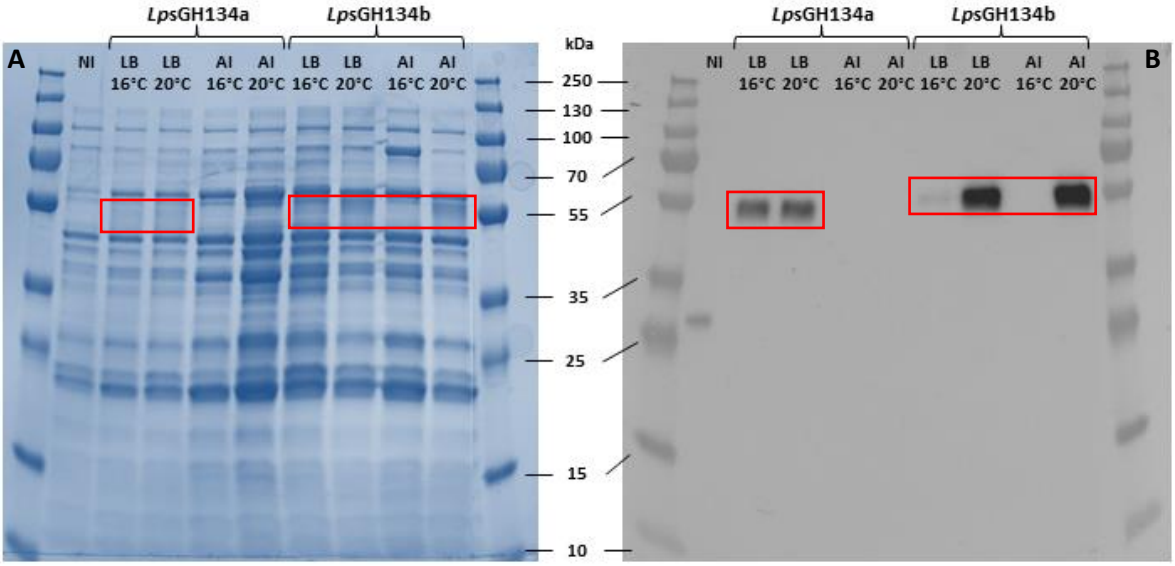


Figure 5.12. Analysis of the expression trial of *Lps*GH134a and *Lps*GH134b. **A.** SDS-PAGE of the soluble fraction obtained from 100 ml culture of non-inducted (NI) and induced cells at 16 °C and 20 °C in LB and AI. **B.** Western blot anti-Strep analysis of the same samples as in A. The expected MW of *Lps*GH134a is ~40.4 kDa, and of *Lps*GH134b is ~44.9 kDa, which is slightly smaller than the product seen in both in A and B (red rectangles), but still in line with correct expression. The loaded samples were 15 µl in volume. The protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific.

Expression was scaled up to one litre of LB medium, following the same protocol and incubating only at 20 °C after induction. The resulting soluble cell fraction was filtered and was purified by affinity chromatography by running through a 5 ml StrepTrap column and eluting with 2.5 mM desthiobiotin in PBS. Figure 5.13A-B shows the purification chromatogram and the SDS-PAGE analysis of the eluted fractions containing absorbance peaks.

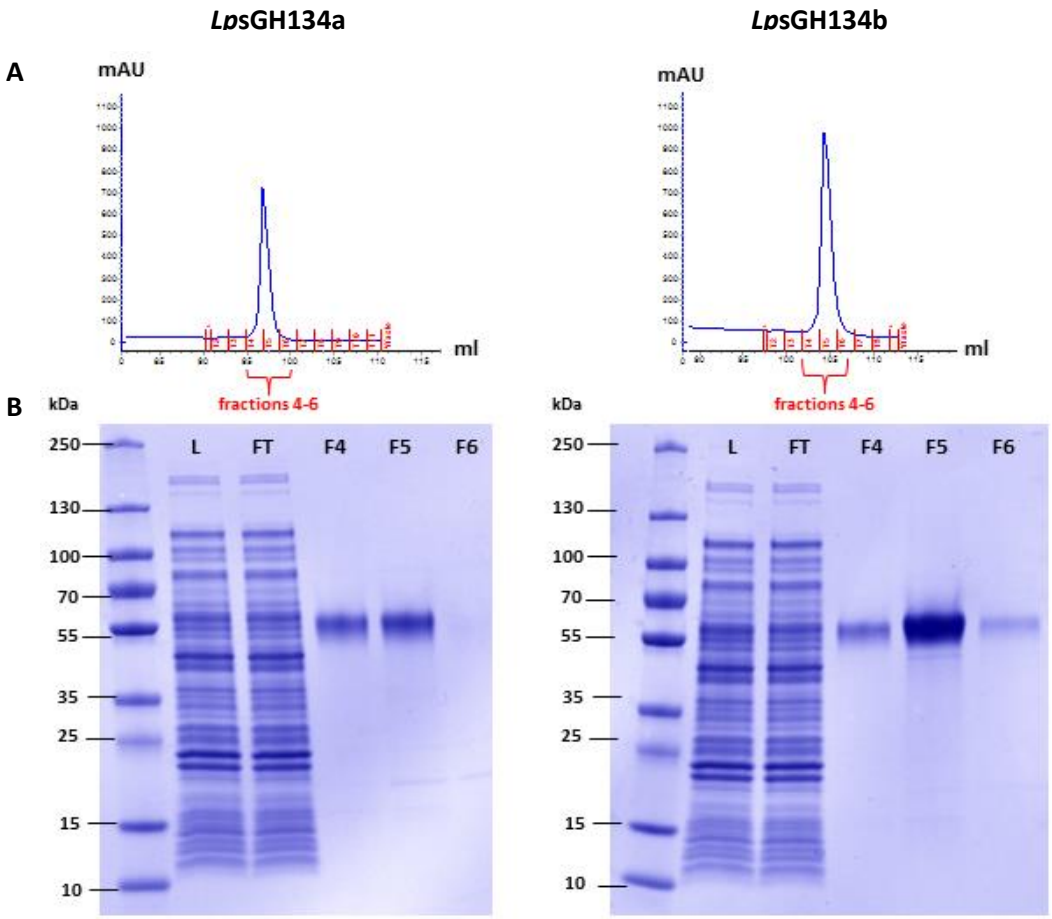


Figure 5.13. Affinity purification of *Lp*sGH134a (left) and *Lp*sGH134b (right). **A.** Chromatogram of the purification performed with a 5 ml StrepTrap column, showing only the elution phase. **B.** SDS-PAGE analysis of the purification presented in A. L=lysate, FT=flow through, F4 to F6 are the fractions from A. The volume of the loaded samples was 20 µl and the protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific. The expected MW of *Lp*sGH134a is ~40.4 kDa, and of *Lp*sGH134b is ~44.9 kDa, which is slightly smaller than the product seen in both in A and B (red rectangles), but is comparable the expression trials presented in Figure 5.12.

Fractions 4 and 5 contained most of the purified protein for both *Lp*sGH134a and *Lp*sGH134b, therefore they were combined together, subject to buffer exchange in PBS to remove the desthiobiotin and concentrated to 2 ml at 0.5 mg/ml (verified with a Bradford assay), for a total of 1 mg of protein. A thermal shift assay (Fig. 5.14) was performed to confirm the correct folding of the protein, which provided a T_m of 51.6 °C (R^2=0.954) for *Lp*sGH134a and of 50.6 °C (R^2=0.931) for *Lp*sGH134a. The melt curves did not present the typical shape that suggest correct folding, given that their fluorescence started at a high value and that they presented small or multiple peaks; this however could be the effect of the presence of multiple domains in the proteins sequence or the proteins might not be suitable for thermal shift analysis.

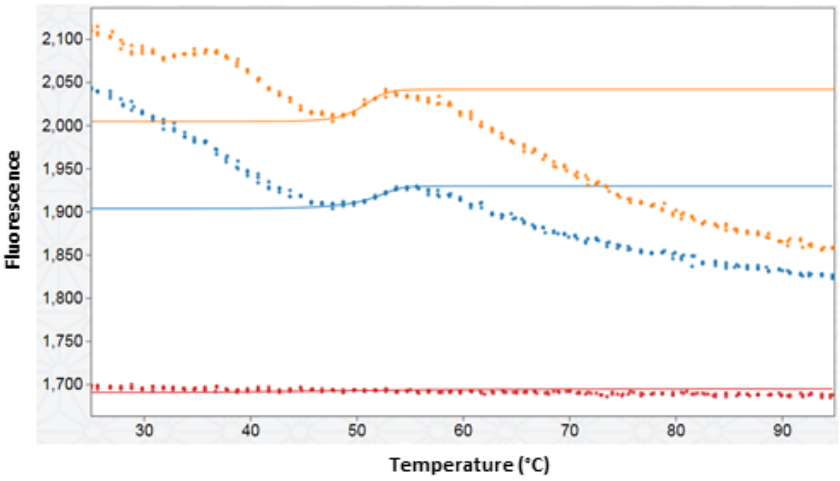


Figure 5.14. Thermal shift assay for *Lps*GH134a (blue line) and *Lps*GH134b (yellow line) after the fractions of the affinity purification were combined together, buffer exchanged and concentrated. Buffer only was used as a negative control (red line).

5.5.3 Activity assays

5.5.3.1 Activity against polysaccharides

The CAZy database lists GH134s as recently discovered enzymes (Shimizu *et al.*, 2015), for which endo-β-1,4-mannanase activity only has been recorded. Endo-β-1,4-mannanases are enzymes that are able to hydrolyse the internal β-1,4-linkage of the backbone of β-mannans such as glucomannan, galactomannan and galactoglucomannan (Moreira and Filho, 2008), which are hemicelluloses found normally in plant cell walls and endosperm and are used as storage polysaccharides (Shimizu *et al.*, 2015). Endo-β-1,4-mannanases are found in many organisms (eukaryotes, prokaryotes, fungi and viruses) and they belong to different GH families of the CAZy database, such as family 2, 5, 26 and 113 (Chauhan *et al.*, 2012). Out of the 138 known GH134s, five are bacterial, while all the others are produced by fungi belonging to various classes. GH134s have a unique inverted stereochemistry and a tertiary structure very similar to that of a lysozyme from hen egg white, suggesting that this structure has been co-opted during evolution for a completely different function (Jin *et al.*, 2016).

To test whether the GH134s from *L. pedicellatus* symbionts are indeed able to degrade β-mannans, DNS reducing sugar assays with *Lps*GH134a and *Lps*GH134b were performed on a range of different polysaccharides, which included mannans but also other hemicelluloses and cellulose in different preparations. The assays were performed despite the unusual melt curves for the proteins obtained with the thermofluor (Fig. 5.14), since Congo red agar plate assays with mannan substrates gave positive results (not shown),

suggesting that the proteins are folded and active. Figure 5.15 presents the results of the DNS reducing sugar assay to test substrate preference, which were performed by incubating the protein for two hours in sodium phosphate buffer pH 8.0 at 30 °C and 320 rpm.

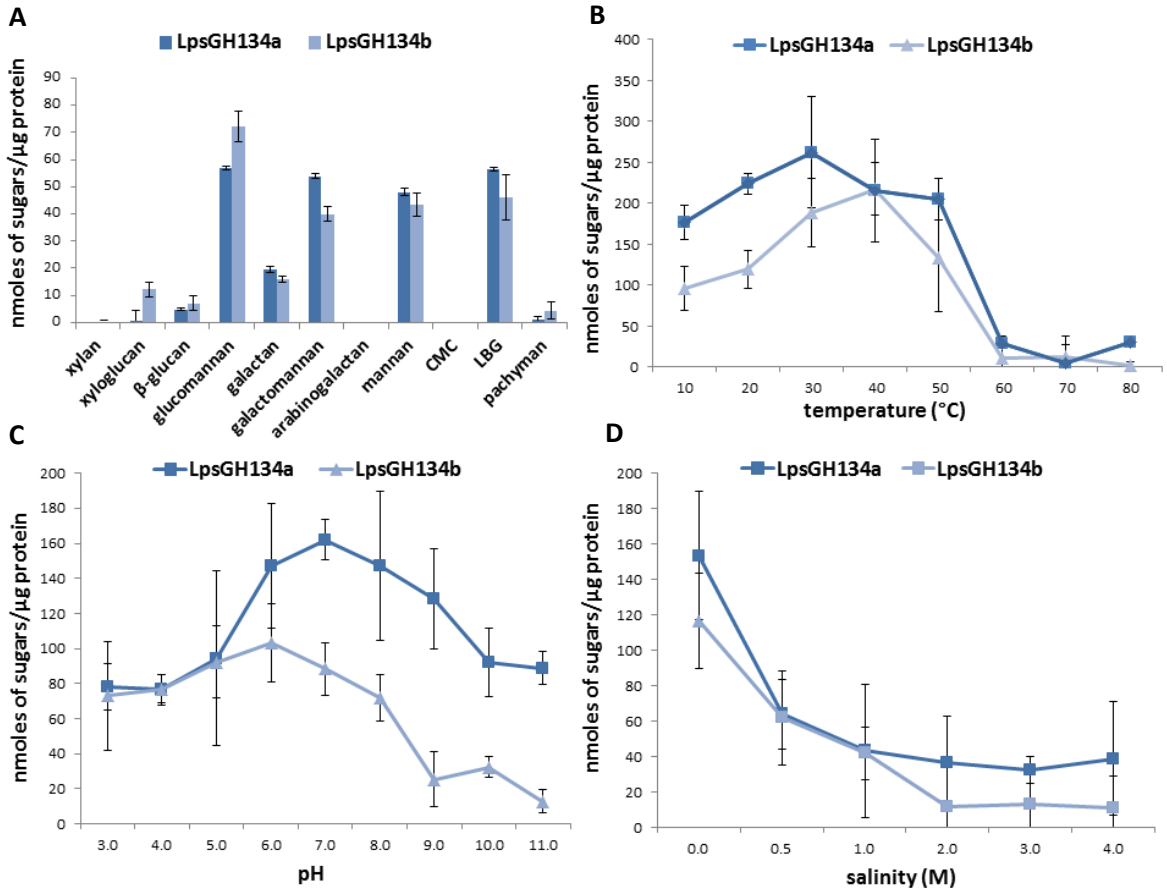


Figure 5.15. DNS reducing sugar assays for *Lps*GH134a and *Lps*GH134b. The protein was tested on a range of polysaccharides. CMC= carboxymethyl cellulose, LBC=locust bean gum, PASC= phosphoric acid swollen cellulose. **B**, **C** and **D** show the protein activity at different temperatures, pH and salinities, respectively.

The assay shows that both *Lps*GH134a and *Lps*GH134b are able to degrade glucomannan, galactomannan, mannan and LBG, with minor activity on galactan and β-glucan, while the amount of reducing sugars produced on pachyman is not sufficient to confirm activity on this substrate. The two proteins seem to have very similar activity, with minor differences apart from the products obtained from the digestion with xyloglucan, which are more convincing for *Lps*GH134b but need further analysis for confirmation. Both proteins do not seem to be able to degrade xylan or arabinogalactan or cellulosic substrates (CMC and PASC), and both seem to have a similar preference for pH around neutrality, despite *Lps*GH134b showing less activity pH above 8.0. Both proteins also seem to be intolerant to

salt, given that best performance was only obtained for buffer without NaCl added. The temperature at which the highest amount of reducing sugars was produced was 30 °C for *Lps*GH134a and 40 °C for *Lps*GH134b; both proteins have lower activity in the 20-30 °C range and a sudden loss of activity between 50 and 60 °C, which is in line with the recorded T_m for the two enzymes (51.6 °C and 50.6 °C, respectively).

5.5.3.2 Heat maps

In order to identify the enzymes optimal activity range at pH and temperature values combined together, heat maps for *Lps*GH5_8 were produced, with the same experimental procedure described for *Lps*GH134a and *Lps*GH134b in section 5.4.2.2. The results are shown in Figure 5.16.

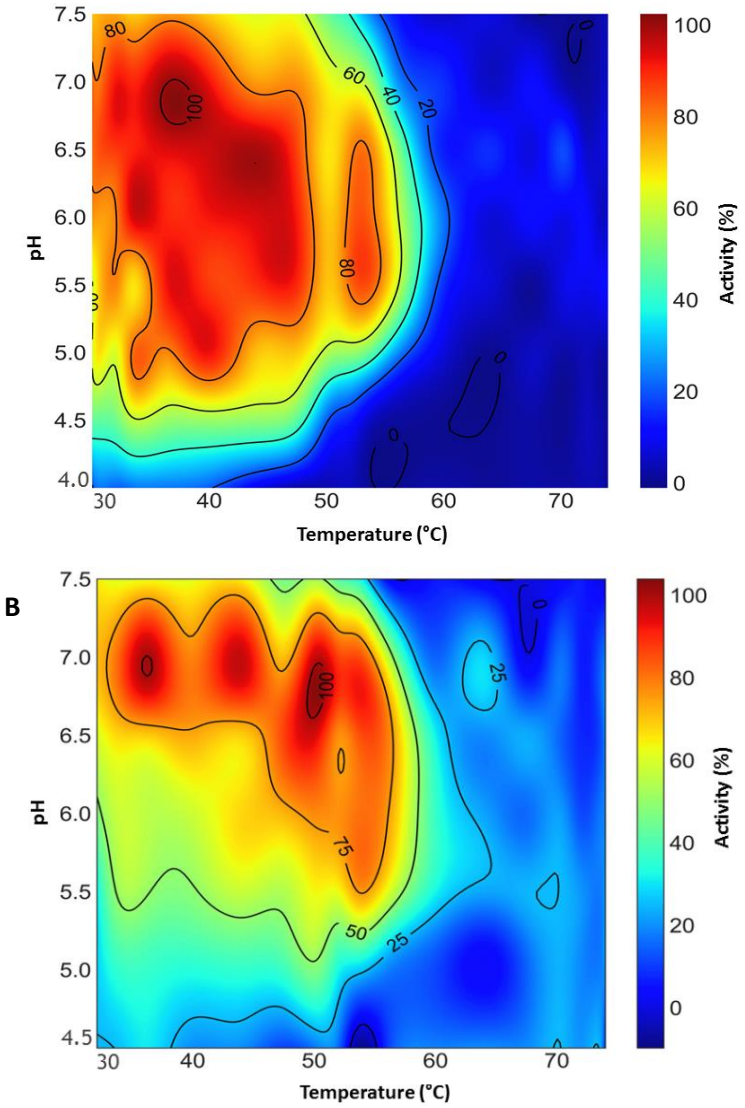


Figure 5.16. Heat maps for *Lps*GH134a (A) and *Lps*GH134b (B). The substrate was glucomannan (1%), the temperature range 30 to 74 °C, the pH range 4.5 to 8.0 for *Lps*GH134b and 4.0 to 7.5 for *Lps*GH134a (citrate buffer). 10 µl of protein (at 0.5 mg/ml) were leaded in each well.

*Lps*GH134a showed best activity for a wide range of pH (from 4.5 to 7.5) and for temperatures between 30 to 50 °C; a hot spot was identified at the pH of around 6.8 and temperature of 38 °C. *Lps*GH134a does not show the same pH tolerance and seems to prefer higher temperatures; a main hot spot was identified, at around pH 6.7 and temperature of 51 °C. For both proteins, the results obtained with the heat maps are similar to those obtained with DNS sugar assays performed with separate measurements for the two parameters.

5.5.3.3 Polysaccharide analysis using carbohydrate gel electrophoresis

The DNS reducing sugar assays confirmed that *Lps*GH134a and *Lps*GH134b are β-1,4-mannanases active on a range of different β-mannans. To deepen our investigation, we performed PACE analysis in order to gather further information on the mechanism of enzyme action and to find out whether the two enzymes are endo- or exo-acting β-mannanases. Similarly to the experiments with *Lps*GH5_8 (section 5.4.3.3), we mixed *Lps*GH134a and *Lps*GH134b were mixed with a panel of mannose substrates, including scots pine, and the reactions were run at 30 °C overnight before removing the undigested polysaccharides and drying the products, which were than separated in a polyacrylamide gel (Figure 5.17). These assays were performed by Marta Busse-Wicher at the Department of Biochemistry, Cambridge University.

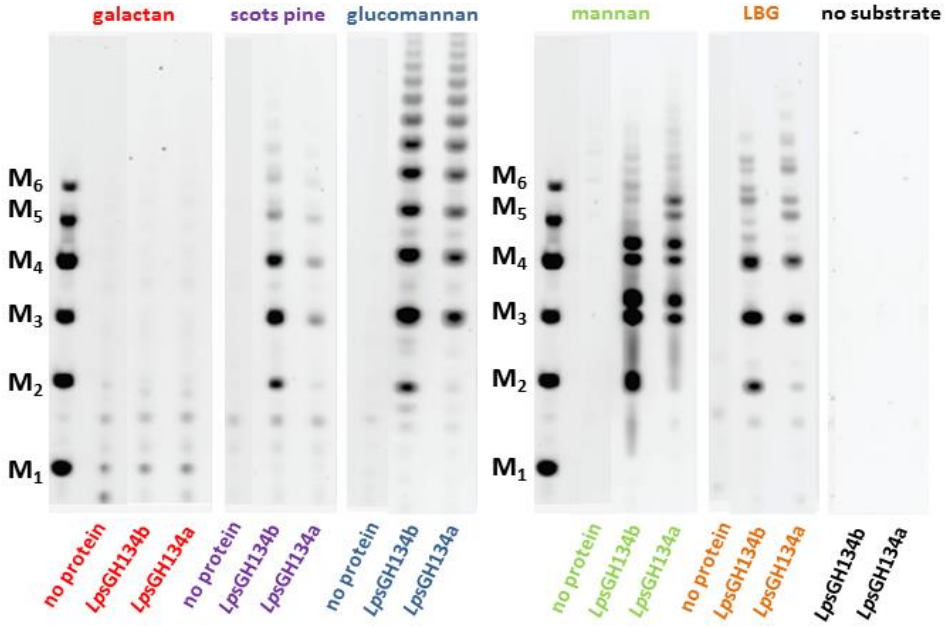


Figure 5.17. PACE analysis performed on selected substrates for *Lps*GH134a and *Lps*GH134b. 1 µg of protein was used for each reaction. M1 to M6 represent the mannose standards containing from one to six molecules of mannose. Negative controls are those with no protein (right panel) or no substrates (first lane of each panel). LBG=locust bean gum.

The assay confirms the results obtained with the DNS reducing sugars assay (Fig. 5.15), indicating that both *Lp*sGH134a and *Lp*sGH134b are able to degrade polysaccharides such as glucomannan, mannan, LBG and galactan. The digestion of scots pine also produced mannose products similar (but not identical) to those obtained with LBG. *Lp*sGH134a and *Lp*sGH134b do not generate identical digestion products, since the former appears to release longer mannoolisaccharides that the latter, with only *Lp*sGH134b generating mannobiose. Both proteins do not appear to have mannosidase activity since no mannose is generated, therefore they can be classified as endo-β-1,4-mannanase.

5.6 The bacterial *Lp*sGH11

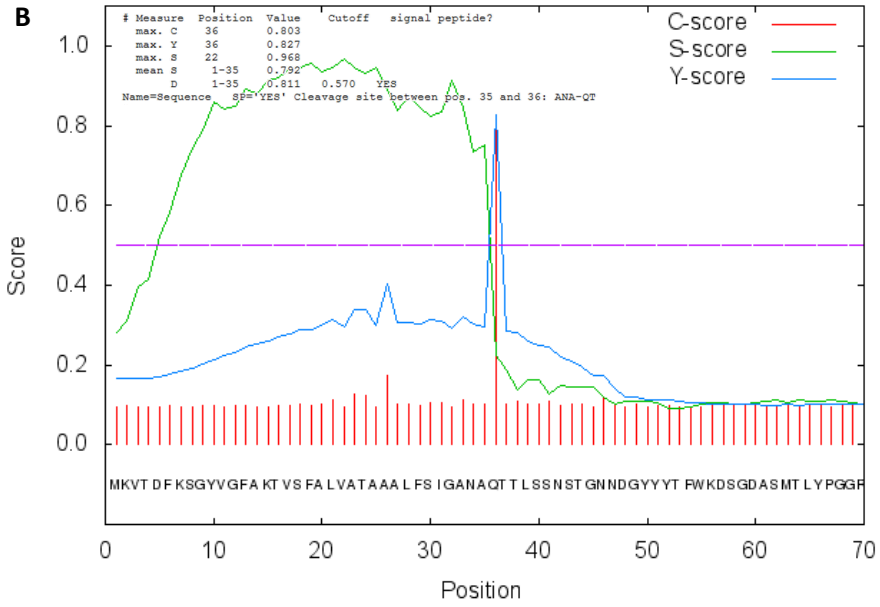
5.6.1 Sequence analysis and cloning

The nucleotide sequence encoding for the bacterial protein *Lp*sGH11 was obtained from the contig c169869_g1_i2 of *L. pedicellatus* transcriptome presented in chapter 4.3. The protein is found among the CAZymes of the caecum and digestive glands, despite not being abundant. It was set as a target for heterologous expression because it is among the most highly expressed bacterial CAZymes of the gills and because it was identified by SDS-polyacrylamide gel electrophoresis performed with the caecum fluids (chapter 3.4.2.4). Table 5.1 and 5.2 show *Lp*sGH11 annotation details and expression attempts. The main protein parameters were predicted using the tools described in materials and methods (section 2.4.3), and are shown in Figure 5.18. *Lp*sGH11 comprises of 419 residues (Fig. 5.18A) and it displays a signal peptide for secretion at the N-terminus, which includes the first 35 amino acids (Fig. 5.18B). The signal peptide is followed by a GH11 domain, a CBM10 and a CBM5 (which was previously described as a cellulose binding domain, while chitin binding activity has recently been reported in several cases), with the three modules being separated by two poly-serine linkers (Fig. 5.18C). The protein has a theoretical pI of 4.68 and an expected MW of 43.3 kDa, and it is predicted to have four possible disulphide bonds.

A

*Lps*GH11
MKVTDFKSGYVGFAKTVSFALVATAAALFSIGANAQTTLSSNSTGNNDGYYYTFWKDSGDASMTLYPGGRYTSSWTS STNNWVGGYGWNPGSRRTVSYSGSYSANGTSYLALYGWTRNPLIEYYVVENWVNYNPSSGADTYGTVNIDGSTYTL ARSQRVQQPSIDGTATFYQYWSVRSNKRTSGTIDVGAHFDAWASVGLNLGTEFNYMVMATEGYQSGGSSDITVSQG GSSGSTTSSSSSSSSSSSSSSSSSGSTGGQCGNCNWYGSIYPLCNNQNSGWGWESNQSCIGVDTCSGQNGNGGPVDC NGGGSNSSSSSSSSSSSSSSSSSNSSSSSSSSSSSSSSSSSSSSGGNNGGGTNCGGVNEYPNWTAKDWSGGTNNHANS GDQMVYQGQLYQANWYTSTVPGSDASWTTVGSCN-

B



C *Lps*GH11

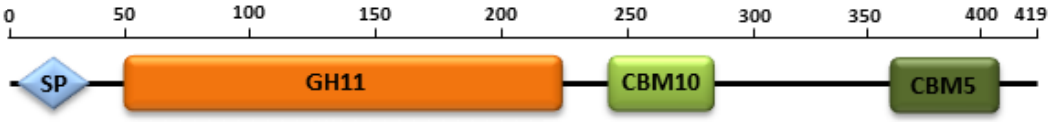


Figure 5.18 *Lps*GH11 predicted sequences and domains. **A.** The amino acid sequences obtained by translating the nucleotide sequences with the ExPASy translate tool. Red indicates the signal peptide and the symbol – the stop codon. **B.** Prediction of signal peptide cleavage site presence and location obtained with the SignalP 4.1 Server. **C.** The domain structure obtained with dbCAN. SP=signal peptide.

Similarly to the other expressed bacterial proteins, the sequence for *Lps*GH11 (with its two CBMs) was amplified from *L. pedicellatus* gill's cDNA and purified (Fig. 5.17A). PCR, ligation cloning into the cloning vector pSC-B-amp/kan, transformed cells screening and colony PCR was also performed following the same methods (Fig. 5.18B). The correct sequence was verified by Sanger sequencing and the validated amino acid sequence was deposited in the ENA (accession number PRJEB28739).

The expression strategy for *Lps*GH11 was different from the other bacterial proteins. The initial attempts to express the protein in *E. coli* cytoplasm and periplasm failed (not shown) despite trials with different temperatures, induction concentrations, culture media, competent cells and solubility tags. It was therefore decided to attempt expression in insect

cells with the help of the Oxford Protein Production Facility UK (OPPF-UK), which kindly allowed us to use their high throughput expression and purification systems for recombinant proteins (https://www.oppf.rc-harwell.ac.uk/OPPF/). The vector used for expression was pOPINSC3, developed by OPPF-UK (Bird, 2011), which includes an N-terminal His-tag, followed by a SUMO (small ubiquitin like modifier) solubility tag and by the rhinovirus 3C cleavage site just before the protein cloning site. The *Lps*GH11 sequence was therefore amplified with primers carrying a 15 bp extensions complementary to the ends the pOPINSC3 vector and the vector was linearised by reverse PCR. In-Fusion cloning was used to clone the gene into the vector, and the obtained construct was transformed into OmniMaxII competent cells (Fig. 5.19C) and the correct sequence was verified by Sanger sequencing.

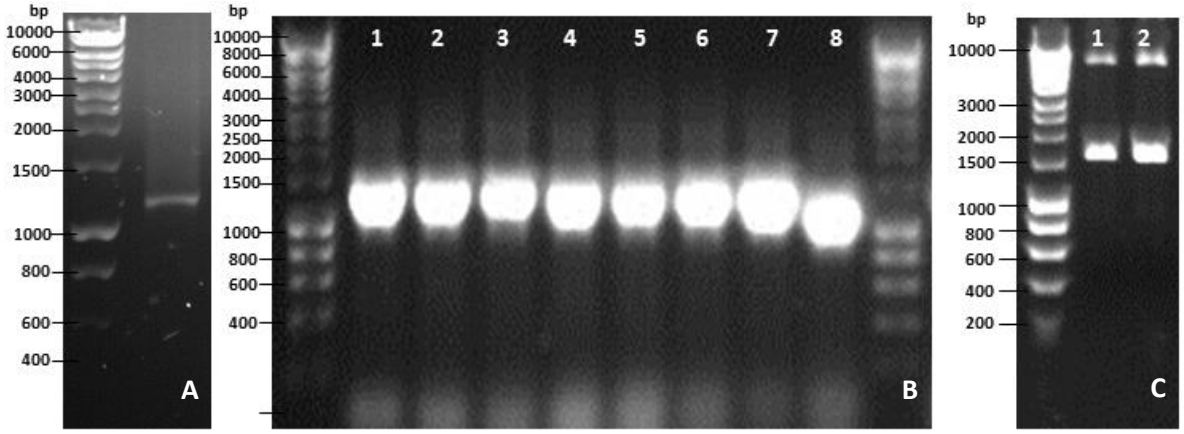


Figure 5.19. Gel electrophoresis of the cloning steps of *Lps*GH11. **A**. Initial amplification from *L. pedicellatus* gill's cDNA. The PCR product lies at the right height for the predicted gene sizes of 1,260 base pairs. **B.** Colony PCR performed to confirm the presence of the *Lps*GH11 insert in the cells transformed with the cloning vector pSC-B-amp/kan. Most of the screened colonies contained an insert of the right length. **C.** Second colony PCR performed to confirm the presence of the *Lps*GH11 insert in the cells transformed with the expression vector pOPINSC3. Only two colonies were screened and both contained *Lps*GH11 insert with the SUMO tag (for a total of 1,605 base pairs). The ladder used for the three gels was the HyperLadder 1kb from Bioline.

5.6.2 Expression

Infection and expression screening into Sf9 (*Spodoptera frugiperda*) insect cells was performed using the baculovirus expression system (see section 2.6.4). Figure 5.20 shows the initial expression screening and the results of the scale up into 0.5 litres of culture medium.

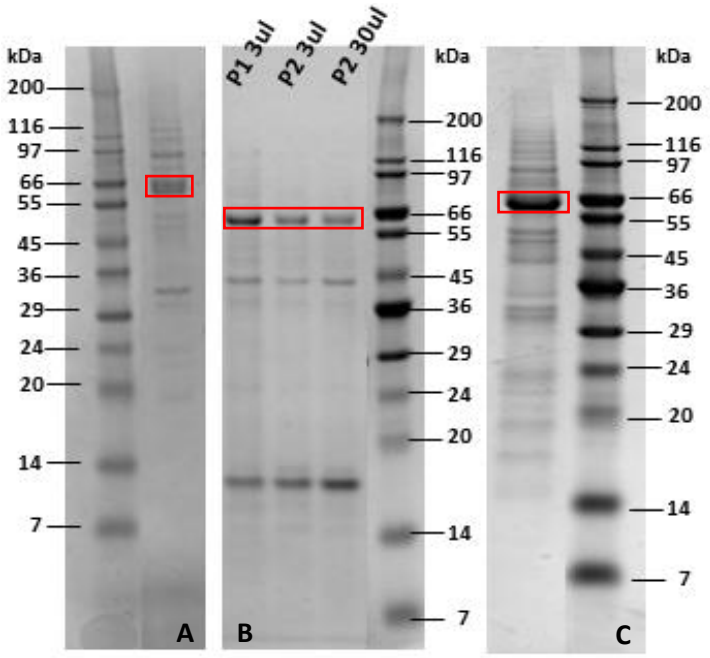


Figure 5.20. Gel electrophoresis of the transfection of *Lp*sGH11 into Sf9 cells. **A.** Initial screening after transfection with P1 virus, showing expression at the MW of around 60 kDa (red rectangle), which is corresponds closely to the protein MW (43.3 kDa) together with the SUMO tag MW (13.2 kDa). **B.** expression screening with the P2 viruses, showing expression both in 3 and 30 µl of medium. **C**. Expression scale up into 0.5 litres of medium. The ladder used for the three gels was the wide range SigmaMarker from Sigma-Aldrich.

The pellet obtained from the centrifugation of the scale-up was lysed and the supernatant was subject to affinity purification using a 5 ml HisTrap HP column and a 30-500 mM imidazole gradient. The fractions obtained from the purification were pooled together, concentrated to 2.5 ml and subjected to gel filtration. Figure 5.21 shows the chromatogram of the purification and the SDS-PAGE and western blot analysis of the fractions obtained. The obtained protein was of good purity, however two bands with a MW difference of around 30 kDa were identified as *Lp*sGH11 in the SDS-PAGE and the western blot. Protein identification by MALDI-TOF/TOF tandem mass spectrometry was performed on the two bands, and both of them were identified as *Lp*sGH11, but no differences in the amino acids composition could be highlighted, though it is possible that one band represents a shorter form of the protein caused by protease degradation or by the loss of one of the CBMs. All the fractions from peak two of the gel filtration were pooled together and the SUMO tag was successfully cleaved using the Turbo HRV 3C protease, procedure that was confirmed by SDS-PAGE and western blot analysis (not shown). The resulting *Lp*sGH11 was concentrated to 1 ml at 0.1 mg/ml (verified with a Bradford assay), for a total of 100 mg of

protein. A thermal shift assay (Fig. 5.22) was performed to confirm the correct folding of the protein, which resulted to have a T_m of 62.2 °C (R^2=0.999).

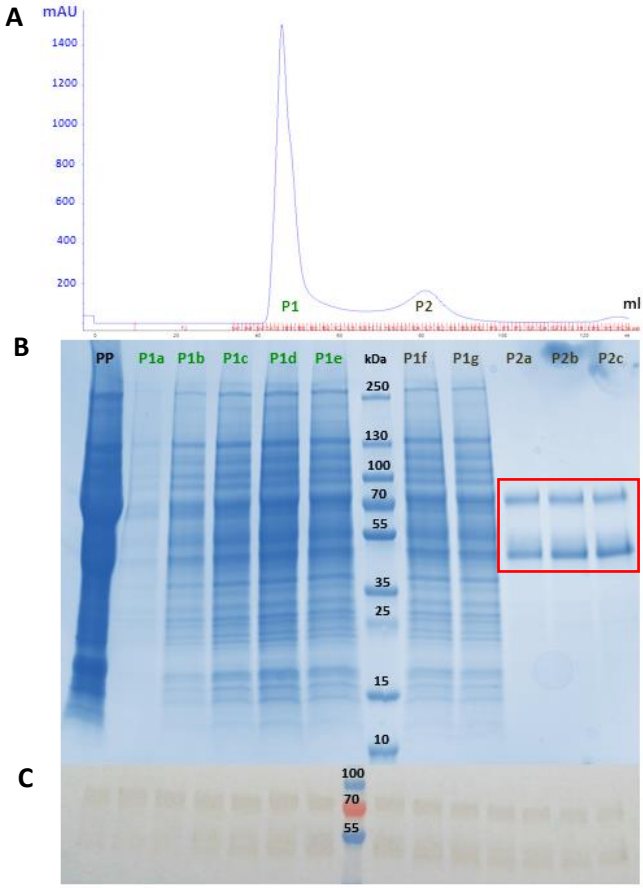


Figure 5.21. Gel filtration of *Lps*GH11. **A.** Chromatogram of the gel filtration performed with a HiLoad 16/600 Superdex 75 pg column in PBS. **B.** SDS-PAGE of the peaks from the purification presented in A. The volume of the loaded samples was 15 µl and the protein ladder used was PageRuler Plus prestained protein ladder from thermo Scientific. PP=pre-purification, P1a to P1g=various fractions from peak 1, P2a to P2c= fractions from peak 2. The expected MW of *Lps*GH11 with the SUMO tag is ~56.5 kDa, which is lower or higher than the two bands seen in the three fractions from peak 2 (red rectangle). **C.** Western blot anti-His analysis of all the fractions from peak 2, showing that the antibody recognised the protein in both bands.

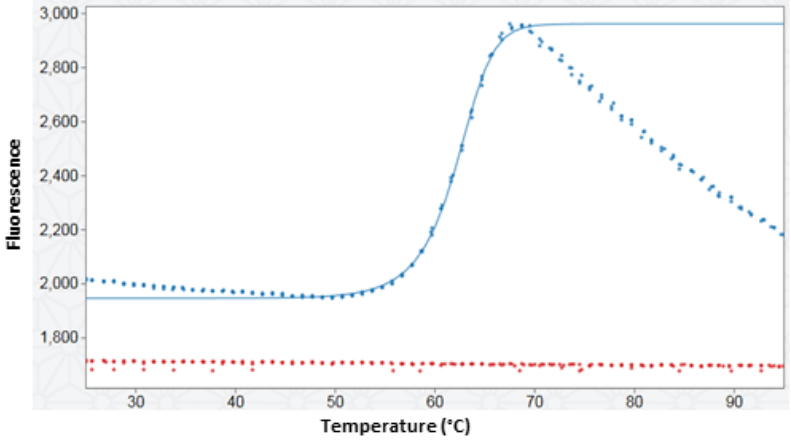


Figure 5.22. Thermal shift assay for *Lps*GH11 (blue line) after the SUMO tag cleavage and concentration. The red line is the negative controls (buffer only).

5.6.3 Activity assays

5.6.3.1 Activity against polysaccharides

Annotation against the NCBI nrdb classifies *Lps*GH11 as a putative 1,4-beta-xylanase, and the CAZy database lists GH11s as enzymes with endo-β-1,4-xylanase or endo-β-1,3-xylanase only. Xylanases are glycosidases that randomly hydrolyse 1-4- or 1-3-β-xylosidic linkages of the cell wall polysaccharide xylan, and are produced (mostly secreted) by many organisms, including bacteria, algae, protozoa, fungi and other eukaryotes (Prade, 1996; Collins *et al.*, 2005). Xylanases are found among the GH families 5, 8, 7 and 43, but mostly belong to the families GH10 and GH11, with the family 11 being the only one that is monospecific, containing only xylanases and being active solely on substrates containing D-xylose (Collins *et al.*, 2005).

To test whether *Lps*GH11 is indeed able to degrade xylans, DNS reducing sugar assays were performed on a range of different polysaccharides, which included xylans but also other hemicelluloses and cellulose in different preparations. Figure 5.23 presents the results of the DNS reducing sugar assay to test substrate preference, which were performed by incubating the protein for two hours in sodium phosphate buffer pH 8.0 at 30 °C and 320 rpm.

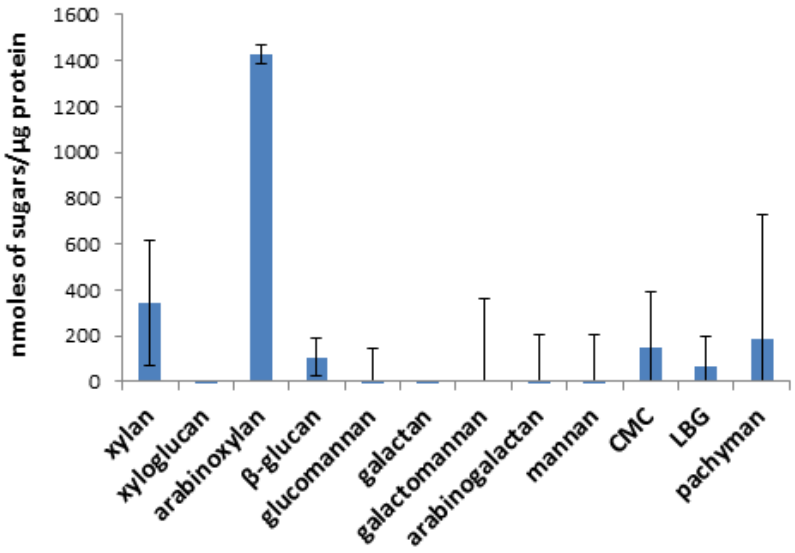


Figure 5.23. DNS reducing sugar assays for *Lps*GH11. The protein was tested on a range of polysaccharides. CMC= carboxymethyl cellulose, LBC=locust bean gum, PASC= phosphoric acid swollen cellulose.

The assay shows that *Lps*GH11 is definitely able to degrade arabinoxylan. The other substrates on which it shows activity, mainly xylan but also CMC, LBG and pachyman, present errors bars that are too big to be able to confirm activity with confidence. The assay

could not be repeated to improve the error bars due to the lack of protein, which was expressed on a difficult platform (insect cells) and only at low yields. The protein does not seem to be able to degrade mannans or galactan. It is therefore possible to confirm the annotation of 1,4-beta-xylanase, though β-1,3-xylanase activity is also possible and should be checked on an appropriate substrate.

5.6.3.2 Polysaccharide analysis using carbohydrate gel electrophoresis

Similarly to the other bacterial enzymes, *Lps*GH11 was tested with PACE in order to gather additional information on the mechanism of enzyme action and to find out if it has a preference for a specific cleavage site along the xylan chain. Different amounts (from 0.3 to 4.0 µg) of *Lps*GH11 were therefore mixed with grass xylan, and the reactions were run at 30 °C overnight before removing the undigested grass xylan and drying the products, which were than separated in a polyacrylamide gel (Figure 5.24). These assays were performed by Marta Busse-Wicher at the Department of Biochemistry, Cambridge University.

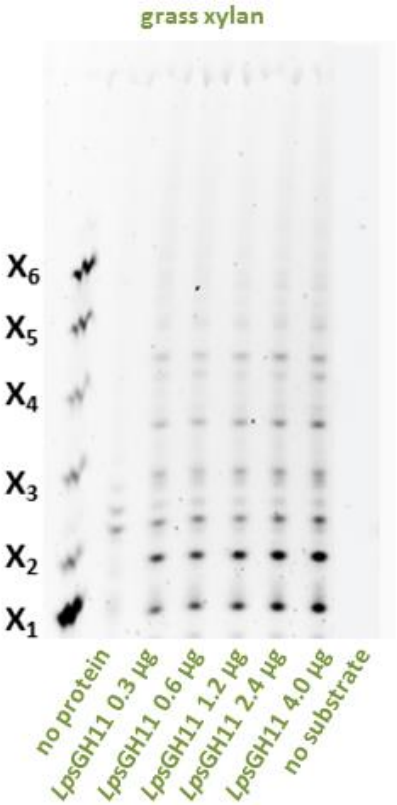


Figure 5.24. PACE analysis performed on grass xylan for *Lps*GH11. From 0.4 to 4.0 µg of protein were used for each reaction. X1 to X6 represent the xylose standards containing from one to six molecules of xylose. Negative controls are those with no protein (first lane) or no substrates (last lane).

The digestion products highlighted by the assay show that *Lp*sGH11 is able to digest grass xylan, producing mainly xylose and xylobiose, and some xylotriose, while it does not generate longer polymers of xylose. This suggests that the enzyme probably has exo- rather than endo-xylanase activity. Bands not corresponding to the xylose standards are also visible in the gel, suggesting that the *Lp*sGH11 is also able to cleave sugar decorations along the xylan chain; the nature of these sugars could be assessed by re-running the PACE assay using different standards.

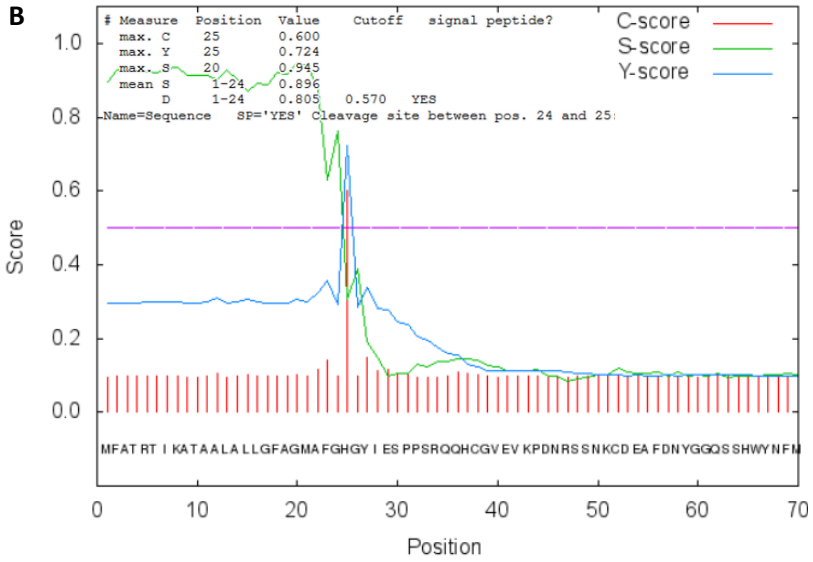
5.7 The bacterial *Lps*AA10

5.7.1 Sequence analysis and cloning

*Lp*sAA10 was set as a target for heterologous expression because it is the fourth most highly expressed bacterial CAZyme of *L. pedicellatus* gills and because it was identified by SDS-polyacrylamide gel electrophoresis performed with the caecum fluids (chapter 3.4.2.4), despite not being very abundant in the caecum proteomics results. It is also the only full length LPMO found in the caecum fluids, where it is likely to have a role in wood digestion. Its nucleotide sequence was obtained from the contig c173837_g2_i2 of the *L. pedicellatus* transcriptome presented in chapter 4.3 and table 5.1 and 5.2 show its annotation details and expression attempts. The main protein parameters were predicted using the tools described in materials and methods (section 2.4.3), and are shown in Figure 5.25. *Lp*sAA10 comprises of 379 amino acid residues (Fig. 5.25A) and it displays a signal peptide for secretion at the N-terminus, which includes the first 24 amino acids (Fig. 5.25B). The signal peptide is followed by an AA10 domain and a CBM10, which are separated by a poly-serine linker (Fig. 5.25C). The protein has a theoretical pI of 5.70 and an expected MW of 40.0 kDa, and it is predicted to have eight possible disulphide bonds.

A

*Lps*AA10
MFATRTIKATAALALLGFAGMAFGHGYIESPPSRQQHCGVEVKPDNRSSNKCDEAFDNYGGQSSHWYNFMSVVAH NEGRKVEKGTTNVCGFDGESFNPAPWDTAADWPTTAISAGNQNFVWNISNGPHFSDTKELVFWITKPGFTFDSNRV LTWDDFEPAPFCDEQVGSGDFGSNPNVTAGSVDITVNCNVPSRSGRHVIYSEWGRDPATYERFFSCVDVTFGGSNG GGSTTSSSSSSSSSSSSSSSSSSSNSSSSSSSSSSSSSSSGSTGGQCVCQWYSEGTYPICENSTGWGWENNQSCQVAS GGSSGGGSSSGSCGTGSCPSSLSCPSGMNCGCYTVFRFGSKQTRLSRCRGRSKLPRVSNDGNRSDGYQLRLW-



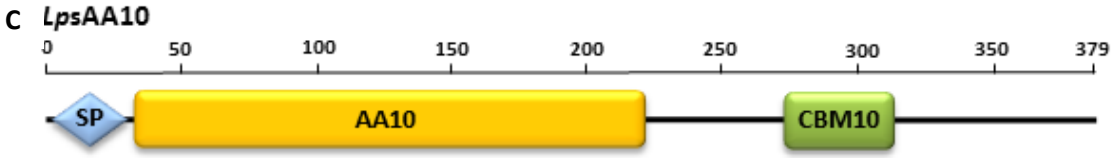


Figure 5.25. *Lps*AA10 predicted sequences and domains. **A.** The amino acid sequences obtained by translating the nucleotide sequences with the ExPASy translate tool. Red indicates the signal peptide and the symbol – the stop codon. **B.** Prediction of signal peptide cleavage site presence and location obtained with the SignalP 4.1 Server. **C.** The domain structure obtained with dbCAN. SP=signal peptide.

The sequence for *Lps*AA10 could not be amplified from *L. pedicellatus* gill's cDNA despite numerous attempts, therefore it was synthetized and codon optimised for expression in *E. coli* by GenArt (www.thermofisher.com). The sequence, without the CBM, was cloned into the vector pETFPP30 (a modification of the vector pET22b) in order to perform expression in the periplasm, an expression system that has more chances to produce a correctly folded and active protein. The vector pETFPP30 has an N-terminal pelB leader sequence which guides to protein to the periplasm (Yoon *et al.*, 2010) and a C-terminal His-Tag preceded by the human rhinovirus (HRV) 3C cleavage site. PCR was used to amplify *Lps*AA10 sequence with primers carrying a 15 base pairs extensions complementary to the ends the vector pETFPP30, adding a Strep-tag at the N-terminus (Fig. 5.26A), and to linearise the vector (see materials and methods section 2.5). The obtained construct was transformed in stellar competent cells and screened by colony PCR (Fig. 5.26B), the correct sequence was verified

by Sanger sequencing and the validated amino acid sequence was deposited in the ENA (accession number PRJEB28739). The plasmids were then transformed into *E. coli* Rosetta 2(DE3) competent cells for expression, with colony PCR screening to confirm the presence of the constructs containing the *Lps*AA10 insert (Fig. 5.26C);

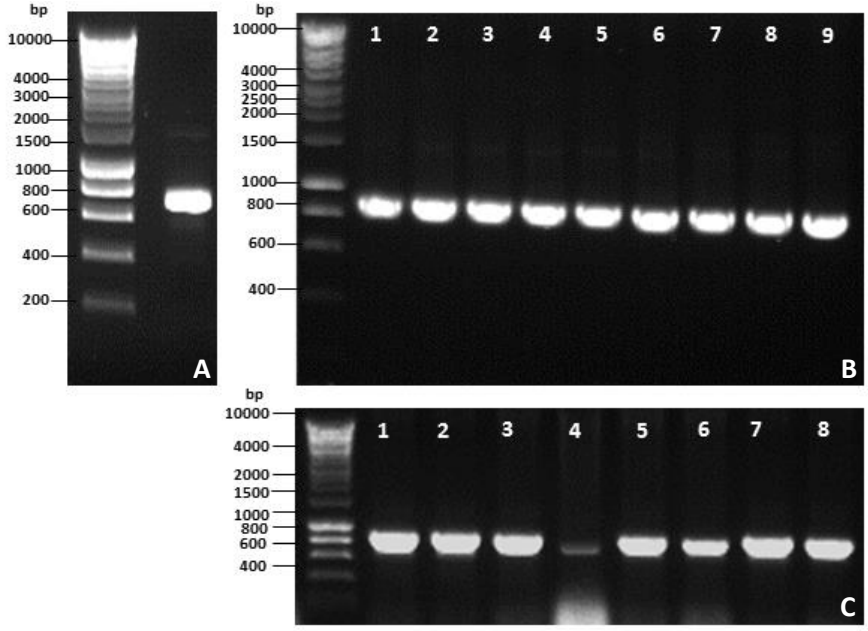


Figure 5.26. Gel electrophoresis of the cloning steps of *Lps*AA10. **A**. Initial amplification from the vector containing the codon optimised sequence. The PCR product represents *Lps*AA10 without the CBM, which has a predicted size of 696 base pairs. **B.** Colony PCR performed to confirm the presence of the *Lps*AA10 insert in the cells transformed with the vector pETFPP30. All the screened colonies contained an insert of the right length. **C.** Second colony PCR performed to confirm the presence of the *Lps*AA10 insert in the Rosetta-Gami 2(DE3) competent cells transformed with the expression vector pETFPP30. All the colonies screened contained *Lps*AA10. The ladder used for the three gels was the HyperLadder 1kb from Bioline.

5.7.2 Expression

The Rosetta 2(DE3) cells from colony one of Fig. 5.25C were grown in 500 ml of M9 minimal medium, induced with 0.1 mM IPTG and harvested after incubation at 20 °C for 20 hours. The periplasmic fraction was recovered and the supernatant was affinity purified with a 5 ml StrepTrap HP column and eluted with 2.5 mM desthiobiotin (Fig. 5.26A-B. Details of the expression, harvesting and purification steps are in section 2.6). The resulting fractions containing the eluted proteins were pooled together, excess $CuSO_4$ was added to provide the metal ion for the protein active centre, and the unbound copper and desthiobiotin were removed by performing gel filtration (Fig. 5.27C-D).

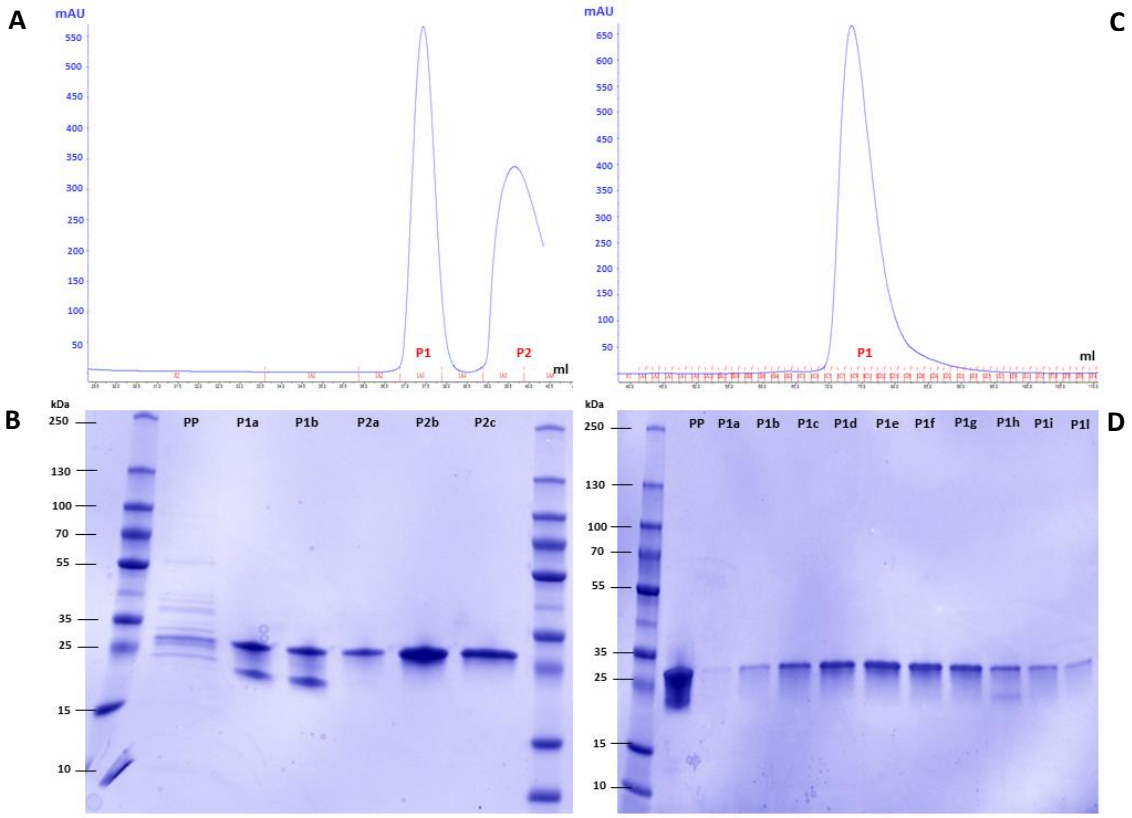


Figure 5.27. Purification of *Lps*AA10. **A.** Chromatogram of the purification performed with a 5 ml StrepTrap HP column. The program stopped before the end of the peak P2 and the rest of the elutions were collected by hand. **B.** SDS-PAGE analysis of the purification presented in A. PP=pre purification, P1 (a to b) and P2 (a to c) are the fractions from A. The volume of the loaded samples was 20 µl. **C.** Chromatogram of the gel purification performed with a HiLoad 16/600 Superdex 75 pg column in PBS. **D.** SDS-PAGE analysis of the purification presented in C. PP=pre purification, P1 (a to l) are the fractions from C. The volume of the loaded samples was 10 µl. The expected MW of *Lps*AA10 without CBM is ~25.2 kDa, which is in line with the band see in the two gels. The protein ladder used for both gels was PageRuler Plus prestained protein ladder from thermo Scientific.

The obtained fractions were combined together and concentrated to 2 ml at 0.7 mg/ml (verified with a Bradford assay), for a total of 1.4 mg of protein. A thermal shift assay was performed to confirm the correct folding of the protein and the binding of the copper (Fig. 5.28). The T_m before adding of copper was of 45.9 °C (R^2=0.988) and of 49.1 °C, which increased to 49.1 °C (R^2 = 0.997) upon addition of $Cu2^+$ and was retained after gel filtration (50.1 °C, R^2 = 0.997), indicating tight specific binding and protein-fold rearrangement.

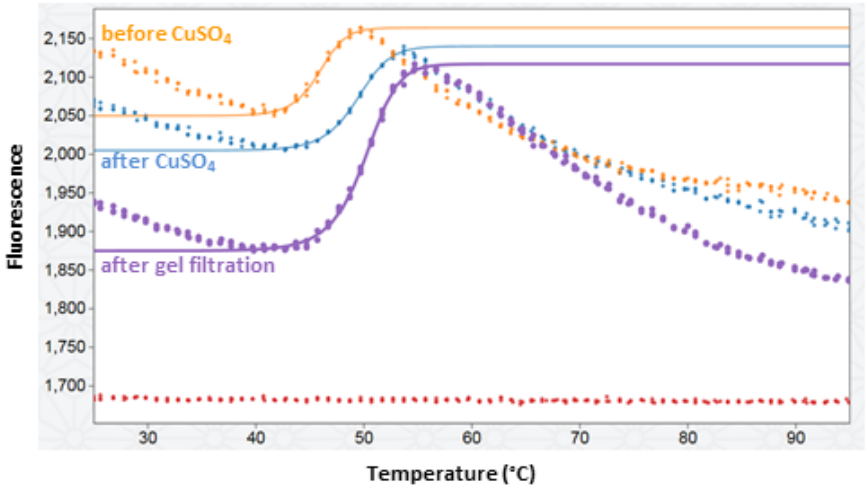


Figure 5.28. Thermal shift assay for *Lps*AA10 before (yellow line) and after (blue line) loading with $CuSO_4$ and after performing gel filtration (purple line). Buffer only was used as a negative control (red line).

5.7.3 Activity assays

5.7.3.1 MALDI/TOF-TOF tandem mass spectroscopy

*Lps*AA10 is annotated by the NCBI nrdb as an "auxiliary activity family 10 domain-containing protein". The Auxiliary Activities (AA) family was recently introduced in the CAZy database to incorporate those proteins that do not hydrolyse glycosidic bonds or build complex carbohydrates such as GHs, PLs, CEs or GTs (Levasseur *et al.*, 2013). The AA class includes enzymes that are either active on lignin or that oxidatively break down crystalline cellulose or xylan, and are classed as lytic polysaccharide mono-oxygenases (Quinlan *et al.*, 2011). The AA10 family in particular (formerly CBM33) has been shown to have activity on chitin or cellulose and has been found in eukaryotes, bacteria, archaea and viruses (Vaaje-Kolstad *et al.*, 2010; Forsberg *et al.*, 2011).

In order to test whether *Lps*AA10 is indeed an LPMO active on chitin or cellulose, reactions were set up with the purified enzymes and different substrates that included crystalline cellulose (PASC and Avicel), chitin (squid, shrimp and colloidal chitin) and hemicelluloses (glucomannan, β-glucan and xylan) using ascorbic acid or gallic acid as electron donor for the reaction. The reactions were run overnight at 30 °C with shaking and the products were analysed by MALDI/TOF-TOF to detect any oxidation product (details in section 2.8.4). No activity was observed for the chitin and hemicellulose substrates, while oxidation products were detected for both forms of crystalline cellulose (the relative peaks are shown in Fig. 5.29 and 5.30), indicating activity on both PASC and Avicel.

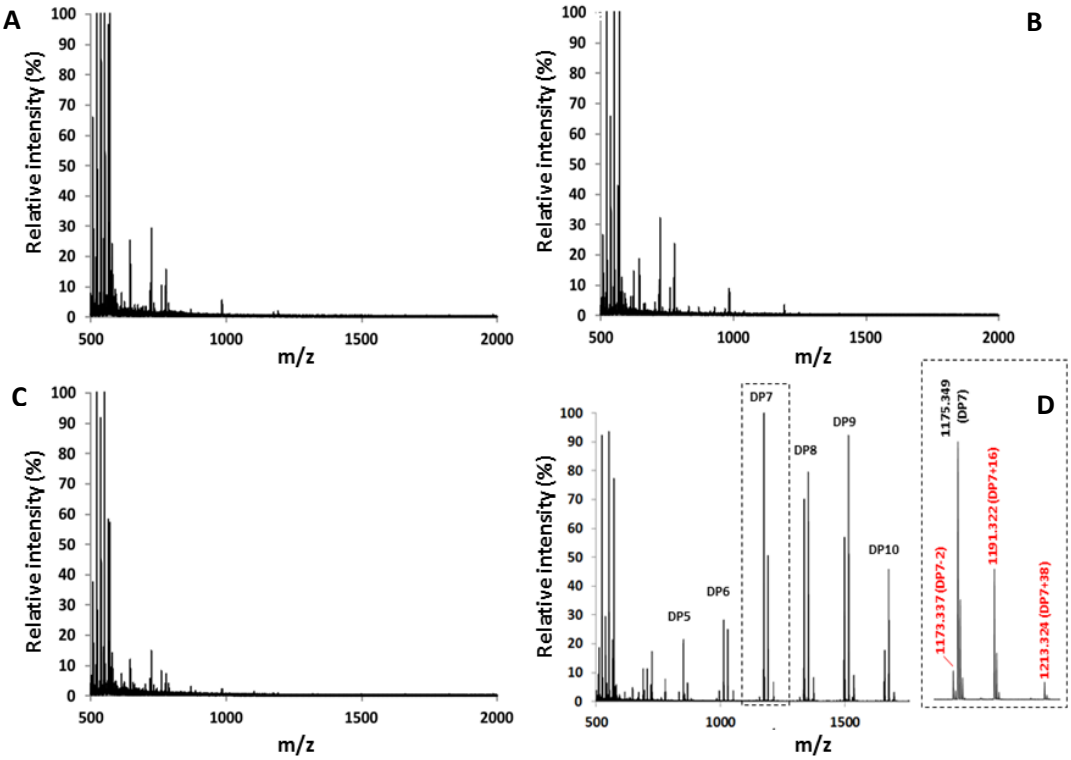


Figure 5.29. MALDI/TOF-TOF tandem mass spectroscopy analysis of activity assays with purified *Lp*sAA10 and PASC, with gallic acid as electron donor. The panels A-C show the spectra of the negative controls after incubation of: **A.** 4 mg/ml PASC; **B.** 4 mg/ml PASC and 4 mM gallic acid; **C.** 4 mg/ml PASC and 2 µM *Lp*sAA10. **D**. Actual reaction with 4 mg/ml PASC, 2 µM L *Lp*sAA10 and 4 mM gallic acid (**D**), where the dotted inset is an expanded mass spectrum for DP7 products. The oxidized products are marked in red. In all panels, 100% relative intensity represents 1.0 x 10^4 arbitrary units (a.u.) and m/z is the mass/charge ratio.

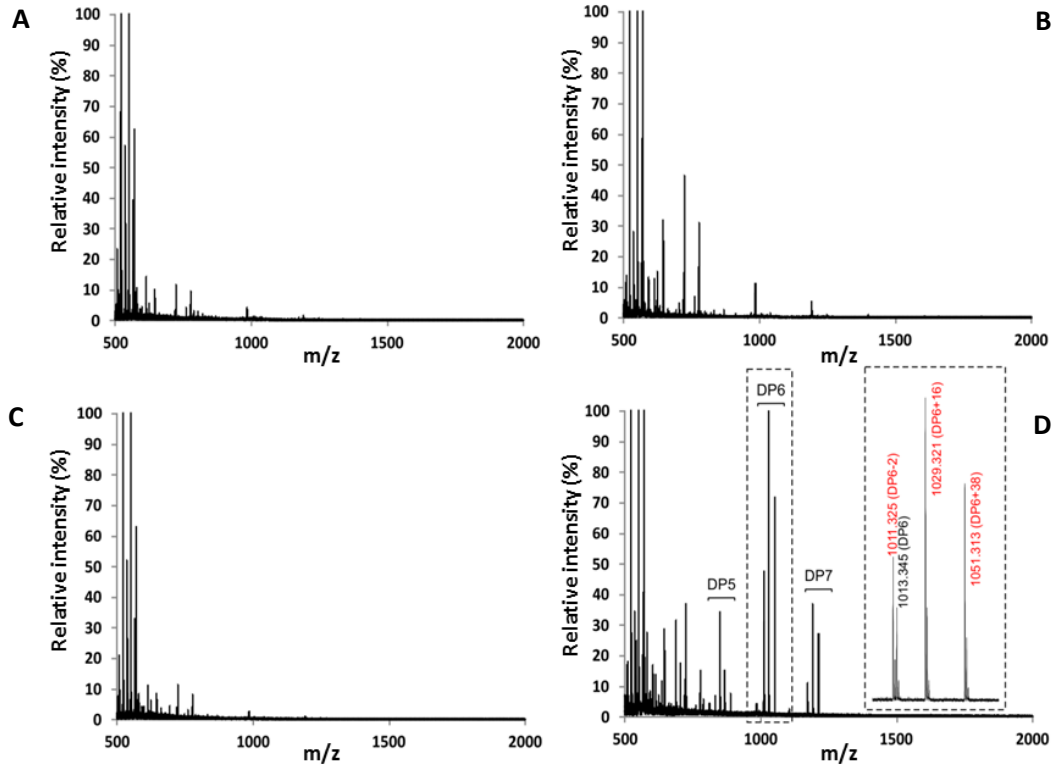


Figure 5.30. MALDI/TOF-TOF tandem mass spectroscopy analysis of activity assays with purified *Lp*sAA10 and Avicel, with gallic acid as electron donor. The panels A-D show the same as in Figure 5.26 but with PASC instead of Avicel and the 100% relative intensity represents 1.3 x 10^4 arbitrary units (a.u.).

The assays revealed a predominant C1-oxidation pattern and generation of C1-aldonic acids on both PASC and Avicel in the presence of an external donor, imparting =16 or =38 m/z respectively, relative to the mono-sodiated unoxidaside form. Smaller peaks for the mono-sodiated lactone (-2) were also identified. Both electron donors produced oxidation, while only the results obtained with gallic acid are shown, as this seemed to be a more effective reductant than ascorbic acid.

5.8 Discussion

The attempts to express the shipworm lignocellulolytic enzymes in heterologous hosts were time consuming and challenging, with the majority of the endogenous proteins not being possible to express in a soluble form and characterise, and only a few bacterial sequences producing successful results. The reason for these difficulties can probably be identified in the nature and environment of the proteins under study, namely the fact these proteins are found in a challenging environment such as seawater. Indeed, the characteristics of this environment shape the protein's structure, resulting in the presence of many disulphide bridges and possibly other post-translational modifications

(Hatahet *et al.*, 2014), making them quite difficult to express. Similar difficulties were posed to colleagues in our research institute when they tried to express CAZymes produced by the marine woodborer *Limnoria quadripunctata* (Besser, personal communication); after many years of attempts to heterologously express some of this crustacean's enzymes, only a GH7 was successfully produced and characterised (Kern *et al.*, 2013). O'Connor and colleagues (2014), who challenged themselves with the expression of endosymbiont-encoded proteins from the shipworm *Bankia setacea*, had more success but were able to produce only very small protein amounts and mainly by using in-vitro transcription/translation systems.

Despite these challenges, it was possible to successfully produce a number of bacterial and endogenous proteins, which are among the most expressed or abundant in the animal's digestive organs. The annotations produced by the NCBI nrdb and by dbCAN suggested that three of these enzymes are glycoside hydrolases active on mannans, and one on xylans. The characterisation performed on the recombinant proteins confirmed that *Lp*sGH134a and *Lp*sGH134b are indeed β-mannanases with an endo- rather than exo-activity, and that *Lp*sGH5_8 is also active on gluco and galactomannans, but it was not possible to confirm whether it has a β-mannosidase (as suggested by the annotation) or β-mannanase activity. *Lp*sGH11 was confirmed to have activity on xylans and to have most likely an exo-xylanase activity, though it is not clear if it can also cleave β-1,3 linked xylans. The fact that these CAZymes found in abundance in the shipworm digestive organs are active on mannans or xylans is not surprising, since animals that digest wood need to have an extensive range of enzymes active on hemicelluloses, which constitutes around 25-30% of the total wood dry weight (Pérez *et al.*, 2002); mannan and xylan in particular are the major carbohydrates found in hemicelluloses (Pérez *et al.*, 2002), and they were indeed the most abundant hemicelluloses revealed by the wood compositional analysis presented in this thesis in chapter 3.4.2.2. Furthermore, the animals used in these studies were grown exclusively on Scots pine, which is a softwood and therefore particularly rich in galactoglucomannan and arabinoglucoronoxylans (Fengel and Wegener, 1989; Moreira and Filho, 2008), therefore it is appropriate for the shipworm to produce different types of mannanases and xylanases in substantial amount, to be able to perform efficient wood deconstruction. Indeed, Shipway (2013) identified numerous mannanases (GH families 2, 5 and 26) and xylanases (GH families 5, 7, 8, 10, 11 and 43) in *L. pedicellatus* digestive gland transcriptome, and O'Connor and colleagues (2014) pulled mainly mannanases and xylanases out of the main

band found in the SDS-PAGE gel performed with the total caecum content extract of the shipworm *B. setacea*.

The fifth bacterial protein that was heterologously expressed was *Lp*sAA10, an LPMO with suggested activity on chitin or cellulose, and characterisation confirmed that it is able to oxidase crystalline cellulose in the form of both Avicel and PASC. LPMOs have been reported in the gill endosymbiont metagenome and in the caecum content of the shipworm *B. setacea* (O'Connor *et al.*, 2014), but their annotation as AA10 enzymes has not been tested to confirm the preferred substrate. AA10s are abundant in the genome of bacteria and viruses, while relatively rare in archaea and eukaryotes such as fungi and plants (Moser *et al.*, 2008; Vaaje-Kolstad *et al.*, 2010). It is therefore not surprising that the shipworm's endosymbiotic bacteria contain an AA10 active on cellulose in their repertoire of lignocellulolytic enzymes, but it is interesting to see how *L. pedicellatus* has taken advantage of this symbiont encoded CAZyme for its own benefit (= digestion of wood), rather than producing one of its own. Indeed, the shipworm transcriptome was searched for LPMOs, but no AA10s or other LPMO families were found to be endogenously produced in this species, suggesting that endogenous LPMOs are not involved in wood digestion in these molluscs. This is different from what has been recently discovered for the firebrat *Thermobia domestica*, an hexapod that recruits many different endogenous LPMOs enzymes to be able to perform and extremely efficient digestion of cellulose (Sabbadin *et al.*, 2018a). *Lp*sAA10 is also the first LPMO to be expressed and characterised from a shipworm endosymbiont.

The bacterial proteins *Lp*sGH5_8, *Lp*sGH134a and *Lp*sGH134b were tested to verify how their activity would change depending on different environmental parameters, such as pH, temperature and salinity, which affect protein stability and therefore performance (Jaenicke and Zavodszky, 1990; Jaenicke, 1991) and are the most important to be considered in industry when preparing enzyme cocktails for lignocellulose digestion. Organisms normally adapt their proteins so that they perform most efficiently within the range of parameters of the environment in which they live (Hochachka and Somero, 1973; Low and Somero, 1976; Feeney and Osuga, 1976; Argos *et al.*, 1979). This is likely to be case also for *L. pedicellatus* from the Atlantic lineage (the population sampled), where their proteins are likely to have maximum activity within the temperatures, pH and salinity range found in the Atlantic sea. A study from Borges and colleagues (Borges *et al.*, 2014) revealed that the Atlantic form of *L. pedicellatus* is found at temperatures ranging

from 5 to 25 °C and in waters with salinity from 33 to 37 PSU, while the tolerance for the Mediterranean shipworms is broader, since they live at temperatures that vary from 7 to 40 °C and with a salinity from 17 to 40 PSU. The pH of Atlantic seawater is reported to be 8.0-8.2 (Salt *et al.*, 2013), while in the Mediterranean Sea is slightly lower at 7.8-7.9, and has recently decreased due to acidification caused by increased CO_2 emissions (Flecha *et al.*, 2015). *Lp*sGH5_8 did not appear to be active at acidic pH, with activity only starting from pH around 6.0 and peaking at pH of 7.0-8.0, showing a tolerance to basic pH up to 11.0, the maximum that was tested. These results fit well with the theoretical isoelectric point obtained for the protein (4.65) and with studies performed by homology modelling and surface charge analysis on GHs from marine wood borers, which showed how these enzymes exhibit strong acidic charges on their surfaces and are able to retain activity at high salt concentrations (Kern *et al.*, 2013). This seems to be an adaptation to the marine environment and a characteristic of halofilic proteins, which need to remain active at high ionic strength values (Kumar, 1998; Delgado-García *et al.*, 2012). *Lp*sGH134a and *Lp*sGH134b instead showed a marked preference for pH around neutrality (6.0-8.0), with a clear peak at 7.0, and activity was still present at both ends of the pH spectrum, even if it is reduced, despite their pI being quite acidic (4.51 and 4.63 respectively). All these three proteins however show best activity levels at pH similar to that of seawater, though no research has been done so far to test the actual pH level of the shipworm digestive system and in particular of the caecum, where the proteins are operational. Indeed, research performed on the woodborer *Limnoria quadripunctata* revealed that the pH of its digestive system was different to that of the seawater, reaching the value of 5.6 in the distal lobes of the hepatopancreas (Dr Katrin Besser, University of York, personal communication, 2018). The preferred salinity of *Lp*sGH5_8 (0.5 M) reflects the one of the seawater of the Atlantic, while *Lp*sGH134a and *Lp*sGH134b have no tolerance for salt and are active only in its absence; this is difficult to explain in a marine animal, which should be well adapted to live at high concentrations of salts. Optimal temperature for activity also showed a different trend in the proteins tested, with *Lp*sGH5_8 showing low activity at 10-30°C, optimal conditions between 30-50 °C, and a sudden loss of function at higher temperatures; these results are broadly in line with its melting temperature of 43 °C, which however has to take into consideration the presence of more than one module in the protein structure. *Lp*sGH5_8 temperature range therefore appears to be more adaptive for the Mediterranean Sea rather than the Atlantic, where water temperatures

up to 40 °C can be reached during the summer in coastal areas, where the shipworms are found. However, correlations between *L. pedicellatus* enzyme activity and the environment of one or the other of the lineages are not appropriate; indeed, despite the two populations having been reported as possible cryptic species (Borges *et al.*, 2012), there is no information on when the separation happened and whether the proteins could have had the time to evolve and adapt to the new environment. *Lp*sGH134a and *Lp*sGH134b instead preferred a broader temperature range, showing a best activity between 10 and 50 °C, with a peak at 30-40 °C, and sudden inactivity after 50 °C. Despite these differences, both proteins seem to be able to work at the temperatures broadly in the range of those recorded in both the Atlantic and Mediterranean Sea, while they are not adapted to extreme temperatures.

To complete the analysis of the expressed proteins, and evaluate their possible industrial application in the contest of the production of second-generation biofuels, experiments to identify their optima in terms of temperature and pH combined together were performed. Indeed, making separate measurements for the above parameters (which was performed initially in this chapter) does not allow to predict performance when the parameters are combined together (Herlet *et al.*, 2017). The enzymes optimal activity range is a useful parameter to determine when evaluating the possible use of an enzyme in industrial processes that have to happen at precise temperature and pH values, and often in a mixture with different enzymes. None of the bacterial proteins tested seem to be particularly stable under the harsh conditions that are required for the saccharification process in industry, which involves high temperatures and acidic pH (Khare *et al.*, 2015). *Lp*sGH5_8 though seems to be adapted to alkali conditions, which are also used in industry to perform biomass digestion. The three proteins also do not appear to be adapted to perform at high salt concentrations, despite originating from a marine animal; salt tolerance would have been useful trait given the current search for halophilic enzymes to use in seawater, therefore avoiding the usage of the declining freshwater reserves (Fingerman *et al.*, 2010). However, the hot spots of *Lp*sGH134b and *Lp*sGH5_8 were found to be quite similar, which would allow to use the two proteins in the same enzyme cocktail at set conditions.

Chapter 6: Final discussion

6.1 General discussion

Anthropogenic global warming, diminishing fossil fuel reserves and food security are the three main drivers for the utilisation of plant biomass as a feedstock for the production of second-generation biofuels. The recalcitrance of the plant cell wall to deconstruction into simple sugars has so far been the major hindrance for the establishment of a successful sustainable biofuel industry. However, nature has long ago learnt how to overcome this problem, since organisms such as bacteria, fungi, protists, marine and land invertebrates are able to utilise lignocellulosic material for their metabolism. The work presented in this thesis aims to increase our understanding of wood digestion in one of these organisms, the shipworm *Lyrodus pedicellatus,* to shed new light on its effective biomass digestion system.

L. pedicellatus was selected for this study because it is the only shipworm (possibly together with *T. navalis*) able to obtain all its nutritional requirements from wood only (Becker, 1959; Gallager *et al.*, 1981), while other shipworms rely also on filter feeding, though in different degrees depending on the species. Since the discovery of lignocellulolytic bacteria in the shipworm gills, the scientific community has attributed almost entirely to the endosymbionts the wood-degrading abilities, suggesting that they are indispensable to the shipworms to perform a complete and effective digestion of wood (O'Connor *et al.*, 2014). While it is true that the bacteria are crucial to the shipworm survival because of their nitrogen-fixing role, the work presented in this thesis suggests how they play a smaller role in wood digestion than previously thought, and that the shipworm instead produces many of its own endogenous lignocellulolytic enzymes. Indeed, proteomic analysis of the enzymes found inside the caecum has revealed that 87.6% of the total of the CAZymes are of endogenous origin, while only 12.4% are prokaryotic (12.4%). SDS-PAGE analysis of the caecum fluid total protein has also highlighted a prevalence of shipworm produced CAZymes rather than bacterial ones. Nevertheless, the bacteria appear to be specialised in producing certain CAZymes families, such as GH5, GH6, GH11, CE15 and AA10, while the shipworm produces mostly GH9, GH1, GH45, GH2 and GH30, as shown by the transcriptomic data. This arrangement probably allows both parties to save energy and the result is a more efficient degradation of wood than it would have been possible without the symbiosis. A similar strategy is probably used by other bivalve molluscs closely related

to shipworms, such as the sea piddocks (Pholadidae), which live at depths greater than the Teredinids and are known to bore into sunken wood. The presence of bacterial endosymbionts in the gills of the species *Xylophaga atlantica* and *X. washingtona*, together with a caecum packed with wood chips, suggests that they have probably evolved a symbiotic solution to terrestrially derived wood digestion similar to that of shipworms, which has though been adapted to a deep-sea environment (Distel and Roberts, 1997). A comparison between these deep-sea wood-eating bivalves and those inhabiting shallow waters (in terms of the genes, enzymes and organs involved in the digestive processes) could provide more insights into the details wood digestion.

The existence of a gill endosymbiotic bacterial community is not uncommon in bivalves, and indeed many species of non-wood eating bivalves are known to receive nutritional benefits from the association with symbiotic bacteria (Savazzi, 2001). Bivalves belonging to the Bathymodiolinae mytilid subfamily have symbionts that are not able to degrade cellulose, but are chemoautotrophic or methanotrophic, enabling the host to colonise environments such as cold seeps or hydrothermal vents (Kenk, 1985). The species *Idas washingtonia* is found on whale carcasses at the bottom of the sea, where the chemoautotrophic gill bacteria allow the bivalve to make use of the hydrogen sulphide produced by the decomposition of the lipids found in the bones (Bennett *et al.*, 1994; Smith *et al.*, 1989). Similarly, the shipworm *Kuphus polythalamia* lives in marine sediments assisted by the sulphur oxidizing chemosynthetic gills symbionts (Distel *et al.*, 2017). It has been suggested that the bivalves that thrive on decomposing wood or whales carcasses could be the evolutionary link that allowed the introduction of the chemosynthetic mussels in the vents and seeps (Distel *et al.*, 2000; Distel *et al.*, 2017). The wood boring shipworms inhabiting shallow water such as *L. pedicellatus* are closely related, and they share a similar type of the symbiosis, with bacteria resident in the gill's bacteriocytes.

The analysis of the transcriptome of the digestive organs of *L. pedicellatus* has revealed which enzymes have been recruited by the shipworm for wood digestion. CAZymes belonging to the GH9 family, as well as GH45, GH2 and GH1, are produced in the digestive glands and represent 88.9% of the CAZymes expressed in this organ. These enzymes are annotated as putative cellulases, endoglucanases, β-glucosidases or β-mannosidases. When these results are compared to the fully annotated genome of the model suspension-feeding bivalve *Crassostrea gigas*, it appears that this oyster is much more adapted to

digest polysaccharides abundant in phytoplankton cell walls (Pomin and Mourão, 2008), since enzymes belonging to the families of GH16, GH19, GH17, GH35, GH38, GH47 and GH3 are differentially expressed by the digestive gland of this species. *L. pedicellatus* is on the contrary more adapted to the digestion of polysaccharides contained in wood, which could indicate a preference for wood boring rather than filter feeding. In this context, it would be interesting to compare *L. pedicellatus* CAZymes expression levels to that of another shipworm species, *Zachsia zenkewitschi.* This small Teredinid has evolved to live in a highly-specialised environment, the rhizomes of the seagrass *Zostera* and *Phyllospadix* (Shipway *et al.*, 2016). These marine plants are angiosperms dicots, which have retained the ability, contrary to other marine angiosperms, to produce sulphated polysaccharide. Their cell wall composition is quite different to that of land plants, being composed by 57% of cellulose, 38% of non-cellulosic polysaccharides (mainly xylan) and only 5% of lignin (Davies *et al.*, 2007). The transcriptome of Z. *zenkewitschi* is, therefore, likely to be more similar to that of *C. gigas* rather than *L. pedicellatus,* possibly showing production of GHs and sulfatases involved in the deconstruction of sulphated polysaccharides, enzymes that are not found in *Lyrodus* transcriptome. It is also possible that the shipworm digestive glands express different types of CAZymes during different phases of the animal's life cycle, with transcripts related to filter feeding being more expressed during the early larval stages prior to the attachment to the timber, when the animals are not yet relying for nutritional purposes on the presence of wood, but rather on plankton. Different enzymes could also be expressed when feeding on different types of wood, for example hardwood instead of softwood.

When comparing the mechanisms developed by *L. pedicellatus* for the digestion of terrestrial lignocellulose to those of others wood-degrading systems, interesting differences emerge. Compositional analysis, comparing the components of wood to the shipworm frass, revealed how *Lyrodus* utilises mainly the cellulose and only partially the hemicellulose fractions. The analysis has not highlighted any reduction of the lignin fraction, which on the contrary accumulates in the frass. These results are not surprising, since only fungi have so far been recognised as able to mineralise or at least modify the lignin. Fungi possess lignin-modifying enzymes or use Fenton chemistry to target the lignin fraction, but none of these systems have been discovered by the analysis of *L. pedicellatus* transcriptome, though we cannot exclude they are found among those transcripts for

which an annotation was not possible (around 38% of the total). Only one transcript, annotated as a "hephaestin-like protein" and similar to the protein of various bird species, was found to be expressed in the shipworm caecum and digestive glands and to have a structure that resembles a laccase (with six copper-containing domains). The shipworm hephaestin could, therefore, be involved in lignin degradation, even if the amounts found in the caecum contents proteome are quite low. Unfortunately, this enzyme could not be recombinantly expressed and purified in quantities sufficient for characterisation.

Termites are the most efficient wood-degrading organisms, but this task is so proficiently performed because their digestive system hosts a diverse community of microorganisms that includes aerobic and anaerobic bacteria, fungi and protozoa, all of which contribute to the production of lignocellulolytic enzymes (Warnecke *et al.*, 2007). Complex communities able to digest lignocellulose are also hosted by wood-eating fish (Watts *et al.*, 2013) cows and other ruminants. The meta-transcriptomic and meta-proteomic analysis of *Lyrodus* digestive system revealed that the digestion of wood in this species is instead the work of only two types of organisms, the shipworm and the gill's endosymbiotic bacteria, as no enzymes produced by archaea, fungi or protozoa were identified. The symbionts belong to different (but closely related) species and are present in great numbers; however, it is remarkable how this relatively simple prokaryote-eukaryote system has evolved to perform such an effective digestion of wood. Even more striking are the isopod crustaceans belonging to the genus *Limnoria*, which are able to digest wood and possibly even modify the lignin fraction without the assistance of any symbionts. This task is likely to be accomplished by the action of hemocyanins found in abundance in the digestive system. These proteins are normally found in the haemolymph and are involved in oxygen transport, but in these isopods they could function as phenoloxidases and, therefore, have a role in the degradation of lignin (King *et al.*, 2010). Similar proteins have not been identified in the shipworm transcriptome. These bivalve wood borers seem to have evolved quite a complex and multifaceted system, which allows them to live on a diet of wood. Not only do they host the already mentioned endosymbiotic bacterial communities in the gills, but they also have developed a system (mostly likely localised in the food groove) to transport the bacterial enzymes to the distant digestive system. With the help of the crystalline style, they appear to be able to release the bacterial CAZymes from the bacteriocytes and to utilise them for wood digestion, while at the same time they make use of the abundant bacterial cells as a source of nitrogen. Amoebic cells found in the white

portion of the digestive glands also give a contribution – though so far not fully understood – to wood digestion by phagocytising wood particles.

The heterologous expression of shipworm and bacterial wood degrading enzymes selected from the proteome and/or transcriptome of the digestive organs was the most challenging and time-consuming task of the project. Numerous host organisms were used, as well as different combinations of competent cells, culture media, expression temperatures, induction time and concentration, solubility and purification tags, and other variables. Only one endogenous shipworm enzyme was successfully expressed in quantities that allowed downstream analysis, a multimodular GH1, however the work is not part of this thesis (Sabbadin *et al.*, 2018b). More positive outcomes were obtained from the expression of the bacterial proteins, whose characterisation revealed that they are capable of breaking down hemicelluloses, particularly galacto-glucomannans, and xylans. The most abundant CAZymes found in the crystalline style is a GH5_8, which proved to be active on galactomannans and glucomannans. A bacterial AA10 LPMO was also recombinantly expressed and characterised, revealing its ability to oxidatively cleave cellulose at the C1 position. To our knowledge, this is the first example of a characterized LPMO from an animal endosymbiont, a finding that carries interesting evolutionary implications. Recently, Sabbadin and colleagues (2018a) have identified and characterized the first endogenous LPMOs (family AA15) in animals and shown how some primitive insects, such as silverfish and firebrats, have co-opted these enzymes to digest plant biomass. Phylogenetic analysis also revealed that AA15s are widespread among insects as well as other invertebrates, including crustaceans, chelicerates and molluscs. Indeed, several putative AA15 sequences were identified in the *L. pedicellatus* transcriptome, although their expression appeared to be low in all analysed tissues and, therefore, the transcripts were not among those presented in chapter four. Unlike firebrats, where proteomic analysis showed that endogenous AA15s are abundant in the digestive system and are crucial for cellulose digestion (Sabbadin *et al.*, 2018a), no significant amounts of AA15s were detected in the proteomic studies of the digestive organs of *L. pedicellatus*, suggesting that endogenous LPMOs are not involved in wood digestion in these molluscs. Perhaps anatomical or energetic constraints in shipworms have hindered the recruitment of endogenous AA15s in wood digestion, while the existing food groove turned out to be a convenient route for bacterial AA10 LPMOs coming from the gills, together with the other bacterial CAZymes.

This thesis presents a multifaceted study of the digestive system of *L. pedicellatus*, revealing the complex molecular and physiological processes of wood digestion of these bivalves. Particular emphasis was given to the study of the symbiosis with the gills bacteria, given that this association was probably the driver that allowed the shipworm to colonise a new ecological niche by developing novel metabolic capacities. Shipworms carry out and important ecological function in the coastal marine ecosystem by unlocking the carbon found in wood, but they are also formidable destroyers of man-made wooden structures and cause of serious economic damage. The results presented in thesis are therefore valuable, since the information on the range and relative abundance of bacterial and endogenous CAZymes can be used both in the second-generation biofuel industry (by helping to formulate the optimal enzymatic cocktail for the industrial break-down of lignocellulose) and in the management of coastal environments afflicted by the problem of shipworm damage.

6.2 Future work

While the work presented in this thesis gives a deeper view of the mechanisms of wood digestion in *L. pedicellatus*, at the same time it generates new questions about the fine details of these processes.

The food groove was identified as the route followed by the bacterial CAZymes from the gills to the caecum. However, it is still not clear whether only bacteriocytes or also single bacteria travel down this path, and whether the enzymes secreted by the bacteria remain inside the bacteriocytes-containing vesicles or are released outside the eukaryotic cells. The difficulty of fixing tissues that contain high amount of mucous, such as the food groove, has been the main obstacle for a definitive answer to this question, alternative fixatives, such as Carnoy's solution (Kamphuis *et al.*, 2017), should be explored, as well as other imaging or molecular techniques.

The mechanism by which the bacterial CAZymes are preferentially selected for transport to the caecum among the other bacterial proteins has also so far remained unexplained. Some work (not presented in this thesis) was carried out to try to link the presence of CBM10s to this selection mechanism, given that all the bacterial CAZymes that are found in the gills, crystalline style or caecum proteome possess from one to three of these modules, which are known to bind to cellulose. However, this hypothesis could not be verified and further work is required to give an answer to this interesting question. Evidence is also needed for

the proposed hypothesis of the involvement of the crystalline style in the mechanical and enzymatic breakage of the bacteriocytes that have reached the stomach, as well as the contribution given by the bacteria to the nutritional needs of the shipworm.

The presence of modifications in the lignin fraction after digestion by the shipworm has also not been explored. The compositional analysis highlighted quantitative changes, indicating that lignin accumulates in the frass and is not utilised for metabolism. However, there is the possibility that, as it happens in fungi, the lignin is modified by the action of enzymes or free radicals, in order to expose the carbohydrates of the cellulose and hemicellulose fractions. These types of enzymes or chemistry could not be identified with the work presented in this thesis, mainly due to the difficulties encountered with heterologous protein expression. Indeed, some interesting targets were identified in the proteome and transcriptome of the shipworm caecum, such as the above-mentioned laccase-like hephaestin-like proteins, some DBH-like monooxygenase proteins and a thioester-containing protein. Further efforts should be put into trying to characterise these promising proteins, by attempting expression in modified or new host systems or by purifying native proteins directly from the animal tissues. At the same time, evidence of lignin modification could be explored with the use of techniques such as FTIR spectroscopy or ssNMR.

Work to complete the characterisation of the bacterial enzymes could also be carried out to produce some useful results. *Lp*sGH11 incubated with the favourite substrates could be studied with PACE and heat maps analysis, in order to determine the nature of the digestion products and the range of temperature, pH and salinity it is most active at, similarly to what was done for the other bacterial GHs. The kinetics of the enzymes could also be explored, determining their V_{max}, K_m and k_{cat}/K_m, to obtain further information on the enzymes performance and determine whether they have more efficient than those

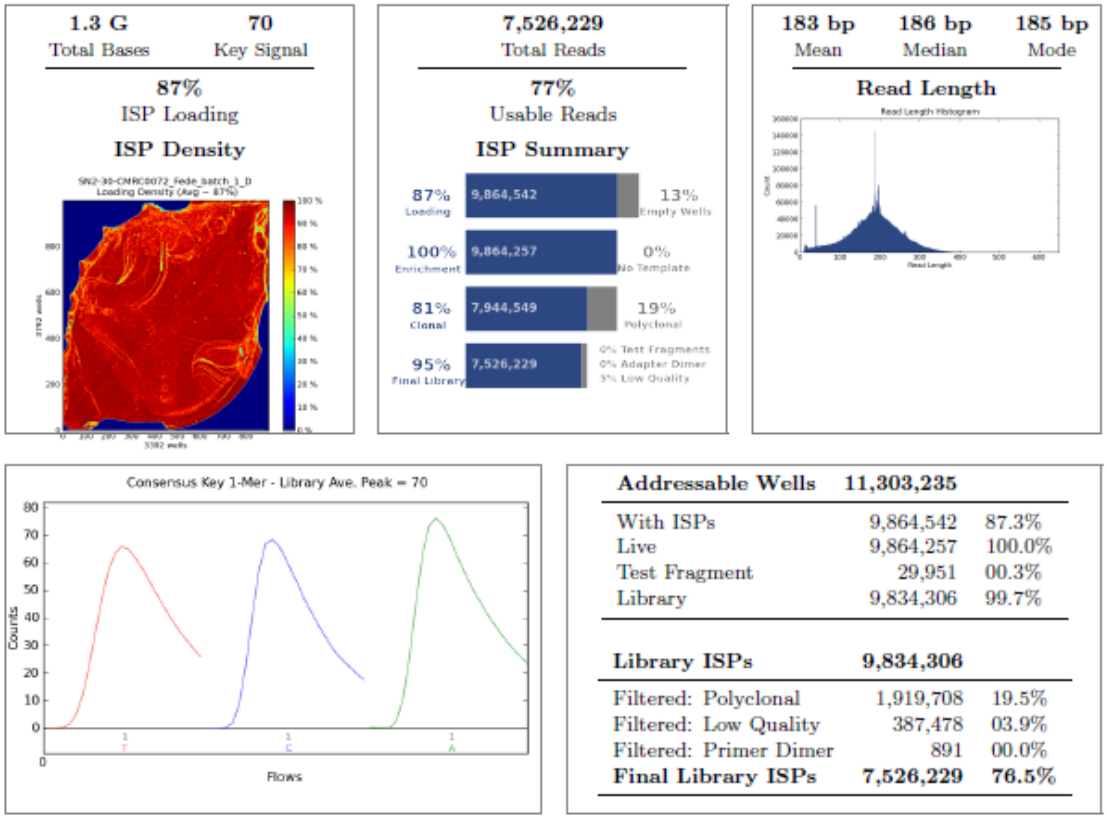
added to the enzymatic cocktails used in industry. Co-incubation experiments could be performed with *Lp*sGH5_8 and *Lp*sGH134a or *Lp*sGH134b, to look for the existence of possible synergisms, as well as with *Lp*sAA10 and any other cellulase such as GH5, GH6, GH7, GH1 or GH8, to search for the presence of a boosting action that increases the reaction yields, as has been demonstrated for other LPMOs (Sabbadin *et al.*, 2018a). The enzymes performance at different salinities could also be tested on real or artificial seawater, rather than on water with different molarity of just the salt sodium chloride, to assess the performance in conditions more similar to the shipworm's natural environment.

Appendices

Appendix A

Ion Torrent run summary for the digestive glands, caecum and gills samples, respectively. The figures report information on the quality and quantity of the Ion Sphere Particle (ISP) load onto a 318 chip, as well as the properties of the obtained EST library.

Run Summary



Addressable Wells	11,303,235	
With ISPs	9,864,542	87.3%
Live	9,864,257	100.0%
Test Fragment	29,951	00.3%
Library	9,834,306	99.7%
Library ISPs	**9,834,306**	
Filtered: Polyclonal	1,919,708	19.5%
Filtered: Low Quality	387,478	03.9%
Filtered: Primer Dimer	891	00.0%
Final Library ISPs	**7,526,229**	**76.5%**

Barcode Name	Sample	Bases	$\geq Q20$	Reads	Mean Read Length
No barcode	None	7,127,705	5,761,511	52,305	136 bp
IonXpressRNA_008	D2	470,837,208	409,617,769	2,626,160	179 bp
IonXpressRNA_009	D3	411,392,142	353,727,538	2,176,127	189 bp
IonXpressRNA_010	D5	490,217,093	423,855,938	2,671,556	183 bp

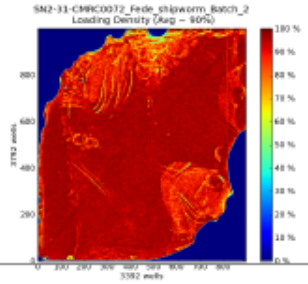
Test Fragment	Reads	Percent 50AQ17	Read Length Histogram
TF_A	**28,204**	**95%**	

Run Summary

1.2 G	67
Total Bases	Key Signal

90%
ISP Loading

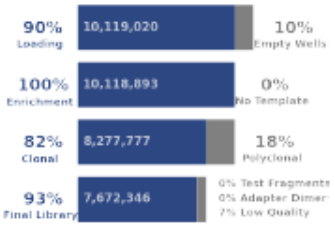
ISP Density



7,672,346
Total Reads

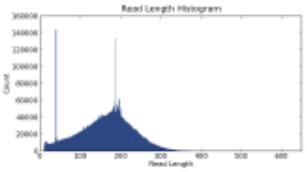
76%
Usable Reads

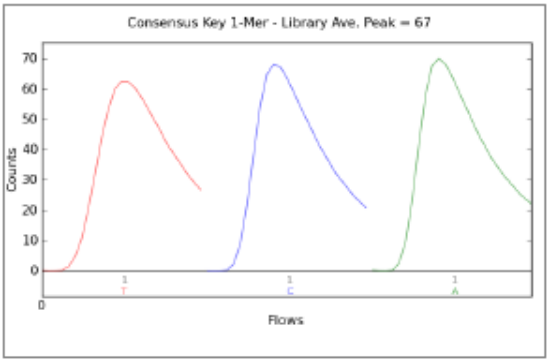
ISP Summary



161 bp	169 bp	39 bp
Mean	Median	Mode

Read Length



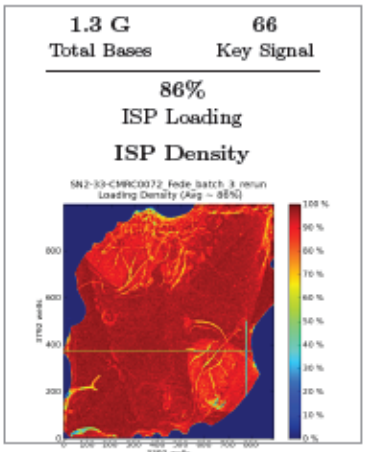


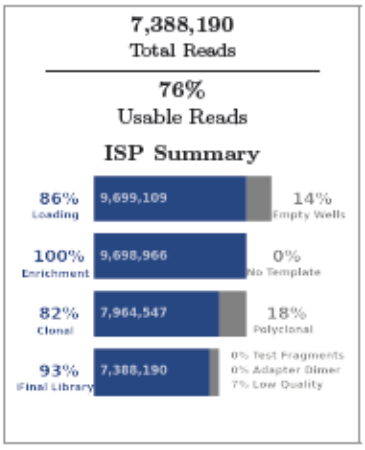
Addressable Wells	**11,305,324**	
With ISPs	10,119,020	89.5%
Live	10,118,893	100.0%
Test Fragment	23,307	00.2%
Library	10,095,586	99.8%
Library ISPs	**10,095,586**	
Filtered: Polyclonal	1,841,116	18.2%
Filtered: Low Quality	580,953	05.8%
Filtered: Primer Dimer	1,171	00.0%
Final Library ISPs	**7,672,346**	**76.0%**

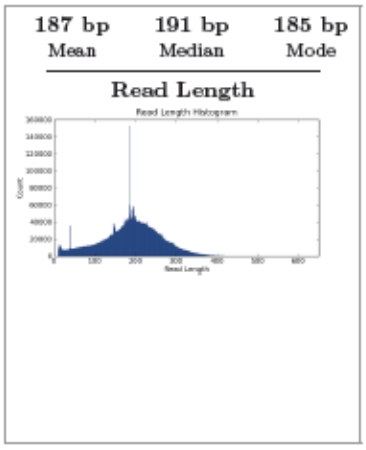
Barcode Name	Sample	Bases	$\geq Q20$	Reads	Mean Read Length
No barcode	None	6,742,047	5,177,317	52,539	128 bp
IonXpressRNA_011	C2	394,887,924	317,468,393	2,552,540	154 bp
IonXpressRNA_012	C3	412,673,191	334,080,906	2,535,732	162 bp
IonXpressRNA_013	C5	422,965,509	348,592,723	2,531,457	167 bp

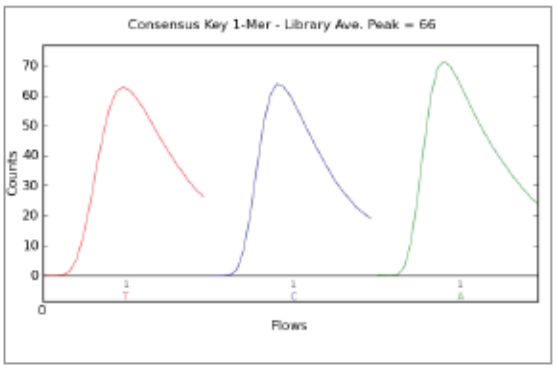
Test Fragment	Reads	Percent 50AQ17	Read Length Histogram
TF_A	**20,333**	**81%**	

Run Summary









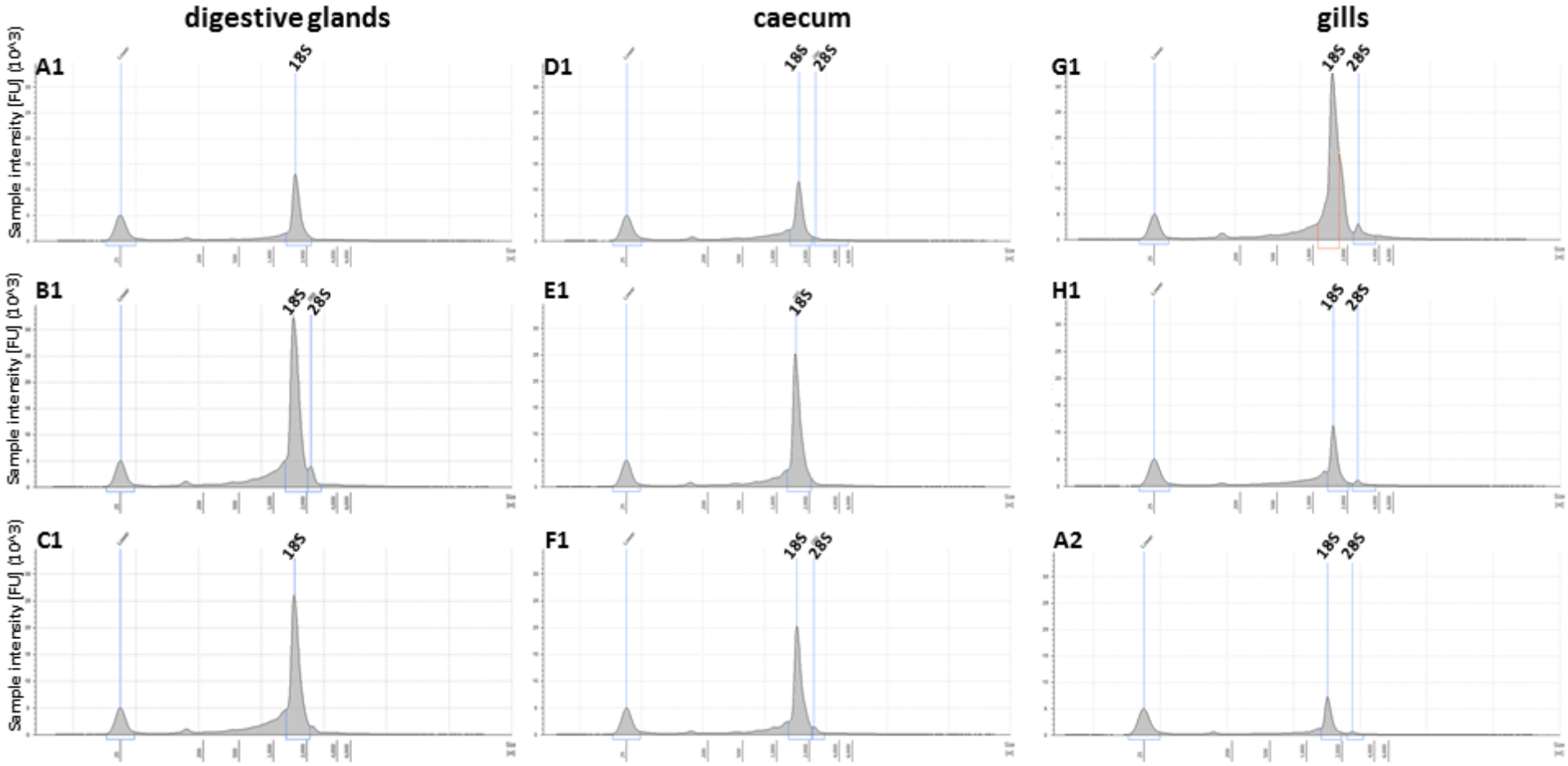
Addressable Wells	**11,307,713**	
With ISPs	9,699,109	85.8%
Live	9,698,966	100.0%
Test Fragment	17,361	00.2%
Library	9,681,605	99.8%
Library ISPs	**9,681,605**	
Filtered: Polyclonal	1,734,419	17.9%
Filtered: Low Quality	558,459	05.8%
Filtered: Primer Dimer	537	00.0%
Final Library ISPs	**7,388,190**	**76.3%**

Barcode Name	Sample	Bases	$\geq Q20$	Reads	Mean Read Length
No barcode	None	13,463,279	9,923,938	89,973	149 bp
IonXpressRNA_014	G2	402,705,382	323,645,988	2,111,183	190 bp
IonXpressRNA_015	G3	604,256,681	478,443,170	3,206,565	188 bp
IonXpressRNA_016	G5	362,074,792	295,105,374	1,979,545	182 bp

Test Fragment	Reads	Percent 50AQ17	Read Length Histogram
TF_A	**14,251**	**78%**	

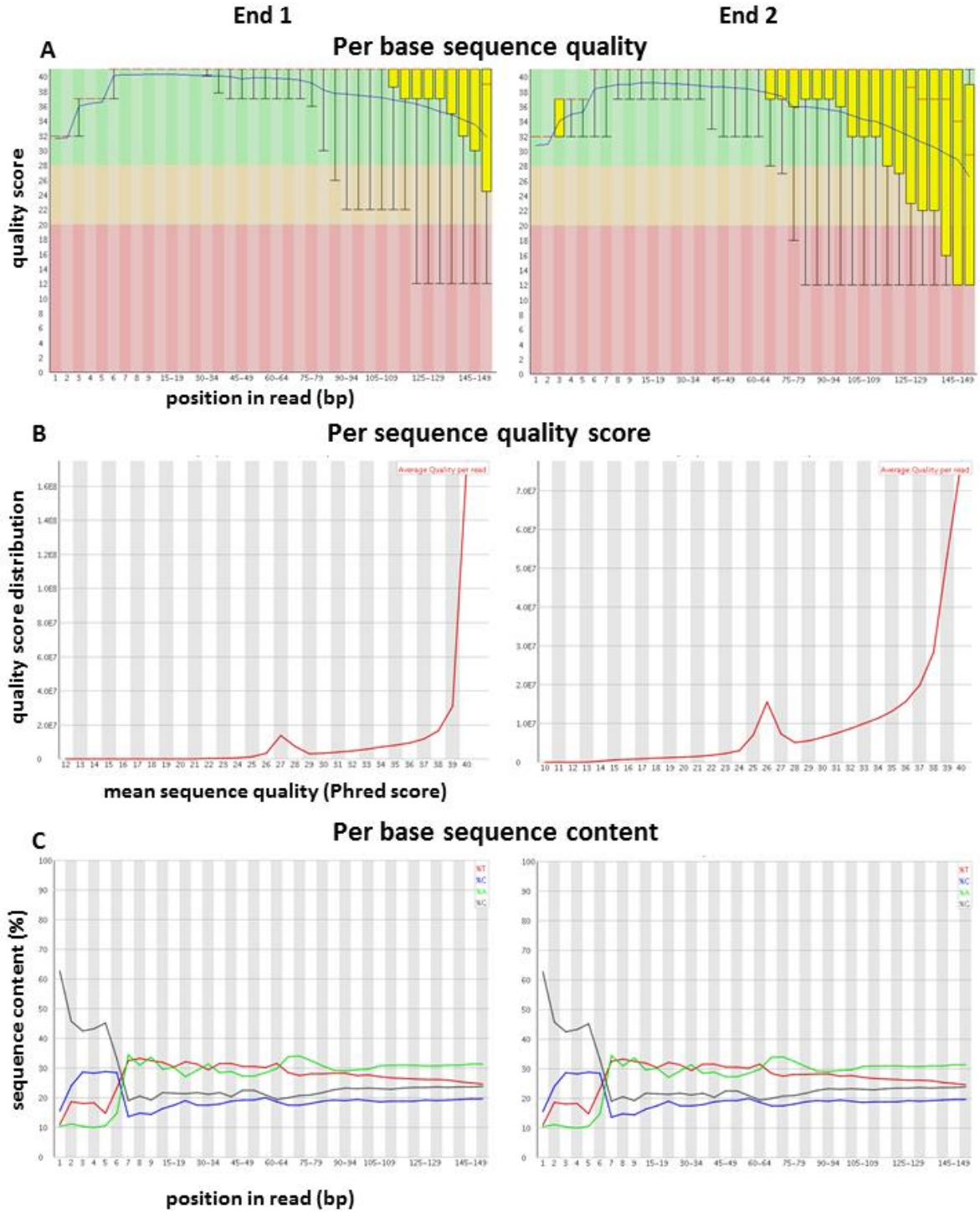
Appendix B

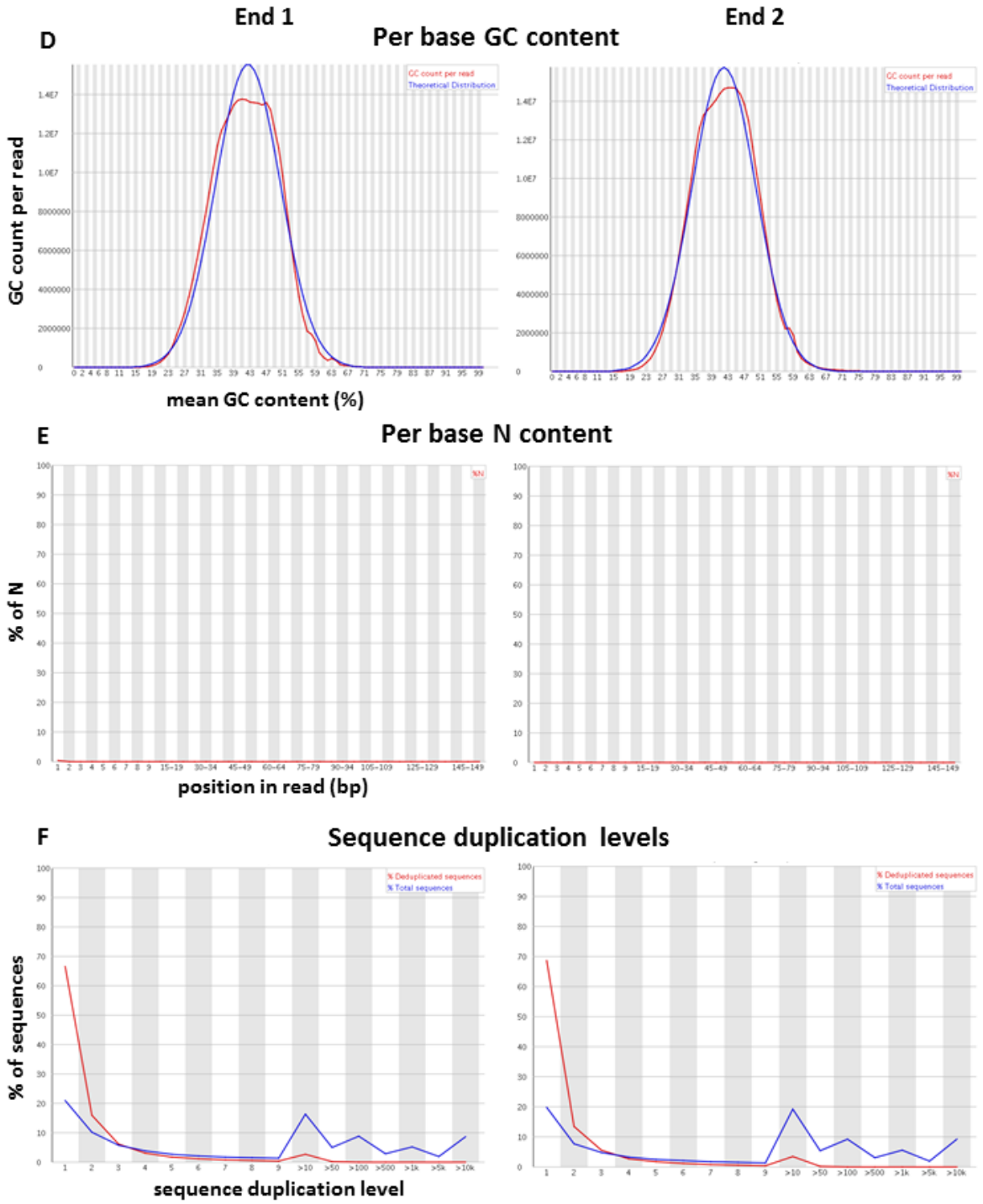
Electropherograms for the RNA extracted from the digestive glands, caecum and gills of three adult *L. pedicellatus.* Each electropherogram corresponds to one well in Figure 4.1 and it has been named accordingly.



Appendix C

FastQC quality check for the crystalline style RNA sequencing performed with the HiSeq3000 system. **A**. Per base sequence quality shows an overview of the quality values across all bases, highlighting in green very good quality calls, in orange reasonable ones and in red poor ones. **B.** Per sequence quality score displays subsets of sequences with low quality values (<27 is less than optimal). **C.** Per base sequence content highlights any imbalance in base content (failure occurs when difference between any bases is greater than 20%). **D**. Per base GC content shows how the GC content differs from a theoretical distribution, indicating a contaminated library or a biased subset (concerns are raised if 15% of the reads deviates from the normal distribution). **E.** Per base N content reports uncalled bases, with values >5% issuing a warning. **F.** Sequence duplication levels counts the degree of duplication for every sequence in the set, which should be below 20% for optimal quality.


End 1
End 2
A
Per base sequence quality
quality score
position in read (bp)
B
Per sequence quality score
quality score distribution
mean sequence quality (Phred score)
C
Per base sequence content
sequence content (%)
position in read (bp)


End 1
End 2
D
Per base GC content
GC count per read
Theoretical Distribution
GC count per read
mean GC content (%)
E
Per base N content
% of N
position in read (bp)
F
Sequence duplication levels
% Deduplicated sequences
% Total sequences
% of sequences
sequence duplication level

List of abbreviations and symbols

µM	microMolar
AA	Auxiliary Activities
ABSL	Acetyl Bromide Soluble Lignin
AI medium	Auto-Induction medium
AIR	Alcohol Insoluble Residue
Amp	Ampicillin
BLAST	Basic Local Alignment Search Tool
bp	base pairs
BSA	Bovine Serum Albumin
CAP	Catabolite Activator Protein binding site
CAZy	Carbohydrate-Active enZymes database
CAZyme	CAZy enzyme
CBM	Carbohydrate Binding Module
cDNA	complementary DeoxyriboNucleic Acid
CE	Carbohydrate Esterase
CMC	CarboxyMethyl Cellulose
CMV	CytoMegaloVirus
CPB	Citrate Phosphate Buffer
C-terminus	Carboxyl-group terminus
dbCAN	database for automated Carbohydrate-Active enzyme aNnotation
dH_20	deionised water
DMF	DiMethylFormamide
DMSO	DiMethyl SulfOxide
DNase	DeoxyriboNuclease
DNS	3,5-DiNitroSalicylic acid
dNTP	deoxyriboNucleotide TriPhosphate
dsDNA	double stranded DeoxyriboNucleic Acid
DTE	DiThioErythritol
DTT	DiThioThreitol
EDTA	EthyleneDiamineTetraacetic Acid
emPAI	exponentially modified Protein Abundance Index
ENA	European Nucleotide Archive

ESi-MS	ElectroSpray ionisation-Mass Spectrometry
EST	Expressed Sequence Tag
ExPASY	Expert Protein Analysis SYstem
FISH	Fluorescence *In Situ* Hybridisation
FLAG	FLAG-tag
FTIR	Fourier-Transform InfraRed spectroscopy
G	Guaiacyl unit
GH	Glycosyl Hydrolase
GHG	GreenHouse Gas
GST	Glutathione *S*-Transferases
H	p-Hydroxyphenyl unit
Halo	Haloalkane dehalogenase
HEK cells	Human Embryonic Kidney cells
HMDS	HexaMethylDiSilazane
HPAEC	High-Performance Anion-Exchange Chromatography
HRP	HorseRadish Peroxidase
HRV	Human RhinoVirus
IgG	Immunoglobulin G
IPTG	IsoPropyl β-D-1-ThioGalactopyranoside
ISP	Ion Sphere Particle
Kan	Kanamicin
kb	kilobase
kDa	kiloDalton
LB	Lysogeny Broth
LBG	Locust Bean Gum
LC-MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LPMO	Lytic Polysaccharide MonOxygenase
M	Molar
MACPF	Membrane Attack Complex/PerForin
MALDI/TOF-TOF	Matrix Assisted Laser Desorption Ionization/tandem Time Of Flight
MBP	Maltose Binding Protein
MCS	Multiple Cloning Site
mRNA	messanger RiboNucleic Acid

MW	Molecular Weight
NBUS	Natural Biomass Utilisation Systems
NCBI	National Center for Biotechnology Information
NCBI nrdb	National Center for Biotechnology Information non-redundant database
Neo	Neomicin
N-terminus	amine-group terminus
OD	Optical Density
oligo-DT	oligo-DeoxyThymine
OPPF-UK	Oxford Protein Production Facilty-UK
ORF	Open Reading Frame
PACE	Polysaccharide Analysis using CArbohydrate gel electrophoresis
PASC	Phosphoric Acid Swollen Cellulose
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PL	Polysaccharide Lyase
poly(A)	poly Adenosine
PSU	Practical Salinity Unit
qPCR	quantitative Polymerase Chain Reaction
RACE	Rapid Amplification of Complementary DNA Ends
RBS	Ribosomal Binding Site
RIN	RNA Integrity Number
RNase	RiboNuclease
RPK	Reads Per Kilobase
rpm	revolutions per minute
rRNA	ribosomal RiboNucleic Acid
S	Syringyl unit
SDS	Sodium Dodecyl Sulfate
SDS-PAGE	Sodium Dodecyl Sulphate PolyAcrylamide Gel Electrophoresis
SEM	Scanning Electron Microscoopy
SOC	Super Optimal broth with Catabolite repression
ssDNA	single stranded DeoxyriboNucleic Acid
ssNMR	Solid-State ^{13}C Nuclear Magnetic Resonance

SUMO	Small Ubiquitin-like Modifier
TBE	Tris-Borate-EDTA buffer
TEM	Transmission Electron Microscopy
TF	Trigger Factor
TFA	TriFluoroacetic Acid
T_m	melting Temperature
TPM	Transcripts Per kilobase Million
Tris	Tris(hydroxymethyl)aminomethane
TRX	ThioRedoXin
UV	UltraViolet
v/v	volume to volume
w/v	weight to volume
x-cellobiose	5-Bromo-4-chloro-3-indolyl β-D-cellobioside

References

Adman, E.T. 1991. Copper protein structures. *Advances in protein chemistry.* Elsevier.

Agger, J.W., Isaksen, T., Várnai, A., Vidal-Melgosa, S., Willats, W.G., Ludwig, R., Horn, S.J., Eijsink, V.G. and Westereng, B. 2014. Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation. *Proceedings of the National Academy of Sciences*, 201323629.

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

Alessi, A.M., Bird, S.M., Bennett, J.P., Oates, N.C., Li, Y., Dowle, A.A., Polikarpov, I., Young, J.P.W., McQueen-Mason, S.J. and Bruce, N.C. 2017. Revealing the insoluble metasecretome of lignocellulose-degrading microbial communities. *Scientific Reports,* 7, 2356.

Alexander, C.G., Cutler, R.L. and Yellowless, D. 1978. Studies on the composition and enzyme content of the crystalline style of *Telescopium telescopium* (L.) Gastropoda. *Comparative Biochemistry and Physiology,* 64b, 83-89.

Allam, B. and Raftos, D. 2015. Immune responses to infectious diseases in bivalves. *Journal of Invertebrate Pathology,* 131, 121-136.

Alyakrinskaya, I.O. 2001. The dimentions, characteristics and functions of the crystalline style of molluscs. *Biology Bulletin of the Academy of Sciences of the USSR,* 28, 523-535.

Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A. and Ouzounis, C. 1999. Automated genome sequence analysis and annotation. *Bioinformatics,* 15, 391-412.

Arantes, V. and Goodell, B. 2014. Current understanding of brown-rot fungal biodegradation mechanisms: a review. *Deterioration and protection of sustainable biomaterials.* ACS Publications.

Argos, P., Rossmann, M.G., Grau, U.M., Zuber, H., Frank, G. and Tratschin, J.D. 1979. Thermal stability and protein structure. *Biochemistry,* 18, 5698-5703.

Aspeborg, H., Coutinho, P.M., Wang, Y., Brumer, H. and Henrissat, B. 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evolutionary Biology,* 12, 186.

Ausec, L., Zakrzewski, M., Goesmann, A., Schluter, A. and Mandic-Mulec, I. 2011. Bioinformatic analysis reveals high diversity of bacterial genes for laccase-like enzymes. *PLoS ONE,* 6, e25724.

Bachere, E., Mialhe, E., Noel, D., Boulo, V., Morvan, A. and Rodriguez, J. 1995. Knowledge and research prospects in marine mollusc and crustacean immunology. *Aquaculture,* 132, 17-32.

Bailey, K. and Worboys, B.D. 1960. The lamellibranch crystalline style. *Biochemical Journal,* 76, 487.

Baldi, P. and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics,* 17, 509-519.

Bartsch, P. 1922. A monograph of the American shipworms. *Bulletin of the United states National Museum,* 122, 1-51.

Bartsch, P. 1927. New species of shipworms from Siam. *The Journal of the Siam Society. Natural History Supplement.,* 7, 59-63.

Baumann, P., Moran, N.A. and Baumann, L. 2006. Bacteriocyte-associated endosymbionts of insects. *In:* Martin Dworkin, Stanley Falkow, Eugene Rosenberg, Karl-Heinz Schleifer & Stackebrandt, E. (eds.) *The Prokaryotes. Volume 1: Symbiotic associations, Biotechnology, Applied Microbiology.* New York: Springer.

Bayer, E.A., Lamed, R., White, B.A. and Flint, H.J. 2008. From cellulosomes to cellulosomics. *The Chemical Record,* 8, 364-77.

Bazylinski, D. and Rosenberg, F. 1983. Occurrence of a brush-border in the cecum (appendix) of several *Teredo* and *Bankia* species (Teredinidae, bivalvia, mollusca). *Veliger,* 25, 251-&.

Becker, G. 1959. Biological investigations on marine borers in Berlin-Dahlem. *In:* Ray, D.L. (ed.) *Marine Boring and Fouling Organisms.* University of Washington Press, Seattle.

Bedford, J.J. and Reid, M.S. 1969. Gel electrophoresis of the proteins in the crystalline style of certain mollusca. *Comparative Biochemistry and Physiology,* 29, 659-664.

Beg, Q., Kapoor, M., Mahajan, L. and Hoondal, G. 2001. Microbial xylanases and their industrial applications: a review. *Applied Microbiology and Biotechnology,* 56, 326-338.

Bélaïch, J.P., Tardif, C., Bélaïch, A. and Gaudin, C. 1997. The cellulolytic system of *Clostridium cellulolyticum*. *Journal of Biotechnology,* 57, 3-14.

Bennett, B.A., Smith, C.R., Glaser, B. and Maybaum, H.L. 1994. Faunal community structure of a chemoautotrophic assemblage on whale bones in the deep northeast Pacific Ocean. *Marine Ecology-Progress Series,* 108, 205-205.

Berger, I., Fitzgerald, D.J. and Richmond, T.J. 2004. Baculovirus expression system for heterologous multiprotein complexes. *Nature Biotechnology,* 22, 1583.

Besser, K., Malyon, G.P., Eborall, W.S., Paro da Cunha, G., Goncalves Filgueiras, J., Dowle, A., Cruz Garcia, L., Page, S.J., Dupree, R. and Kern, M.F. 2018. Hemocyanin facilitates lignocellulose digestion by wood-boring marine crustaceans. *Nature Communications,* 9, 5125.

Betcher, M.A., Fung, J.M., Han, A.W., O'Connor, R., Seronay, R., Concepcion, G.P., Distel, D.L. and Haygood, M.G. 2012. Microbial distribution and abundance in the digestive system of five shipworm species (Bivalvia: Teredinidae). *PLoS ONE,* 7, e45309.

Bienhold, C., Ristova, P.P., Wenzhöfer, F., Dittmar, T. and Boetius, A. 2013. How deep-sea wood falls sustain chemosynthetic life. *PLoS ONE,* 8, e53590.

Bignell, D.E. and Eggleton, P. 2000. Termites in ecosystems. *Termites: evolution, sociality, symbioses, ecology.* Springer.

Bird, L.E. 2011. High throughput construction and small scale expression screening of multi-tag vectors in Escherichia coli. *Methods,* 55, 29-37.

Blackman, L.M., Cullerne, D.P. and Hardham, A.R. 2014. Bioinformatic characterisation of genes encoding cell wall degrading enzymes in the *Phytophthora* parasitica genome. *BMC Genomics,* 15, 785.

Blifernez-Klassen, O., Klassen, V., Doebbe, A., Kersting, K., Grimm, P., Wobbe, L. and Kruse, O. 2012. Cellulose degradation and assimilation by the unicellular phototrophic eukaryote *Chlamydomonas reinhardtii*. *Nature Communications,* 3, 1214.

Boerjan, W., Ralph, J. and Baucher, M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology,* 54, 519-46.

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30, 2114-2120.

Borges, L. 2007. Biogeography of wood boring organisms in European Coastal waters and new approaches to controlling borer attack. *Portsmouth University, Portsmouth, United Kingdom (PhD thesis)*.

Borges, L.M., Cragg, S.M., Bergot, J., Williams, J.R., Shayler, B. and Sawyer, G.S. 2008. Laboratory screening of tropical hardwoods for natural resistance to the marine borer *Limnoria quadripunctata*: the role of leachable and non-leachable factors. *Holzforschung,* 62, 99-111.

Borges, L.M. and Merckelbach, L.M. 2018. *Lyrodus mersinensis* sp. nov.(Bivalvia: Teredinidae) another cryptic species in the Lyrodus pedicellatus (Quatrefages, 1849) complex. *Zootaxa,* 4442, 441-457.

Borges, L.M.S., Merckelbach, L.M., Sampaio, I. and Cragg, S.M. 2014. Diversity, environmental requirements, and biogeography of bivalve wood-borers (Teredinidae) in European coastal waters. *Frontiers in Zoology,* 11, 13.

Borges, L.M.S., Sivrikaya, H., le Roux, A., Shipway, J.R., Cragg, S.M. and Costa, F.O. 2012. Investigating the taxonomy and systematics of marine wood borers (Bivalvia: Teredinidae) combining evidence from morphology, DNA barcodes and nuclear locus sequences. *Invertebrate Systematics,* 26, 572.

Bork, P. and Bairoch, A. 1996. Go hunting in sequence databases but watch out for the traps. *Trends in Genetics,* 12, 425-427.

Boucias, D.G., Cai, Y., Sun, Y., Lietze, V.U., Sen, R., Raychoudhury, R. and Scharf, M.E. 2013. The hindgut lumen prokaryotic microbiota of the termite *Reticulitermes flavipes* and its responses to dietary lignocellulose composition. *Molecular Ecology,* 22, 1836-1853.

Boudet, A.M., Lapierre, C. and Grima-Pettenati, J. 1995. Biochemistry and molecular biology of lignification. *New Phytologist,* 129, 203-236.

Bradford, M.M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry,* 72, 248-254.

Brito, T.L., Campos, A.B., von Meijenfeldt, F.B., Daniel, J.P., Ribeiro, G.B., Silva, G.G.Z., Morais, D., Wilke, D.V., Dutilh, B.E. and Meirelles, P.M. 2018. The gill-associated symbiont microbiome is a main source of woody-plant polysaccharide hydrolase genes and secondary metabolite gene clusters in *Neoteredo reynei*, a unique shipworm from south Atlantic mangroves. *BioRxiv*, 357616.

Brock, V.K., V.S. and Kennedy, V.S. 1992. Quantitative analysis of crystalline style carbohydrases in five suspension and deposit feeding bivalves. *Journal of Experimental Marine Biology and Ecology,* 159, 51-58.

Bromley, J.R., Busse-Wicher, M., Tryfona, T., Mortimer, J.C., Zhang, Z., Brown, D.M. and Dupree, P. 2013. GUX1 and GUX2 glucuronyltransferases decorate distinct domains of glucuronoxylan with different substitution patterns. *The Plant Journal,* 74, 423-434.

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

Buckeridge, M.S. 2010. Seed cell wall storage polysaccharides: models to understand cell wall biosynthesis and degradation. *Plant Physiology,* 154, 1017-23.

Bugg, T.D., Ahmad, M., Hardiman, E.M. and Rahmanpour, R. 2011. Pathways for degradation of lignin in bacteria and fungi. *Natural Product Reports,* 28, 1883-96.

Bui, T.H. and Lee, S.Y. 2015. Endogenous cellulase production in the leaf litter foraging mangrove crab *Parasesarma erythodactyla*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology,* 179, 27-36.

Bull, A.T. and Chesters, C. 1966. The biochemistry of laminarin and the nature of laminarinase. *Advances in Enzymology and Related Areas of Molecular Biology,* 28, 325-364.

Busk, P.K. and Lange, L. 2015. Classification of fungal and bacterial lytic polysaccharide monooxygenases. *BMC Genomics,* 16, 368.

Caffall, K.H. and Mohnen, D. 2009. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrare Research,* 344, 1879-900.

Calloway, C.B. 1982. Parentally derived extra-embryonic nutrition in *Lyrodus pedicellatus* (Bivalvia, Teredinidae). *American Zoologist,* 22, 860-860.

Calloway, C.B. and Turner, R.D. 1988. Brooding in the Teredinidae (Mollusca: Bivalvia). *Marine biodeterioration: advanced techniques applicable to the Indian Ocean.* Rotterdam: AA Balkema.

Calman, W.T. 1919. *Marine boring animals injurious to submerged structures.*, British Museum Natural History Economy Series.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. 2008. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research,* 37, D233-D238.

Carpenter, E.J. and Culliney, J.L. 1975. Nitrogen fixation in marine shipworms. *Science,* 187, 551-2.

Carroll, A. and Somerville, C. 2009. Cellulosic biofuels. *Annual Review of Plant Biology,* 60, 165-82.

Cassab, G. 1998. Plant cell wall proteins. *Annual Review of Plant Physiology and Plant Molecular Biology,* 49, 281-309.

Charles, F., Sauriau, P.-G., Aubert, F., Lebreton, B., Lantoine, F. and Riera, P. 2018. Sources partitioning in the diet of the shipworm *Bankia carinata* (Gray, 1827): An experimental study based on stable isotopes. *Marine Environmental Research,* 142, 208-213.

Chauhan, P.S., Puri, N., Sharma, P. and Gupta, N. 2012. Mannanases: microbial sources, production, properties and potential biotechnological applications. *Applied Microbiology and Biotechnology,* 93, 1817-1830.

Chen, F. and Dixon, R.A. 2007. Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology,* 25, 759.

Ciucanu, I. and Kerek, F. 1984. A simple and rapid method for the permethylation of carbohydrates. *Carbohydrate Research,* 131, 209-217.

Cobb, K. 2002. Return of a castaway. The gripping story of a boring clam. *Science News,* 162, 72-74.

Coleman, G. 1978. The metabolism of cellulose, glucose and starch by the rumen ciliate protozoon *Eudiplodinium maggii*. *Microbiology,* 107, 359-366.

Collins, T., Gerday, C. and Feller, G. 2005. Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiology Reviews,* 29, 3-23.

Coughlan, J. 1977. Marine borers in Southampton water. *Proceedings of the Hampshire Field Club & Archaeological Society,* 33, 5-15.

Coupin, H. 1900. Sur les fonctions de la tige cristalline des Acephales. *Comptes Rendus de l'Académie des Sciences,* 130, 1214.

Couturier, M., Ladevèze, S., Sulzenbacher, G., Ciano, L., Fanuel, M., Moreau, C., Villares, A., Cathala, B., Chaspoul, F. and Frandsen, K.E. 2018. Lytic xylan oxidases from wood-decay fungi unlock biomass degradation. *Nature Chemical Biology,* 14, 306.

Cowling, E.B. and Merrill, W. 1966. Nitrogen in wood and its role in wood deterioration. *Canadian Journal of Botany,* 44, 1539-1554.

Coy, M., Salem, T., Denton, J., Kovaleva, E., Liu, Z., Barber, D., Campbell, J., Davis, D., Buchman, G. and Boucias, D. 2010. Phenol-oxidizing laccases from the termite gut. *Insect Biochemistry and Molecular Biology,* 40, 723-732.

Cragg, S. 2008. Marine wood boring invertebrates of New Guinea and its surrounding waters. *The Ecology of Papua, Part 1.* Periplus.

Cragg, S.M., Beckham, G.T., Bruce, N.C., Bugg, T.D., Distel, D.L., Dupree, P., Etxabe, A.G., Goodell, B.S., Jellison, J., McGeehan, J.E.*, et al.* 2015. Lignocellulose degradation mechanisms across the Tree of Life. *Current Opinion in Chemical Biology,* 29, 108-19.

Cragg, S.M., Jumel, M.C., Al-Horani, F.A. and Hendy, I.W. 2009. The life history characteristics of the wood-boring bivalve *Teredo bartschi* are suited to the elevated salinity, oligotrophic circulation in the Gulf of Aqaba, Red Sea. *Journal of Experimental Marine Biology and Ecology,* 375, 99-105.

Dahiya, N., Tewari, R. and Hoondal, G.S. 2006. Biotechnological aspects of chitinolytic enzymes: a review. *Applied Microbiology and Biotechnology,* 71, 773-782.

Davies, P., Morvan, C., Sire, O. and Baley, C. 2007. Structure and properties of fibres from sea-grass (*Zostera marina*). *Journal of Materials Science,* 42, 4850-4857.

de Gannes, V., Eudoxie, G. and Hickey, W.J. 2013. Prokaryotic successions and diversity in composts as revealed by 454-pyrosequencing. *Bioresource Technology,* 133, 573-580.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. 2009. Global signatures of protein and mRNA expression levels. *Molecular BioSystems,* 5, 1512-1526.

DeCicco, J.M., Liu, D.Y., Heo, J., Krishnan, R., Kurthen, A. and Wang, L. 2016. Carbon balance effects of US biofuel production and use. *Climatic Change,* 138, 667-680.

Delgado-García, M., Valdivia-Urdiales, B., Aguilar-González, C.N., Contreras-Esquivel, J.C. and Rodríguez-Herrera, R. 2012. Halophilic hydrolases as a new tool for the biotechnological industries. *Journal of the Science of Food and Agriculture,* 92, 2575-2580.

Deshayes, G.P. 1848. Histoire naturelle des Mollusques. i.Mollusques acephales. *Exploration scientifique de l'Algerie, 1840-2.* Paris.

Dimarogona, M., Topakas, E. and Christakopoulos, P. 2012. Cellulose degradation by oxidative enzymes. *Computational and Structural Biotechnology Journal,* 2, e201209015.

Distel, D.L. 2003. The biology of marine wood boring bivalves and their bacterial endosymbionts. *Wood Deterioration and Preservation,* 845, 253-271.

Distel, D.L., Altamia, M.A., Lin, Z., Shipway, J.R., Han, A., Forteza, I., Antemano, R., Limbaco, M., Tebo, A.G., Dechavez, R.*, et al.* 2017. Discovery of chemoautotrophic symbiosis in the giant

shipworm *Kuphus polythalamia* (Bivalvia: Teredinidae) extends wooden-steps theory. *Proceedings of the National Academy of Sciences of the United States of America*.

Distel, D.L., Amin, M., Burgoyne, A., Linton, E., Mamangkey, G., Morrill, W., Nove, J., Wood, N. and Yang, J. 2011. Molecular phylogeny of Pholadoidea Lamarck, 1809 supports a single origin for xylotrophy (wood feeding) and xylotrophic bacterial endosymbiosis in Bivalvia. *Molecular Phylogenetics and Evolution,* 61, 245-54.

Distel, D.L., Baco, A.R., Chuang, E., Morrill, W., Cavanaugh, C. and Smith, C.R. 2000. Marine ecology: do mussels take wooden steps to deep-sea vents? *Nature,* 403, 725-726.

Distel, D.L., Beaudoin, D.J. and Morrill, W. 2002a. Coexistence of multiple proteobacterial endosymbionts in the gills of the wood-boring bivalve *Lyrodus pedicellatus* (Bivalvia : Teredinidae). *Applied and Environmental Microbiology,* 68, 6292-6299.

Distel, D.L., DeLong, E.F. and Waterbury, J.B. 1991. Phylogenetic characterization and in situ localization of the bacterial symbiont of shipworms (Teredinidae: Bivalvia) by using 16S rRNA sequence analysis and oligodeoxynucleotide probe hybridization. *Applied and Environmental Microbiology,* 57, 2376-82.

Distel, D.L., Morrill, W., MacLaren-Toussaint, N., Franks, D. and Waterbury, J. 2002b. *Teredinibacter turnerae* gen. nov., sp. nov., a dinitrogen-fixing, cellulolytic, endosymbiotic gamma-proteobacterium isolated from the gills of wood-boring molluscs (Bivalvia: Teredinidae). *International Journal of Systematic and Evolutionary Microbiology,* 52, 2261-9.

Distel, D.L. and Roberts, S.J. 1997. Bacterial endosymbionts in the gills of the deep sea woodboring bivalves *Xylophaga atlantica* and *Xylophaga washingtona*. *Biological Bulletin (Woods Hole),* 192, 253-261.

Dodd, D. and Cann, I.K. 2009. Enzymatic deconstruction of xylan for biofuel production. *Global change biology. Bioenergy,* 1, 2-17.

Dodd, D., Moon, Y.H., Swaminathan, K., Mackie, R.I. and Cann, I.K. 2010. Transcriptomic analyses of xylan degradation by *Prevotella bryantii* and insights into energy acquisition by xylanolytic bacteroidetes. *Journal of Biological Chemistry,* 285, 30261-73.

Doerks, T., Bairoch, A. and Bork, P. 1998. Protein annotation: detective work for function prediction. *Trends in Genetics,* 14, 248-250.

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

Domon, B. and Costello, C.E. 1988. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate Journal,* 5, 397-409.

Dore, W.H. and Miller, R.C. 1923. The digestion of wood by *Teredo navalis*. *University of California Publications in Zoology,* 22, 383-400.

E.U. 2007. Brussels European Council, 8/9 March 2007. Presidency Conclusions. Brusells: Council of the European Union.

Eborall, W.S. 2013. *Discovering novel lignocellulose degrading enzymes from the marine wood borer, Limnoria quadripunctata.* Doctor of Philosophy, University of York.

Eckelbarger, K.J. and Reish, D.J. 1972. Effects of varying temperature and salinities on settlement, growth, and reproduction of the wood-boring pelecypod, *Lyrodus pedicellatus*. *Bulletin of the Southern California Academy of Sciences,* 71, 116-127.

Edmondson, C.H. 1920. The reformation of the crystalline style in *Mya arenaria* after extraction. *Journal of Experimental Zoology,* 30, 259-291.

Eichinger, L., Pachebat, J., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K. and Xu, Q. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature,* 435, 43.

Ekborg, N.A., Morrill, W., Burgoyne, A.M., Li, L. and Distel, D.L. 2007. CelAB, a multifunctional cellulase encoded by *Teredinibacter turnerae* T7902T, a culturable symbiont isolated from the wood-boring marine bivalve *Lyrodus pedicellatus*. *Applied and Environmental Microbiology,* 73, 7785-8.

Ekenler, M. and Tabatabai, M. 2004. β-Glucosaminidase activity as an index of nitrogen mineralization in soils. *Communications in Soil Science and Plant Analysis,* 35, 1081-1094.

Elshahawi, S.I., Trindade-Silva, A.E., Hanora, A., Han, A.W., Flores, M.S., Vizzoni, V., Schrago, C.G., Soares, C.A., Concepcion, G.P., Distel, D.L*., et al.* 2013. Boronated tartrolon antibiotic produced by symbiotic cellulose-degrading bacteria in shipworm gills. *Proceedings of the National Academy of Sciences of the United States of America,* 110, E295-304.

Eriksson, K.-E.L., Blanchette, R.A. and Ander, P. 1990. Biodegradation of cellulose. *Microbial and enzymatic degradation of wood and wood components.* Berlin, Heidelberg: Springer.

Eriksson, T., Karlsson, J. and Tjerneld, F. 2002. A model explaining declining rate in hydrolysis of lignocellulose substrates with cellobiohydrolase I (Cel7A) and endoglucanase I (Cel7B) of *Trichoderma reesei*. *Applied Biochemistry and Biotechnology,* 101, 41-60.

Ertan, H., Siddiqui, K.S., Muenchhoff, J., Charlton, T. and Cavicchioli, R. 2012. Kinetic and thermodynamic characterization of the functional properties of a hybrid versatile peroxidase using isothermal titration calorimetry: insight into manganese peroxidase activation and lignin peroxidase inhibition. *Biochimie,* 94, 1221-1231.

Feeney, R.E. and Osuga, D.T. 1976. Comparative biochemistry of Antarctic proteins. *Comparative Biochemistry and Physiology Part A: Physiology,* 54, 281-286.

Fengel, D. and Wegener, G. 1989. *Wood: chemistry, ultrastructure, reactions,* Berlin, Walter de Gruyter.

Fernandes, A.N., Thomas, L.H., Altaner, C.M., Callow, P., Forsyth, V.T., Apperley, D.C., Kennedy, C.J. and Jarvis, M.C. 2011. Nanostructure of cellulose microfibrils in spruce wood. *Proceedings of the National Academy of Sciences of the United States of America,* 108, E1195-203.

Fingerman, K.R., Torn, M.S., O'Hare, M.H. and Kammen, D.M. 2010. Accounting for the water impacts of ethanol production. *Environmental Research Letters,* 5, 014020.

Flecha, S., Pérez, F.F., García-Lafuente, J., Sammartino, S., Ríos, A.F. and Huertas, I.E. 2015. Trends of pH decrease in the Mediterranean Sea through high frequency observational data: indication of ocean acidification in the basin. *Scientific Reports,* 5, 16770.

Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R.A., Henrissat, B., Martínez, A.T., Otillar, R., Spatafora, J.W. and Yadav, J.S. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science,* 336, 1715-1719.

Forsberg, Z., Vaaje-Kolstad, G., Westereng, B., Bunæs, A.C., Stenstrøm, Y., MacKenzie, A., Sørlie, M., Horn, S.J. and Eijsink, V.G. 2011. Cleavage of cellulose by a CBM33 protein. *Protein Science,* 20, 1479-1483.

Foster, C.E., Martin, T.M. and Pauly, M. 2010. Comprehensive compositional analysis of plant cell walls (lignocellulosic biomass) part II: carbohydrates. *Journal of Visualized Experiments*, e1745.

Fukushima, R.S. and Hatfield, R.D. 2001. Extraction and isolation of lignin for utilization as a standard to determine lignin concentration using the acetyl bromide spectrophotometric method. *Journal of Agricultural and Food Chemistry,* 49, 3133-3139.

Gabaldón, T. and Huynen, M.A. 2004. Prediction of protein function and pathways in the genome era. *Cellular and Molecular Life Sciences CMLS,* 61, 930-944.

Galbe, M. and Zacchi, G. 2007. Pretreatment of lignocellulosic materials for efficient bioethanol production. *Biofuels.* Springer.

Gallager, S.M., Turner, R.D. and Berg, C.J. 1981. Physiological aspects of wood consumption, growth, and reproduction in the shipworm *Lyrodus pedicellatus* Quatrefages (Bivalvia, Teredinidae). *Journal of Experimental Marine Biology and Ecology,* 52, 63-77.

Geib, S.M., Filley, T.R., Hatcher, P.G., Hoover, K., Carlson, J.E., Jimenez-Gasco Mdel, M., Nakagawa-Izumi, A., Sleighter, R.L. and Tien, M. 2008. Lignin degradation in wood-feeding insects. *Proceedings of the National Academy of Sciences of the United States of America,* 105, 12932-7.

George, W.C. 1952. The digestion and absorption of fat in lamellibranchs. *The Biological Bulletin,* 102, 118-127.

Gerdol, M. and Venier, P. 2015. An updated molecular basis for mussel immunity. *Fish & Shellfish Immunology,* 46, 17-38.

Gilbertson, R.L. 1980. Wood-rotting fungi of North America. *Mycologia,* 72, 1-49.

Gírio, F.M., Fonseca, C., Carvalheiro, F., Duarte, L.C., Marques, S. and Bogel-Łukasik, R. 2010. Hemicelluloses for fuel ethanol: a review. *Bioresource Technology,* 101, 4775-4800.

Gomes, A.R., Byregowda, S.M., Veeregowda, B.M. and Balamurugan, V. 2016. An overview of heterologous expression host systems for the production of recombinant proteins. *Advances in Animal and Veterinary Sciences,* 4, 346-356.

Gomez, L.D., Steele-King, C.G. and McQueen-Mason, S.J. 2008. Sustainable liquid biofuels from biomass: the writing's on the walls. *New Phytologist,* 178, 473-85.

Gosling, E. 2015. *Marine bivalve molluscs,* Oxford, John Wiley & Sons.

Goubet, F., Jackson, P., Deery, M.J. and Dupree, P. 2002. Polysaccharide analysis using carbohydrate gel electrophoresis: a method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Analytical Biochemistry,* 300, 53-68.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R. and Zeng, Q. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology,* 29, 644.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., Knapp, S., Oppermann, U., Arrowsmith, C. and Hui, R. 2008. Protein production and purification. *Nature Methods,* 5, 135.

Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology,* 4, 117.

Greene, R.V., Griffin, H.L. and Freer, S.N. 1988. Purification and characterization of an extracellular endoglucanase from the marine shipworm bacterium. *Archives of Biochemistry and Biophysics,* 267, 334-341.

Greenfield, L.J. and Lane, C.E. 1953. Cellulose digestion in *Teredo*. *Journal of Biological Chemistry,* 204, 669-672.

Guerriero, G., Hausman, J.F., Strauss, J., Ertan, H. and Siddiqui, K.S. 2015. Destructuring plant biomass: focus on fungal and extremophilic cell wall hydrolases. *Plant Science,* 234, 180-93.

Gustafsson, C., Govindarajan, S. and Minshull, J. 2004. Codon bias and heterologous protein expression. *Trends in Biotechnology,* 22, 346-353.

Hameed, P.S. and Paulpandian, A.L. 1987. Quantitative-analysis of carbohydrases in the crystalline style of some intertidal bivalve mollusks. *Proceedings of the Indian Academy of Sciences-Animal Sciences,* 96, 41-47.

Han, A.W., Sandy, M., Fishman, B., Trindade-Silva, A.E., Soares, C.A., Distel, D.L., Butler, A. and Haygood, M.G. 2013. Turnerbactin, a novel triscatecholate siderophore from the shipworm endosymbiont *Teredinibacter turnerae* T7901. *PLoS ONE,* 8, e76151.

Hanquet, A.-C., Jouaux, A., Heude, C., Mathieu, M. and Kellner, K. 2011. A sodium glucose co-transporter (SGLT) for glucose transport into *Crassostrea gigas* vesicular cells: Impact of alimentation on its expression. *Aquaculture,* 313, 123-128.

Hansen, J., Ruedy, R., Sato, M. and Lo, K. 2010. Global surface temperature change. *Reviews of Geophysics,* 48.

Harington, C.R. 1921. A note on the physiology of the ship-worm (*Teredo norvegica*). *Biochemical Journal,* 15, 736-41.

Harris, D. and DeBolt, S. 2010. Synthesis, regulation and utilization of lignocellulosic biomass. *Plant Biotechnology Journal,* 8, 244-62.

Hashimoto, H. 2006. Recent structural studies of carbohydrate-binding modules. *Cellular and Molecular Life Sciences,* 63, 2954-2967.

Hatahet, F., Boyd, D. and Beckwith, J. 2014. Disulfide bond formation in prokaryotes: history, diversity and design. *Biochimica et Biophysica Acta,* 1844, 1402-14.

Hatai, K.M. 1951. A lower Cretaceous *Teredo*. *Institute of Geology and Paleontology of Sendai,* 3, 29-32.

Hatakka, A. and Hammel, K.E. 2011. Fungal biodegradation of lignocelluloses. *Industrial applications.* Springer.

Haygood, M., Altamia, M., Elshahawi, S., Han, A., Lin, Z., Betcher, M., O'Connor, R., Sandy, M., Trinidade-Silva, A. and Butler, A. 2015. Versatile bacterial symbionts of shipworms contribute to wood digestion, fix nitrogen and produce secondary metabolites. *Planta Medica,* 81, IL43.

Heide, A.d. 1892. *De 1686 Experimenta circa sanguinis missionem, fibras motrices, urticam marinam, etc.—accedunt ejusdem auctoris observationes medicoe nec non anatome Mytuli. 35–36*, Editio Nova, Amsterdam.

Hemsworth, G.R., Johnston, E.M., Davies, G.J. and Walton, P.H. 2015. Lytic polysaccharide monooxygenases in biomass conversion. *Trends in Biotechnology,* 33, 747-61.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal,* 280, 309-316.

Heredia, A., Jimonez, A. and Guillon, R. 1995. Composition of plant cell wall. *Z Lebensm Unters Forsch,* 200, 24-31.

Herlet, J., Kornberger, P., Roessler, B., Glanz, J., Schwarz, W., Liebl, W. and Zverlov, V. 2017. A new method to evaluate temperature vs. pH activity profiles for biotechnological relevant enzymes. *Biotechnology for Biofuels,* 10, 234.

Herts, W. 1983. Arrays of plasma-membrane "rosettes" involved in cellulose microfibril formation of *Spirogyra*. *Planta,* 159, 347-356.

Higuchi, T. 2004. Microbial degradation of ligning: role of lignin peroxidase, manganese peroxidase, and laccase. *The Proceedings of the Japan Academy,* 80, 204-214.

Hill, J., Nelson, E., Tilman, D., Polasky, S. and Tiffany, D. 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences,* 103, 11206-11210.

Hochachka, P.W. and Somero, G.N. 1973. *Strategies of biochemical adaptation,* Philadelphia and London, W. B. Sauders Company.

Honein, K., Kaneko, G., Katsuyama, I., Matsumoto, M., Kawashima, Y., Yamada, M. and Watabe, S. 2012. Studies on the cellulose-degrading system in a shipworm and its potential applications. *Energy Procedia,* 18, 1271-1274.

Horiuchi, S. and Lane, E.C. 1966. Carbohydrases of the crystalline style and hepatopancreas of *Strombus gigas* Linné. *Comparative Biochemistry and Physiology,* 17, 1189–1197.

Horwitz, J.P., Chua, J., Curby, R.J., Tomson, A.J., Da Rooge, M.A., Fisher, B.E., Mauricio, J. and Klundt, I. 1964. Substrates for cytochemical demonstration of enzyme activity. I. Some substituted 3-indolyl-β-D-glycopyranosides1a. *Journal of Medicinal Chemistry,* 7, 574-575.

Huang, G., Huang, S., Yan, X., Yang, P., Li, J., Xu, W., Zhang, L., Wang, R., Yu, Y. and Yuan, S. 2014. Two apextrin-like proteins mediate extracellular and intracellular bacterial recognition in amphioxus. *Proceedings of the National Academy of Sciences,* 111, 13469-13474.

Hungate, R. 1940. Nitrogen content of sound and decayed coniferous woods and its relation to loss in weight during decay. *Botanical Gazette,* 102, 382-392.

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering,* 9, 90-95.

I.E.A. 2008. 1st-to 2nd-generation biofuel technologies. An overview of current industry and RD&D activities. . Paris: International Energy Agency (IEA).

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M. 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics,* 4, 1265-1272.

Ito, Y., Yoshikawa, A., Hotani, T., Fukuda, S., Sugimura, K. and Imoto, T. 1999. Amino acid sequences of lysozymes newly purified from invertebrates imply wide distribution of a novel class in the lysozyme family. *The FEBS Journal,* 259, 456-461.

Jaenicke, R. 1991. Protein stability and molecular adaptation to extreme conditions. *European Journal of Biochemistry,* 202, 715-728.

Jaenicke, R. and Zavodszky, P. 1990. Proteins under extreme physical conditions. *FEBS Letters,* 268, 344-9.

Jenner, H., Rajagopal, S., Van der Velde, G. and Daud, M. 2003. Perforation of ABS pipes by boring bivalve *Martesia striata*: a case study. *International Biodeterioration & Biodegradation,* 52, 229-232.

Jensen, K.A., Houtman, C.J., Ryan, Z.C. and Hammel, K.E. 2001. Pathways for extracellular Fenton chemistry in the brown rot basidiomycete *Gloeophyllum trabeum*. *Applied and Environmental Microbiology,* 67, 2705-2711.

Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D'Andrea, D., Lepore, R., Funk, C.S., Kahanda, I., Verspoor, K.M. and Ben-Hur, A. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology,* 17, 184.

Jin, Y., Petricevic, M., John, A., Raich, L., Jenkins, H., Portela De Souza, L., Cuskin, F., Gilbert, H.J., Rovira, C. and Goddard-Borger, E.D. 2016. A β-mannanase with a lysozyme-like fold and a novel molecular catalytic mechanism. *ACS Central Science,* 2, 896-903.

Jing, D. 2010. Improving the simultaneous production of laccase and lignin peroxidase from *Streptomyces lavendulae* by medium optimization. *Bioresource Technology,* 101, 7592-7.

Jing, D. and Wang, J. 2012. Controlling the simultaneous production of laccase and lignin peroxidase from *Streptomyces cinnamomensis* by medium formulation. *Biotechnology for Biofuels,* 5, 15.

Johnson, R.A., Iredale, T. and McNeill, F.A. 1936. *Destruction of timber by marine organisms in the port of Sydney. Supplementary report no. 1*, Maritime Services Board of New South Wales.

Jones, E., Oliphant, T. and Peterson, P. 2014. *SciPy: open source scientific tools for Python* [Online]. Available: http://www.scipy.org/ [Accessed August 2018].

Jones, E.G., Turner, R., Furtado, S. and Kuhne, H. 1976. Marine biodeteriogenic organisms. I: Lignicolous fungi and bacteria and wood boring mollusca and crusctacea. *International Biodeterioration Bulletin,* 12, 120-34.

Joost, H.-G. and Thorens, B. 2001. The extended GLUT-family of sugar/polyol transport facilitators: nomenclature, sequence characteristics, and potential function of its novel members. *Molecular Membrane Biology,* 18, 247-256.

Jørgensen, H., Kristensen, J.B. and Felby, C. 2007. Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuels, Bioproducts and Biorefining,* 1, 119-134.

Judd, W. 1979. The secretions and fine structure of bivalve crystalline style sacs. *Ophelia,* 18, 205-233.

Judd, W. 1987. Crystalline style proteins from bivalve mollusks. *Comparative Biochemistry and Physiology, B. Biochemistry & Molecular Biology,* 88, 333-339.

Kamphuis, J.B., Mercier-Bonin, M., Eutamène, H. and Theodorou, V. 2017. Mucus organisation is shaped by colonic content; a new view. *Scientific Reports,* 7, 8527.

Karande, A., Balasubramanyan, K. and Prema, S. 1968. Development of a laboratory method for bioassay of candidate toxins against teredid wood-borers. *Proceedings of the Symposium on Mollusca, Marine Biological Association of India*.

Kärkäs, M.D., Matsuura, B.S., Monos, T.M., Magallanes, G. and Stephenson, C.R. 2016. Transition-metal catalyzed valorization of lignin: the key to a sustainable carbon-neutral future. *Organic & biomolecular Chemistry,* 14, 1853-1914.

Kataoka, M. and Ishikawa, K. 2014. A new crystal form of a hyperthermophilic endocellulase. *Acta Crystallographica Section F,* 70, 878-883.

Kawabata, S.-i., Nagayama, R., Hirata, M., Shigenaga, T., Agarwala, K.L., Saito, T., Cho, J., Nakajima, H., Takagi, T. and Iwanaga, S. 1996. Tachycitin, a small granular component in horseshoe crab hemocytes, is an antimicrobial protein with chitin-binding activity. *The Journal of Biochemistry,* 120, 1253-1260.

Kenk, V.C. 1985. A new mussel (Bivalvia, Mytilidae) from hydrothermal vents in the Galapagos Rift zone. *Malacologia,* 26, 253-271.

Kern, M., McGeehan, J.E., Streeter, S.D., Martin, R.N., Besser, K., Elias, L., Eborall, W., Malyon, G.P., Payne, C.M., Himmel, M.E.*, et al.* 2013. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proceedings of the National Academy of Sciences of the United States of America,* 110, 10189-94.

Khare, S.K., Pandey, A. and Larroche, C. 2015. Current perspectives in enzymatic saccharification of lignocellulosic biomass. *Biochemical Engineering Journal,* 102, 38-44.

King, A.J., Cragg, S.M., Li, Y., Dymond, J., Guille, M.J., Bowles, D.J., Bruce, N.C., Graham, I.A. and McQueen-Mason, S.J. 2010. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proceedings of the National Academy of Sciences of the United States of America,* 107, 5345-50.

Kiuchi, I., Moriya, S. and Kudo, T. 2004. Two different size-distributions of engulfment-related vesicles among symbiotic protists of the lower termite, *Reticulitermes speratus*. *Microbes and Environments,* 19, 211-214.

Klemm, D., Heublein, B., Fink, H.P. and Bohn, A. 2005. Cellulose: fascinating biopolymer and sustainable raw material. *Angewandte Chemie International Edition,* 44, 3358-3393.

Klessmann, C., Held, A., Rathmann, M. and Ragwitz, M. 2011. Status and perspectives of renewable energy policy and deployment in the European Union—What is needed to reach the 2020 targets? *Energy Policy,* 39, 7637-7657.

Kosik, O., Bromley, J.R., Busse-Wicher, M., Zhang, Z. and Dupree, P. 2012. Studies of enzymatic cleavage of cellulose using polysaccharide analysis by carbohydrate gel electrophoresis (PACE). *Methods in Enzymology.* Elsevier.

Kristensen, J.H. 1972. Structure and function of crystalline styles of bivalves. *Ophelia,* 10, 91-108.

Kumar, S. 1998. Enzyme vs. extremozyme. *Resonance,* 3, 32-40.

Lambertz, C., Garvey, M., Klinger, J., Heesel, D., Klose, H., Fischer, R. and Commandeur, U. 2014. Challenges and advances in the heterologous expression of cellulolytic enzymes: a review. *Biotechnology for Biofuels,* 7, 135.

Lane, C.E. 1955. Recent biological studies on *Teredo* - a marine wood-boring mollusc. *The Scientific Monthly,* 80, 286-292.

Langdon, C.J. and Newell, R.I. 1990. Utilization of detritus and bacteria as food sources by two bivalve suspension-feeders, the oyster *Crassostrea virginica* and the mussel *Geukensia demissa*. *Marine Ecology Progress Series*, 299-310.

Langmead, B. and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9, 357.

Lavine, T.F. 1946. A study of the enzymatic and other properties of the crystalline style of clams; evidence for the presence of a cellulase. *Journal of Cellular Physiology,* 28, 183-95.

Lebour, M.V. 1946. The species of *Teredo* from Plymouth waters. *Journal of the Marine Biological Association of the United Kingdom,* 26, 381-389.

Lechene, C.P., Luyten, Y., McMahon, G. and Distel, D.L. 2007. Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science,* 317, 1563-1566.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M. and Henrissat, B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels,* 6, 41.

Leyva, A., Quintana, A., Sánchez, M., Rodríguez, E.N., Cremata, J. and Sánchez, J.C. 2008. Rapid and sensitive anthrone–sulfuric acid assay in microplate format to quantify carbohydrate in biopharmaceutical products: method development and validation. *Biologicals,* 36, 134-141.

Li, Q., Song, J., Peng, S., Wang, J.P., Qu, G.Z., Sederoff, R.R. and Chiang, V.L. 2014. Plant biotechnology for lignocellulosic biofuel production. *Plant Biotechnology Journal,* 12, 1174-92.

Liu, D. and Townsley, P. 1968. Glucose metabolism in the caecum of the marine borer *Bankia setacea*. *Journal of the Fisheries Board of Canada,* 25, 853-862.

Liu, D. and Walden, C. 1970. Enzymes of glucose metabolism in the caecum of the marine borer *Bankia setacea*. *Journal of the Fisheries Board of Canada,* 27, 1141-1146.

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

Lopes, S.G., Domaneschi, O., De Moraes, D.T., Morita, M. and Meserani, G.D.L. 2000. Functional anatomy of the digestive system of *Neoteredo reynei* (Bartsch, 1920) and *Psiloteredo healdi*

(Bartsch, 1931)(Bivalvia: Teredinidae). *Geological Society, London, Special Publications,* 177, 257-271.

Low, P.S. and Somero, G.N. 1976. Adaptation of muscle pyruvate kinases to environmental temperatures and pressures. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology,* 198, 1-11.

Lynd, L.R., Weimer, P.J., van Zyl, W.H. and Pretorius, I.S. 2002. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiology and Molecular Biology Reviews,* 66, 506-577.

Mackenzie, A.L. and Marshall, C. 2014. Proteins in the crystalline styles of the marine mussels *Perna canaliculus* Gmelin and *Mytilus galloprovincialis* Lamarck. *Journal of Shellfish Research,* 33, 673-685.

Mackintosh, N.A. 1925. The crystalline style in gastropods. *The Quarterly Journal of Microscopical Science,* 69, 317-345.

Makela, M.R., Donofrio, N. and de Vries, R.P. 2014. Plant biomass degradation by fungi. *Fungal Genetics and Biology,* 72, 2-9.

Mann, R. and Gallager, S.M. 1985. Growth, morphometry and biochemical composition of the wood boring molluscs *Teredo navalis* L., *Bankia gouldi* (Bartsch), and *Nototeredo knoxi* (Bartsch)(Bivalvia: Teredinidae). *Journal of Experimental Marine Biology and Ecology,* 85, 229-251.

Manwell, C. 1963. The chemistry and biology of hemoglobin in some marine clams - I. Distribution of the pigment and properties of the oxygen equilibrium. *Comparative Biochemistry and Physiology,* 8, 209-218.

Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A. and Ferrari, R. 2016. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics*, bbw114.

Marden, C.L., McDonald, R., Schreier, H.J. and Watts, J.E. 2017. Investigation into the fungal diversity within different regions of the gastrointestinal tract of *Panaque nigrolineatus*, a wood-eating fish. *AIMS microbiology,* 3, 749.

Marriott, P. 2014. *Identifying novel genes to improve lignocellulosic biomass as a feedstock for bioethanol.* Doctor of Philosophy, University of York.

McCann, M.C., Wells, B. and Roberts, K. 1990. Direct visualization of cross-links in the primary plant cell wall. *Journal of Cell Science,* 96, 323-334.

McDonald, R., Schreier, H.J. and Watts, J.E. 2012. Phylogenetic analysis of microbial communities in different regions of the gastrointestinal tract in *Panaque nigrolineatus*, a wood-eating fish. *PLoS ONE,* 7, e48018.

McHenery, J.G. and Birkbeck, T.H. 1979. Lysozyme of the mussel, *Mytilus edulis* (L). *Marine Biology Letters,* 1, 111-119.

McHenery, J.G., Birkbeck, T.H. and Allen, J.A. 1979. The occurrence of lysozyme in marine bivalves. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry,* 63, 25-28.

McLeod, M.P., Warren, R.L., Hsiao, W.W., Araki, N., Myhre, M., Fernandes, C., Miyazawa, D., Wong, W., Lillquist, A.L., Wang, D*., et al.* 2006. The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse. *Proceedings of the National Academy of Sciences of the United States of America,* 103, 15582-7.

McQuiston, R. 1970. Fine structure of the gastric shield in the lamellibranch bivalve, *Lasaea rubra* (Montagu). *Journal of Molluscan Studies,* 39, 69-75.

Mengel, M., Nauels, A., Rogelj, J. and Schleussner, C.-F. 2018. Committed sea-level rise under the Paris Agreement and the legacy of delayed mitigation action. *Nature Communications,* 9, 601.

Milatovič, M. and Štrus, J. 2010. Endogenous origin of endo-β-1, 4-glucanase in common woodlouse *Porcellio scaber* (Crustacea, Isopoda). *Journal of Comparative Physiology B,* 180, 1143-1153.

Miller, G.L. 1959. Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Analytical Chemistry,* 31, 426-428.

Miller, R.C. and Boynton, L.C. 1926. Digestion of wood by the shipworm. *Science,* 63, 524.

Miyauchi, K. 2000. Purification and characterization of lysozyme from brackishwater clam *Corbicula japonica*. *Nippon Suisan Gakkaishi,* 66, 275-281.

Moll, F. 1914. *Die Bohrmuschel (Genus Teredo Linné)*, Ulmer.

Moreira, L.R. and Filho, E.X. 2008. An overview of mannan structure and mannan-degrading enzyme systems. *Applied Microbiology and Biotechnology,* 79, 165-78.

Morton, B. 1970. The functional anatomy of the organs of feeding and digestion of *Teredo navalis* Linnaeus and *Lyrodus pedicellatus* (Quatrefages). *Proceedings of the Malacological Society of London,* 39, 151-168.

Morton, J. 1952. The role of the crystalline style. *Journal of Molluscan Studies,* 29, 85-92.

Moser, F., Irwin, D., Chen, S. and Wilson, D. 2008. Regulation and characterization of *Thermobifida fusca* carbohydrate-binding module proteins E7 and E8. *Biotechnology and Bioengineering,* 100, 1066-1077.

Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y., Holtzapple, M. and Ladisch, M. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresource Technology,* 96, 673-686.

Nair, N.B. and Saraswathy, M. 1971. The biology of wood-boring teredinid molluscs. *Advances in Marine Biology,* 9, 335-509.

Nelson, T.C. 1917. On the origin, nature, and function of the crystalline style of lamellinranchs. *Journal of Morphology,* 31, 53-111.

Nilsen, I.W., Øverbø, K., Sandsdalen, E., Sandaker, E., Sletten, K. and Myrnes, B. 1999. Protein purification and gene isolation of chlamysin, a cold-active lysozyme-like enzyme with antibacterial activity. *FEBS Letters,* 464, 153-158.

Nilsson, T. and Daniel, G. 1989. Chemistry and microscopy of wood decay by some higher ascomycetes. *Holzforschung-International Journal of the Biology, Chemistry, Physics and Technology of Wood,* 43, 11-18.

Nishimoto, A., Mito, S. and Shirayama, Y. 2009. Organic carbon and nitrogen source of sunken wood communities on continental shelves around Japan inferred from stable isotope ratios. *Deep-Sea Research Part II -Topical Studies in Oceanography,* 56, 1683-1688.

Nishiyama, Y., Langan, P. and Chanzy, H. 2002. Crystal structure and hydrogen-bonding system in cellulose Iβ from synchrotron x-ray and neutron fiber diffraction. *Journal of the American Chemical Society,* 124, 9074-9082.

Nozaki, M., Miura, C., Tozawa, Y. and Miura, T. 2009. The contribution of endogenous cellulase to the cellulose digestion in the gut of earthworm (*Pheretima hilgendorfi:* Megascolecidae). *Soil Biology and Biochemistry,* 41, 762-769.

O'Connor, R.M., Fung, J.M., Sharp, K.H., Benner, J.S., McClung, C., Cushing, S., Lamkin, E.R., Fomenkov, A.I., Henrissat, B., Londer, Y.Y.*, et al.* 2014. Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. *Proceedings of the National Academy of Sciences of the United States of America,* 111, E5096-104.

Ochoa-Villarreal, M., Aispuro-Hernndez, E., Vargas-Arispuro, I. and Martinez-Tellez, M. 2012. *Plant cell wall polymers: function, structure and biological activity of their derivatives*, Ailton de Souza Gomes.

Odling-Smee, L. 2007. Biofuels bandwagon hits a rut. *Nature,* 446, 483.

Oevering, P., Pitman, A.J. and Pandey, K.K. 2003. Wood digestion in *Pselactus spadi*x Herbst-a weevil attacking marine timber structures. *Biofouling,* 19, 249-254.

Olsen, Ø.M., Nilsen, I.W., Sletten, K. and Myrnes, B. 2003. Multiple invertebrate lysozymes in blue mussel (*Mytilus edulis*). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology,* 136, 107-115.

Orton, J. 1912. The mode of feeding of *Crepidula*, with an account of the current-producing mechanism in the mantle cavity, and some remarks on the mode of feeding in gastropods and lamellibranchs. *Journal of the Marine Biological Association of the United Kingdom,* 9, 444-478.

Paalvast, P. and van der Velde, G. 2013. What is the main food source of the shipworm (*Teredo navalis*)? A stable isotope approach. *Journal of Sea Research,* 80, 58-60.

Park, J.T. and Johnson, M.J. 1949. A submicrodetemination of glucose. *Journal of Biological Chemistry,* 181, 149-151.

Pauly, M. and Keegstra, K. 2008. Cell-wall carbohydrates and their modification as a resource for biofuels. *Plant Journal,* 54, 559-68.

Payne, C.M., Knott, B.C., Mayes, H.B., Hansson, H., Himmel, M.E., Sandgren, M., Stahlberg, J. and Beckham, G.T. 2015. Fungal cellulases. *Chemical Reviews,* 115, 1308-1448.

Payne, D.W., Thorpe, N.A. and Donaldson, E.M. 1972. Cellulolytic activity and a study of the bacterial population in the digestive tract of *Scrobicularia plana* (Da Costa). *Proceedings of the Malacological Society of London,* 40, 147-160.

Peitsch, M.C. and Tschopp, J. 1991. Assembly of macromolecular pores by immune defense systems. *Current Opinion in Cell Biology,* 3, 710-716.

Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B. and Keasling, J.D. 2012. Microbial engineering for the production of advanced biofuels. *Nature,* 488, 320-8.

Pérez, J., Munoz-Dorado, J., de la Rubia, T. and Martinez, J. 2002. Biodegradation and biological treatments of cellulose, hemicellulose and lignin: an overview. *International Microbiology,* 5, 53-63.

Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8, 785.

Pham, V.H. and Kim, J. 2012. Cultivation of unculturable soil bacteria. *Trends in Biotechnology,* 30, 475-484.

Pollegioni, L., Tonin, F. and Rosini, E. 2015. Lignin-degrading enzymes. *The FEBS Journal,* 282, 1190-1213.

Pomin, V.H. and Mourão, P.A. 2008. Structure, biology, evolution, and medical importance of sulfated fucans and galactans. *Glycobiology,* 18, 1016-1027.

Popham, J.D. and Dickson, M.R. 1973. Bacterial associations in the Teredo *Bankia australis* (Lamellibranchia: Mollusca). *Marine Biology,* 19, 338-340.

Potts, F.A. 1923. The structure and function of the liver of *Teredo*, the shipworm. *Proceedings of the Cambridge Philosophical Society-Biological Sciences,* 1, 1-17.

Poutiers, J. 1998. Bivalves. *FAO species identification guide for fishery purposes. The living marine resources of the Western Central Pacific. Volume 1. Seaweeds, corals, bivalves and gastropods.* Food and Agriculture Organization of the United Nations.

Prade, R.A. 1996. Xylanases: from biology to biotechnology. *Biotechnology and Genetic Engineering Reviews,* 13, 101-132.

Purchon, R. 1955. The structure and function of the British Pholadidae (rock-boring Lamellibranchia). *Proceedings of the Zoological Society of London,* 124, 859-911.

Purchon, R. 1971. Digestion in filter feeding bivalves - a new concept. *Journal of Molluscan Studies,* 39, 253-262.

Purchon, R.D. 1941. On the biology and relationships of the lamellibranch *Xylophaga dorsalis* (Turton). *Journal of the Marine Biological Association of the United Kingdom,* 25, 1-39.

Quadrelli, R. and Peterson, S. 2007. The energy-climate challenge: recent trends in CO_2 emissions from fuel combustion. *Energy Policy,* 35, 5938-5952.

Quinlan, R.J., Sweeney, M.D., Leggio, L.L., Otten, H., Poulsen, J.-C.N., Johansen, K.S., Krogh, K.B., Jørgensen, C.I., Tovborg, M. and Anthonsen, A. 2011. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. *Proceedings of the National Academy of Sciences,* 108, 15079-15084.

Raghothama, S., Simpson, P.J., Szabó, L., Nagy, T., Gilbert, H.J. and Williamson, M.P. 2000. Solution structure of the CBM10 cellulose binding module from Pseudomonas xylanase A. *Biochemistry,* 39, 978-984.

Rapp, P. and Beerman, A. 1991. Bacterial cellulases. Biosynthesis and biodegradation of cellulose. *In:* Haigler, C.W., Pj (ed.) *Biosynthesis and biodegradation of cellulose.* New York: Marcel Dekker, Inc.

Reidinger, S., Ramsey, M.H. and Hartley, S.E. 2012. Rapid and accurate analyses of silicon and phosphorus in plants using a portable X-ray fluorescence spectrometer. *New Phytologist,* 195, 699-706.

Riley, R., Salamov, A.A., Brown, D.W., Nagy, L.G., Floudas, D., Held, B.W., Levasseur, A., Lombard, V., Morin, E., Otillar, R.*, et al.* 2014. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proceedings of the National Academy of Sciences of the United States of America,* 111, 9923-8.

Roch, F. 1932. Einige Beobachtungen zur Ökologie und Physiologie von *Teredo navalis* L. *Arkiv för zoologi,* 24, 1-18.

Roch, F. 1940. *Die Terediniden des Mittelmeeres*, Deutsch-Italienisches Institut für Meeresbiologie zu Rovigno d'Istria.

Roeselers, G. and Newton, I.L. 2012. On the evolutionary ecology of symbioses between chemosynthetic bacteria and bivalves. *Applied Microbiology and Biotechnology,* 94, 1-10.

Rosado, C.J., Buckle, A.M., Law, R.H., Butcher, R.E., Kan, W.T., Bird, C.H., Ung, K., Browne, K.A., Baran, K., Bashtannyk-Puhalovich, T.A.*, et al.* 2007. A common fold mediates vertebrate defense and bacterial attack. *Science,* 317, 1548-51.

Rosado, C.J., Kondos, S., Bull, T.E., Kuiper, M.J., Law, R.H., Buckle, A.M., Voskoboinik, I., Bird, P.I., Trapani, J.A. and Whisstock, J.C. 2008. The MACPF/CDC family of pore-forming toxins. *Cellular Microbiology,* 10, 1765-1774.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O. and Ofran, Y. 2003. Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS,* 60, 2637-2650.

Rubin, E.M. 2008. Genomics of cellulosic biofuels. *Nature,* 454, 841-5.

Sabbadin, F., Hemsworth, G.R., Ciano, L., Henrissat, B., Dupree, P., Tryfona, T., Marques, R.D., Sweeney, S., Besser, K. and Elias, L. 2018a. An ancient family of lytic polysaccharide monooxygenases with roles in arthropod development and biomass digestion. *Nature Communications,* 9, 756.

Sabbadin, F., Pesante, G., Elias, L., Besser, K., Li, Y., Steele-King, C., Stark, M., Rathbone, D.A., Dowle, A.A. and Bates, R. 2018b. Uncovering the molecular mechanisms of lignocellulose digestion in shipworms. *Biotechnology for Biofuels,* 11, 59.

Sahu, A. and Lambris, J.D. 2001. Structure and biology of complement protein C3, a connecting link between innate and acquired immunity. *Immunological Reviews,* 180, 35-48.

Sakamoto, K., Touhata, K., Yamashita, M., Kasai, A. and Toyohara, H. 2007. Cellulose digestion by common Japanese freshwater clam *Corbicula japonica*. *Fisheries Science,* 73, 675-683.

Sakamoto, K. and Toyohara, H. 2009. Putative endogenous xylanase from brackish-water clam *Corbicula japonica*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology,* 154, 85-92.

Sakamoto, K., Uji, S., Kurokawa, T. and Toyohara, H. 2008. Immunohistochemical, in situ hybridization and biochemical studies on endogenous cellulase of *Corbicula japonica*. *Comparative Biochemistry and Physiology B Biochemistry & Molecular Biology,* 150, 216-21.

Sakamoto, K., Uji, S., Kurokawa, T. and Toyohara, H. 2009. Molecular cloning of endogenous β-glucosidase from common Japanese brackish water clam *Corbicula japonica*. *Gene,* 435, 72-79.

Salt, L.A., Thomas, H., Prowe, A., Borges, A.V., Bozec, Y. and Baar, H.J. 2013. Variability of North Sea pH and CO_2 in response to North Atlantic oscillation forcing. *Journal of Geophysical Research: Biogeosciences,* 118, 1584-1592.

Saraswathy, M. and Nair, N.B. 1971. Observations on the structure of the shipworms, *Nausitora hedleyi, Teredo furcifera* and *Teredora princesae* (Bivalvia: Teredinidae). *Earth and Environmental Science Transactions of The Royal Society of Edinburgh,* 68, 507-566.

Sattler, S.E. and Funnell-Harris, D.L. 2013. Modifying lignin to improve bioenergy feedstocks: strengthening the barrier against pathogens? *Frontiers of Plant Science,* 4, 70.

Savazzi, E. 2001. A review of symbiosis in the Bivalvia, with special attention to macrosymbiosis. *Paleontological Research,* 5, 55-73.

Scharf, M.E., Karl, Z.J., Sethi, A. and Boucias, D.G. 2011. Multiple levels of synergistic collaboration in termite lignocellulose digestion. *PLoS ONE,* 6, e21709.

Scheller, H.V. and Ulvskov, P. 2010. Hemicelluloses. *Annual Review of Plant Biology,* 61, 263-89.

Schmieder, R. and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27, 863-864.

Schröder, A. 2016. *Feeding strategies of shipworms (Bivalvia: Teredinidae) in the Lagoon of Venice.* Doctor of Philosophy, Universität Hamburg, Fachbereich Biologie.

Schulz, M.N., Landström, J. and Hubbard, R.E. 2013. MTSA-A Matlab program to fit thermal shift data. *Analytical Biochemistry,* 433, 43-47.

Schwarz, W.H. 2001. The cellulosome and cellulose degradation by anaerobic bacteria. *Applied Microbiology and Biotechnology,* 56, 634-649.

Scofield, J. 1975. Christopher Columbus: the sailor who gave us the New World. *National Geographic Magazine,* 148.

Seiderer, L., Davis, C., Robb, F. and Newell, R. 1984. Utilisation of bacteria as nitrogen resource by kelp-bed mussel *Choromytilus meridionalis*. *Marine Ecology Progress Series*, 109-116.

Seiderer, L. and Newell, R. 1979. Adjustment of the activity of α-amylase extracted from the style of the black mussel *Choromytilus meridionalis* (Krauss) in response to thermal acclimation. *Journal of Experimental Marine Biology and Ecology,* 39, 79-86.

Sellius, G. 1733. *Historia naturalis teredinis seu xlophagi marini tubulo-chonchoidis speciatim Belgici (Trajecti ad Rhenum)*.

Shafiee, S. and Topal, E. 2009. When will fossil fuel reserves be diminished? *Energy Policy,* 37, 181-189.

Shah, M.M., Barr, D.P., Chung, N. and Aust, S.D. 1992. Use of white rot fungi in the degradation of environmental chemicals. *Toxicology Letters,* 64, 493-501.

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

Shi, W., Xie, S., Chen, X., Sun, S., Zhou, X., Liu, L., Gao, P., Kyrpides, N.C., No, E.G. and Yuan, J.S. 2013a. Comparative genomic analysis of the microbiome of herbivorous insects reveals eco-environmental adaptations: biotechnology applications. *PLoS Genetics,* 9, e1003131.

Shi, Y., Chai, L., Tang, C., Yang, Z., Zhang, H., Chen, R., Chen, Y. and Zheng, Y. 2013b. Characterization and genomic analysis of kraft lignin biodegradation by the beta-proteobacterium *Cupriavidus basilensis* B-8. *Biotechnology for Biofuels,* 6, 1.

Shimizu, M., Kaneko, Y., Ishihara, S., Mochizuki, M., Sakai, K., Yamada, M., Murata, S., Itoh, E., Yamamoto, T., Sugimura, Y.*, et al.* 2015. Novel beta-1,4-mannanase belonging to a new glycoside hydrolase family in *Aspergillus nidulans*. *Journal of Biological Chemistry,* 290, 27914-27.

Shipway, J.R. 2013. *Aspects of the life history strategies of the Teredinidae.* Doctor of Philosophy, University of Portsmouth.

Shipway, J.R., O'Connor, R., Stein, D., Cragg, S.M., Korshunova, T., Martynov, A., Haga, T. and Distel, D.L. 2016. *Zachsia zenkewitschi* (Teredinidae), a rare and unusual seagrass boring bivalve revisited and redescribed. *PLoS ONE,* 11, e0155269.

Showalter, A.M. 1993. Structure and function of plant cell wall proteins. *The Plant Cell,* 5, 9-23.

Si, A.A., C.G.: Bellwood, O. 2000. Habitat partitioning by two wood-boring invertebrates in a mangrove system in tropical Australia. *Journal of the Marine Biological Association of the United Kingdom,* 80, 1131-1132.

Siddiqui, K.S., Ertan, H., Charlton, T., Poljak, A., Daud Khaled, A.K., Yang, X., Marshall, G. and Cavicchioli, R. 2014. Versatile peroxidase degradation of humic substances: use of isothermal titration calorimetry to assess kinetics, and applications to industrial wastes. *Journal of Biotechnology,* 178, 1-11.

Sigerfoos, C.P. 1908. Natural history, organization, and late development of the Teredinidae, or ship-worms. *Bulletin of the United States Bureau of Fisheries,* 27, 191-231.

Simmons, T.J., Mortimer, J.C., Bernardinelli, O.D., Pöppler, A.-C., Brown, S.P., Dupree, R. and Dupree, P. 2016. Folding of xylan onto cellulose fibrils in plant cell walls revealed by solid-state NMR. *Nature Communications,* 7, 13902.

Sims, R.E., Mabee, W., Saddler, J.N. and Taylor, M. 2010. An overview of second generation biofuel technologies. *Bioresource Technology,* 101, 1570-1580.

Sippel, A.E. 1973. Purification and characterization of adenosine triphosphate: ribonucleic acid adenyltransferase from *Escherichia coli*. *European Journal of Biochemistry,* 37, 31-40.

Smant, G., Stokkermans, J.P., Yan, Y., de Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L.*, et al.* 1998. Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proceedings of the National Academy of Sciences of the United States of America,* 95, 4906-11.

Smith, C., Kukert, H., Wheatcroft, R., Jumars, P. and Deming, J. 1989. Vent fauna on whale remains. *Nature,* 341, 27.

Smucker, R.A. and Wright, D.A. 1984. Chitinase activity in the crystalline style of the American oyster *Crassostrea virginica*. *Comparative Biochemistry and Physiology a-Physiology,* 77, 239-241.

Smucker, R.A. and Wright, D.A. 1986. Characteristics of *Crassostrea virginica* crystalline style chitin digestion. *Comparative Biochemistry and Physiology a-Physiology,* 83, 489-493.

Song, L., Wang, L., Qiu, L. and Zhang, H. 2010. Bivalve immunity. *Invertebrate immunity.* Springer.

Sova, V.V., Elyakova, L.A. and Vaskovsky, V.E. 1970. Purification and some properties of β-1,3-glucan glucanohydrolase from the crystalline style of bivalvia, *Spisula sachalinensis*. *Biochimica et Biophysica Acta,* 212, 111-115.

Stewart, E.J. 2012. Growing unculturable bacteria. *Journal of Bacteriology,* 194, 4151-4160.

Stips, A., Macias, D., Coughlan, C., Garcia-Gorriz, E. and San Liang, X. 2016. On the causal structure between CO_2 and global temperature. *Scientific Reports,* 6, 21691.

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

Sun, J., Ding, S.-Y. and Doran-Peterson, J. 2013. Biomass and its biorefinery: novel approaches from nature-inspired strategies and technology. *In:* Jianzhong Sun, S.-Y.D., Joy D Peterson (ed.) *Biological conversion of biomass for fuels and chemicals: explorations from natural utilization systems.* Royal Society of Chemistry.

Suzuki, K.-i., Ojima, T. and Nishita, K. 2003. Purification and cDNA cloning of a cellulase from abalone *Haliotis discus hannai*. *European Journal of Biochemistry,* 270, 771-778.

Szallasi, Z. 1999. Genetic network analysis in light of massively parallel biological data acquisition. *Biocomputing'99.* World Scientific.

Takahashi, K.G. and Itoh, N. 2011. Lysozymes in molluscs. *Diseases in Asian aquaculture VII. Malaysia*, 93-102.

Tall, B.D. and Nauman, R.K. 1981. Scanning electron microscopy of *Cristispira* species in Chesapeake Bay oysters. *Applied and Environmental Microbiology,* 42, 336-343.

Taprab, Y., Johjima, T., Maeda, Y., Moriya, S., Trakulnaleamsai, S., Noparatnaraporn, N., Ohkuma, M. and Kudo, T. 2005. Symbiotic fungi produce laccases potentially involved in phenol degradation in fungus combs of fungus-growing termites in Thailand. *Applied and Environmental Microbiology,* 71, 7696-704.

Tartar, A., Wheeler, M.M., Zhou, X., Coy, M.R., Boucias, D.G. and Scharf, M.E. 2009. Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. *Biotechnology for Biofuels,* 2, 25.

Taupp, M., Mewis, K. and Hallam, S.J. 2011. The art and design of functional metagenomic screens. *Current Opinion in Biotechnology,* 22, 465-72.

Teeri, T.T. 1997. Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. *Trends in Biotechnology,* 15, 160-167.

ten Have, R. and Teunissen, P.J.M. 2001. Oxidative mechanisms involved in lignin degradation by white-rot fungi. *Chemical Reviews,* 101, 3397-3414.

Thomas, L.H., Forsyth, V.T., Sturcova, A., Kennedy, C.J., May, R.P., Altaner, C.M., Apperley, D.C., Wess, T.J. and Jarvis, M.C. 2013. Structure of cellulose microfibrils in primary cell walls from collenchyma. *Plant Physiology,* 161, 465-76.

Thomas, W. and Dutcher, R.A. 1924. The colorimetric determination of carbohydrates in plants by the picrid acid reduction method. I. The estimation of reducing sugars and sucrose. *Journal of the American Chemical Society,* 46, 1662-1669.

Tian, J.-H., Pourcher, A.-M., Bouchez, T., Gelhaye, E. and Peu, P. 2014. Occurrence of lignin degradation genotypes and phenotypes among prokaryotes. *Applied Microbiology and Biotechnology,* 98, 9527-9544.

Toulza, E., Shin, M.-S., Blanc, G., Audic, S., Laabir, M., Collos, Y., Claverie, J.-M. and Grzebyk, D. 2010. Gene expression in proliferating cells of the dinoflagellate *Alexandrium catenella* (Dinophyceae). *Applied and Environmental Microbiology,* 76, 4521-4529.

Treneman, N.C., Borges, L.M., Shipway, J.R., Raupach, M.J., Altermark, B. and Carlton, J.T. 2018. A molecular phylogeny of wood-borers (Teredinidae) from Japanese tsunami marine debris. *Aquatic Invasions,* 13, 101-112.

Tripp, M.R. 1963. Cellular responses of mollusks. *Annals of the New York Academy of Sciences,* 113, 467-474.

Tsuji, A., Tominaga, K., Nishiyama, N. and Yuasa, K. 2013. Comprehensive enzymatic analysis of the cellulolytic system in digestive fluid of the sea hare *Aplysia kurodai.* Efficient glucose release from sea lettuce by synergistic action of 45 kDa endoglucanase and 210 kDa ß-glucosidase. *PLoS ONE,* 8, e65418.

Turner, R.D. 1955. The family Pholadidae in the western Atlantic and the eastern Pacific II. Martesiinae, Jouannetinae and Xylophaginae. *Johnsonia,* 3, 65-160.

Turner, R.D. 1959. The status of systematic work in the Teredinidae. *Marine boring and fouling organisms (DL Ray ed)*, 124-136.

Turner, R.D. 1966. A survey and illustrated catalogue of the Teredinidae (Mollusca: Bivalvia). *A survey and illustrated catalogue of the Teredinidae (Mollusca: Bivalvia).*

Turner, R.D. 1971a. Biology of marine wood-boring molluscs. *Marine borers, fungi and fouling organisms of wood*, 259-301.

Turner, R.D. 1971b. Identification of marine wood-boring molluscs. *In:* Jones, E.B.G. & Eltringham, S.K. (eds.) *Marine Borers and Fouling Organisms of Wood.* OECD, Paris: Organisation for economic co-operation and development.

Turner, R.D. and Yakovlev, Y. 1983. Dwarf males in the teredinidae (Bivalvia, Pholadacea). *Science,* 219, 1077-1078.

U.N. 2013. World Population Prospects. The 2012 Revision. *In:* Division, U.N.P. (ed.). New York.

U.N.C.T.A.D. 2016. *Second generation biofuel markets: state of play, trade and developing country perspectives* [Online]. Geneva: United Nations. Available: https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=1455 [Accessed 21 August 2018].

U.N.F.C.C.C. 2015. *The Paris Agreement. United Nations Framework Convention on Climate Change.* [Online]. Paris. Available: http://unfccc.int/paris_agreement/items/9485.php [Accessed 20 August 2018].

UNCTAD. 2016. *Second generation biofuel markets: state of play, trade and developing country perspectives* [Online]. United Nations Geneva. Available: https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=1455 [Accessed 21 August 2018].

Vaaje-Kolstad, G., Westereng, B., Horn, S.J., Liu, Z., Zhai, H., Sørlie, M. and Eijsink, V.G. 2010. An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. *Science,* 330, 219-222.

Vahldiek, B.-W. and Schweigert, G. 2007. Oldest record of wood-boring bivalves. *Neues Jahrbuch für Geologie und Paläontologie-Abhandlungen,* 244, 261-271.

Väljamäe, P., Kipper, K., Pettersson, G. and Johansson, G. 2003. Synergistic cellulose hydrolysis can be described in terms of fractal-like kinetics. *Biotechnology and Bioengineering,* 84, 254-257.

van den Brink, J. and de Vries, R.P. 2011. Fungal enzyme sets for plant polysaccharide degradation. *Applied Microbiology and Biotechnology,* 91, 1477.

Vane, C.H., Drage, T.C. and Snape, C.E. 2006. Bark decay by the white-rot fungus *Lentinula edodes*: polysaccharide loss, lignin resistance and the unmasking of suberin. *International Biodeterioration & Biodegradation,* 57, 14-23.

Vasta, G.R. 2012. Galectins as pattern recognition receptors: structure, function, and evolution. *Current Topics in Innate Immunity II.* Springer.

Vispo, C. and Hume, I.D. 1995. The digestive tract and digestive function in the North American porcupine and beaver. *Canadian Journal of Zoology,* 73, 967-974.

Vogel, J. 2008. Unique aspects of the grass cell wall. *Current Opinion in Plant Biology,* 11, 301-7.

Voight, J.R. 2007. Experimental deep-sea deployments reveal diverse Northeast Pacific wood-boring bivalves of Xylophagainae (Myoida: Pholadidae). *Journal of Molluscan Studies,* 73, 377-391.

Voight, J.R. 2015. Xylotrophic bivalves: aspects of their biology and the impacts of humans. *Journal of Molluscan Studies,* 81, 175-186.

Waldrop, M. 2007. Kill king corn. *Nature,* 449, 637.

Wang, J., Ding, M., Li, Y.-H., Chen, Q.-X., Xu, G.-J. and Zhao, F.-K. 2003. Isolation of a multi-functional endogenous cellulase gene from mollusc, *Ampullaria crossean*. *Sheng wu hua xue yu sheng wu wu li xue bao Acta biochimica et biophysica Sinica,* 35, 941-946.

Wang, L., Song, X. and Song, L. 2018. The oyster immunity. *Developmental & Comparative Immunology,* 80, 99-118.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N.*, et al.* 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature,* 450, 560-5.

Watanabe, H. and Tokuda, G. 2010. Cellulolytic systems in insects. *Annual Review of Entomology,* 55, 609-32.

Watanabe, H.N., H.; Tokuda, G.; Lo, N. 1998. A cellulase gene of termite origin. *Nature,* 394, 330-331.

Waterbury, J.B., Calloway, C.B. and Turner, R.D. 1983. A cellulolytic nitrogen-fixing bacterium cultured from the Gland of Deshayes in shipworms (Bivalvia: Teredinidae). *Science,* 221, 1401-1403.

Watts, J., McDonald, R., Daniel, R. and Schreier, H. 2013. Examination of a culturable microbial population from the gastrointestinal tract of the wood-eating Loricariid catfish *Panaque nigrolineatus*. *Diversity,* 5, 641-656.

Weimer, P.J. 2013. CHAPTER 14. The ruminant animal as a natural biomass-conversion platform and a source of bioconversion agents. 248-281.

Welborn, J.R. and Manahan, D.T. 1990. Direct measurements of sugar uptake from seawater into molluscan larvae. *Marine Ecology Progress Series*, 233-239.

Whisstock, J.C. and Lesk, A.M. 2003. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics,* 36, 307-340.

Wilson, D.B. 2011. Microbial diversity of cellulose hydrolysis. *Current Opinion in Microbiology,* 14, 259-63.

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

Wojtowicz, M.B. 1972. Carbohydrases of the digestive gland and the crystalline style of the Atlantic deep-sea scallop (*Placopecten magellanicus*, Gmelin). *Comparative Biochemistry & Physiology A Comparative Physiology,* 43, 131-41.

Wright, E.M. 2001. Renal Na^+-glucose cotransporters. *American Journal of Physiology-Renal Physiology,* 280, F10-F18.

Wurm, F.M. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature Biotechnology,* 22, 1393.

Wurzinger-Mayer, A., Shipway, J.R., Kristof, A., Schwaha, T., Cragg, S.M. and Wanninger, A. 2014. Developmental dynamics of myogenesis in the shipworm *Lyrodus pedicellatus* (Mollusca: Bivalvia). *Frontiers in Zoology,* 11, 90.

Xie, L., Liu, N. and Huang, Y. 2013. Chapter 8. Lignocellulose degradation in termite symbiotic systems. *Biological Conversion of Biomass for Fuels and Chemicals.*

Xie, S., Syrenne, R., Sun, S. and Yuan, J.S. 2014. Exploration of Natural Biomass Utilization Systems (NBUS) for advanced biofuel--from systems biology to synthetic design. *Current Opinion in Biotechnology,* 27, 195-203.

Xu, B., Hagglund, P., Stalbrand, H. and Janson, J.C. 2002. Endo-beta-1,4-mannanases from blue mussel, *Mytilus edulis*: purification, characterization, and mode of action. *Journal of Biotechnology,* 92, 267-77.

Xu, B., Janson, J.C. and Sellos, D. 2001. Cloning and sequencing of a molluscan endo-β-1, 4-glucanase gene from the blue mussel, *Mytilus edulis*. *European Journal of Biochemistry,* 268, 3718–3727.

Xue, Q.G., Itoh, N., Schey, K.L., Li, Y.L., Cooper, R.K. and La Peyre, J.F. 2007. A new lysozyme from the eastern oyster (*Crassostrea virginica*) indicates adaptive evolution of i-type lysozymes. *Cellular and Molecular Life Sciences,* 64, 82-95.

Yamakawa, M. and Tanaka, H. 1999. Immune proteins and their gene expression in the silkworm, *Bombyx mori*. *Developmental & Comparative Immunology,* 23, 281-289.

Yang, B. and Wyman, C.E. 2008. Pretreatment: the key to unlocking low-cost cellulosic ethanol. *Biofuels, Bioproducts and Biorefining: Innovation for a sustainable economy,* 2, 26-40.

Yang, J.C., Madupu, R., Durkin, A.S., Ekborg, N.A., Pedamallu, C.S., Hostetler, J.B., Radune, D., Toms, B.S., Henrissat, B., Coutinho, P.M.*, et al.* 2009. The complete genome of *Teredinibacter turnerae* T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). *PLoS ONE,* 4, e6085.

Yokoe, Y. and Yasumasu, I. 1964. The distribution of cellulase in invertebrates. *Comparative Biochemistry and Physiology,* 13, 323-338.

Yonge, C.M. 1931. Studies on the comparative physiology of digestion. *British journal of experimental Zoology,* 1, 15-63.

Yonge, C.M. 1937. Review of digestion in Metazoa. *Biological Reviews,* 12, 87-115.

Yoon, S.H., Kim, S.K. and Kim, J.F. 2010. Secretory production of recombinant proteins in *Escherichia coli*. *Recent patents on Biotechnology,* 4, 23-29.

Zannella, C., Mosca, F., Mariani, F., Franci, G., Folliero, V., Galdiero, M., Tiscar, P.G. and Galdiero, M. 2017. Microbial diseases of bivalve mollusks: infections, immunology and antimicrobial defense. *Marine Drugs,* 15, 182.

Zhang, J., Presley, G.N., Hammel, K.E., Ryu, J.-S., Menke, J.R., Figueroa, M., Hu, D., Orr, G. and Schilling, J.S. 2016. Localizing gene regulation reveals a staggered wood decay mechanism for the brown rot fungus *Postia placenta*. *Proceedings of the National Academy of Sciences,* 113, 10968-10973.

Zhong, R., Lee, C. and Ye, Z.H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science,* 15, 625-32.

Publications arising from this work

Sabbadin et al. Biotechnol Biofuels (2018) 11:59
https://doi.org/10.1186/s13068-018-1058-3

Biotechnology for Biofuels

RESEARCH Open Access

Uncovering the molecular mechanisms of lignocellulose digestion in shipworms

Federico Sabbadin[1†], Giovanna Pesante[1†], Luisa Elias[1], Katrin Besser[1], Yi Li[1], Clare Steele-King[1], Meg Stark[2], Deborah A. Rathbone[3], Adam A. Dowle[2], Rachel Bates[2], J. Reuben Shipway[4], Simon M. Cragg[5], Neil C. Bruce[1] and Simon J. McQueen-Mason[1*]

Abstract

Lignocellulose forms the structural framework of woody plant biomass and represents the most abundant carbon source in the biosphere. Turnover of woody biomass is a critical component of the global carbon cycle, and the enzymes involved are of increasing industrial importance as industry moves away from fossil fuels to renewable carbon resources. Shipworms are marine bivalve molluscs that digest wood and play a key role in global carbon cycling by processing plant biomass in the oceans. Previous studies suggest that wood digestion in shipworms is dominated by enzymes produced by endosymbiotic bacteria found in the animal's gills, while little is known about the identity and function of endogenous enzymes produced by shipworms. Using a combination of meta-transcriptomic, proteomic, imaging and biochemical analyses, we reveal a complex digestive system dominated by uncharacterized enzymes that are secreted by a specialized digestive gland and that accumulate in the cecum, where wood digestion occurs. Using a combination of transcriptomics, proteomics, and microscopy, we show that the digestive proteome of the shipworm *Lyrodus pedicellatus* is mostly composed of enzymes produced by the animal itself, with a small but significant contribution from symbiotic bacteria. The digestive proteome is dominated by a novel 300 kDa multi-domain glycoside hydrolase that functions in the hydrolysis of β-1,4-glucans, the most abundant polymers in wood. These studies allow an unprecedented level of insight into an unusual and ecologically important process for wood recycling in the marine environment, and open up new biotechnological opportunities in the mobilization of sugars from lignocellulosic biomass.

Background

Large amounts of wood from terrestrial plants enter the marine environment and support complex ecosystems. Tropical mangrove swamps, for example, provide safe nurseries that support fisheries [1] and are among the most productive ecosystems on the planet. It has been shown that around 70% of dead wood in mangroves is processed by the action of wood-boring bivalve molluscs known as shipworms [2, 3]. Shipworms acquired their name due to their devastating effects on wooden ships prior to the advent of copper-bottoming, a process developed to protect wooden vessels from their attack. Shipworms continue to destroy timber structures and docks around the world but, despite their ecological, historical and economic importance, the process by which these animals digest wood remains poorly understood.

Shipworms burrow cylindrical tunnels using specialized shell valves with abrasive toothed ridges as a rasp and ingest the wood particles as they burrow. Wood particles are transported by ciliary currents through the esophagus and stomach to accumulate in the cecum [4], which occupies a large part of the animal's body (Fig. 1a) and is thought to form the main site of wood digestion [2]. While the digestive systems of most herbivorous and xylophagous animals harbor commensal microbes that assist with digestion, the shipworm cecum is reported to be largely devoid of microbial life [5]. However, large amounts of carbohydrate active enzymes (CAZymes)

*Correspondence: simon.mcqueenmason@york.ac.uk
†Federico Sabbadin and Giovanna Pesante contributed equally to this work
[1] Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

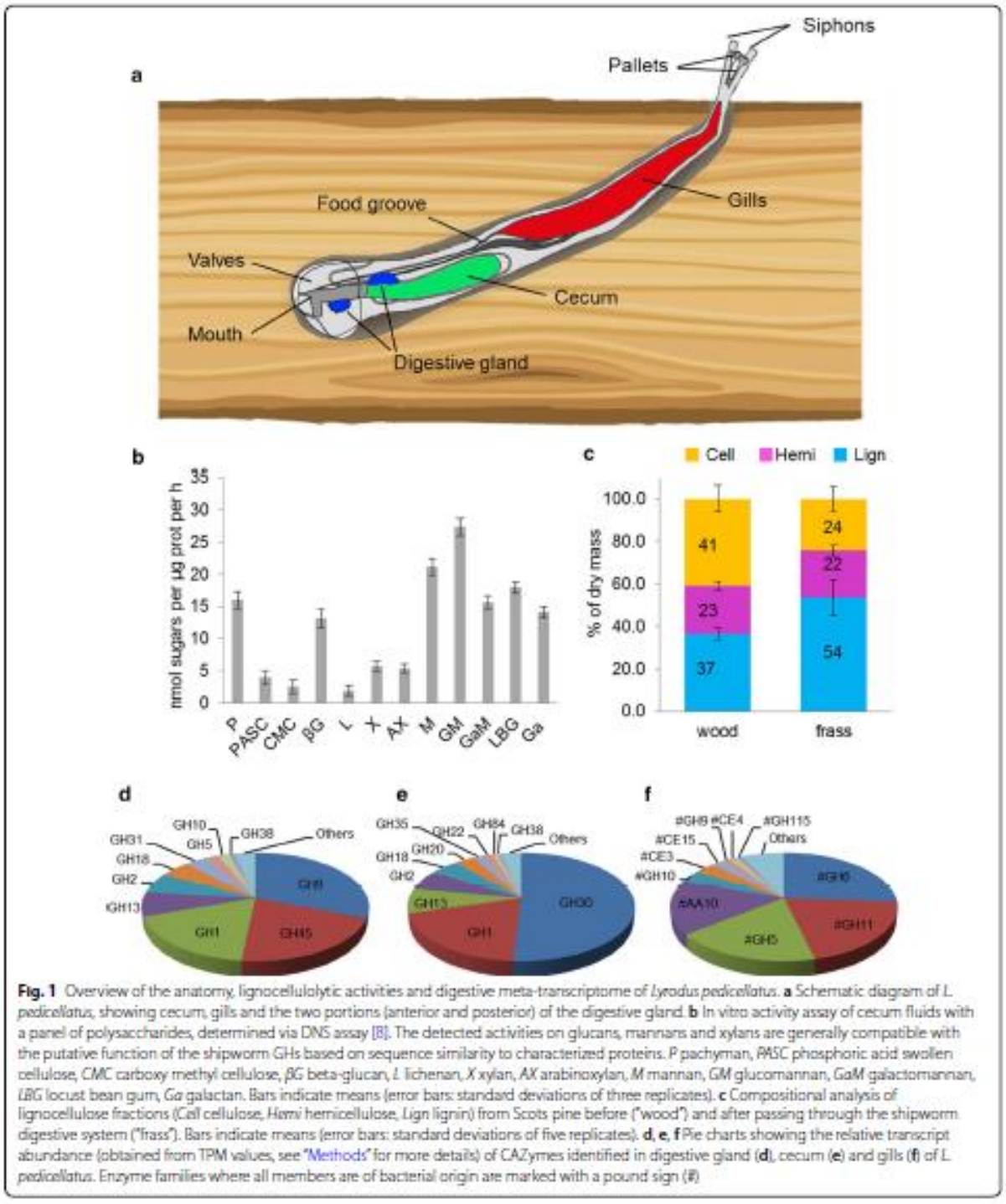


Fig. 1 Overview of the anatomy, lignocellulolytic activities and digestive meta-transcriptome of *Lyrodus pedicellatus*. **a** Schematic diagram of *L. pedicellatus*, showing cecum, gills and the two portions (anterior and posterior) of the digestive gland. **b** In vitro activity assay of cecum fluids with a panel of polysaccharides, determined via DNS assay [8]. The detected activities on glucans, mannans and xylans are generally compatible with the putative function of the shipworm GHs based on sequence similarity to characterized proteins. *P* pachyman, *PASC* phosphoric acid swollen cellulose, *CMC* carboxy methyl cellulose, *βG* beta-glucan, *L* lichenan, *X* xylan, *AX* arabinoxylan, *M* mannan, *GM* glucomannan, *GaM* galactomannan, *LBG* locust bean gum, *Ga* galactan. Bars indicate means (error bars: standard deviations of three replicates). **c** Compositional analysis of lignocellulose fractions (*Cell* cellulose, *Hemi* hemicellulose, *Lign* lignin) from Scots pine before ("wood") and after passing through the shipworm digestive system ("frass"). Bars indicate means (error bars: standard deviations of five replicates). **d**, **e**, **f** Pie charts showing the relative transcript abundance (obtained from TPM values, see "Methods" for more details) of CAZymes identified in digestive gland (**d**), cecum (**e**) and gills (**f**) of *L. pedicellatus*. Enzyme families where all members are of bacterial origin are marked with a pound sign (#)

have been shown to be produced by endosymbiotic bacteria housed in specialized cells (bacteriocytes) in the animal's gills, and have been reported to play a major role in wood digestion by shipworms [6]. The gills are spatially distant from the cecum (Fig. 1a), and the route by which bacterial enzymes move to the site of wood digestion

remains elusive. Previous work in *Bankia setacea* suggests that bacterial CAZymes account for the majority of the digestive proteome in shipworms [6]. We have undertaken studies in *Lyrodus pedicellatus* with the aim of better understanding the digestive processes in shipworms, and here reveal the importance of the shipworm's own enzymes in wood digestion.

Results

Meta-transcriptomic analysis of *L. pedicellatus* and its endosymbionts

In vitro activity assays carried out with cecum fluids of *L. pedicellatus* against a panel of substrates revealed a complex enzymatic cocktail with activity against many polysaccharides associated with lignocellulosic biomass (Fig. 1b). Compositional analysis of wood and shipworm feces (frass) revealed that more than 40% of the cellulose content and lesser amounts of hemicellulose are removed while passing through the shipworm digestive system (Fig. 1c). While compelling evidence has shown that gill bacteriocytes are the main site for the production of bacterial CAZymes in shipworms, several authors have also hypothesized that the digestive gland might be responsible for the synthesis of endogenous enzymes, while the cecum appears to be the site of wood breakdown and could potentially be involved in sugar uptake [2, 4]. In order to identify the key genes involved in wood digestion and absorption of breakdown products in *L. pedicellatus*, we performed meta-transcriptome sequencing from the main organs putatively involved in wood digestion (digestive gland, cecum, and gills) (Fig. 1a, Additional file 1: Table S1) using healthy adult *L. pedicellatus* growing in blocks of Scots pine submerged in sea water.

Our gene expression analysis reveals that the shipworm digestive gland is the major site of transcription of endogenous lignocellulolytic enzymes in *L. pedicellatus*, all carrying a predicted signal peptide for secretion (identified using SingalP). BlastX and functional domain annotation show that the most highly transcribed CAZyme genes in the digestive gland encode putative glycoside hydrolases (GHs) belonging to GH9, GH45, GH1, GH13, GH2, GH18, GH31, GH5, GH10, and GH38 families (Fig. 1d, Additional file 1: Table S2).

Very few bacterial transcripts were detected in the cecum samples, confirming previous reports of the virtual absence of live bacteria in this organ [5]. Only two sequences sharing similarity to putative GH30s from *Bacillus* species were found to be expressed at relatively high levels in the cecum (Fig. 1e, Additional file 1: Table S3). However, manual sequence alignment revealed that the two contigs are actually part of one unique transcript featuring a putative polyadenylation (polyA) tail at the 3′ terminus, and have orthologues in the annotated genomes of model bivalve molluscs (XP_011456351.1 from *Crassostrea gigas*, XP_021348230.1 from *Mizuhopecten yessoensis*), whose intron-exon structure strongly suggests an endogenous nature. This GH30 gene might thus be the result of an ancient horizontal gene transfer (HGT) from bacteria, a phenomenon previously observed for several GH families in multiple invertebrate genomes [7]. The shipworm cecum also features high expression of endogenous GH1s, GH13s, and GH2s (Fig. 1e, Additional file 1: Table S3).

RNA-Seq data show that virtually all bacterial genes found in the shipworm meta-transcriptome are transcribed in the gills, with the most abundant being from GH family 6, 11, 5 and auxiliary activity (AA) family 10 lytic polysaccharide monooxygenases (LPMOs) (Fig. 1f, Additional file 1: Table S4). All these bacterial CAZymes carry a putative N-terminal signal peptide likely involved in protein translocation through the periplasm for secretion and have best BlastX matches to sequences from gammaproteobacteria of the family Alteromonadaceae (mainly *Teredinibacter* and *Saccharophagus* species). The identification of the gills as the main site of expression of bacterial CAZymes is in line with previous work from the shipworm *B. setacea*, where GH5s and GH6s were found to be the dominant CAZymes produced by gill bacteria [6].

Comparison of the digestive gland transcriptome from *L. pedicellatus* and *Crassostrea gigas*

Crassostrea gigas (Japanese oyster) is a model suspension feeding bivalve with a fully annotated genome and numerous transcriptomic resources readily available through open access databases. We have compared the digestive gland transcriptome of the wood-boring *L. pedicellatus*, with publicly available data from *C. gigas*, in order to try and pinpoint the enzymes that shipworms have uniquely recruited for wood digestion. The results show that, while putative endo β-1,4-glucanases (GH9s and GH45s) and β-glucosidases (GH1s) together account for over 70% of all CAZymes expressed in the digestive gland of the shipworm (Fig. 2a), these classes are much less abundantly expressed in the oyster (Fig. 2b). In contrast, the oyster transcriptome shows a greater abundance of putative endo β-1,3-glucanases (GH16), α-L-fucosidases (GH29), α-galactosidases (GH27), β-galactosidases (GH35), α-mannosidases (GH38, GH47), xylanases (GH30), β-xylosidases (GH3) (Fig. 2b) and aryl-sulfatases (Additional file 1: Figure S1), compatible with the digestion of polysaccharides (such as mannans, laminarin, xylan, sulfated fucans, and galactans) abundant in the cell walls of phytoplankton [9], which represents the staple diet of *C. gigas* [10].

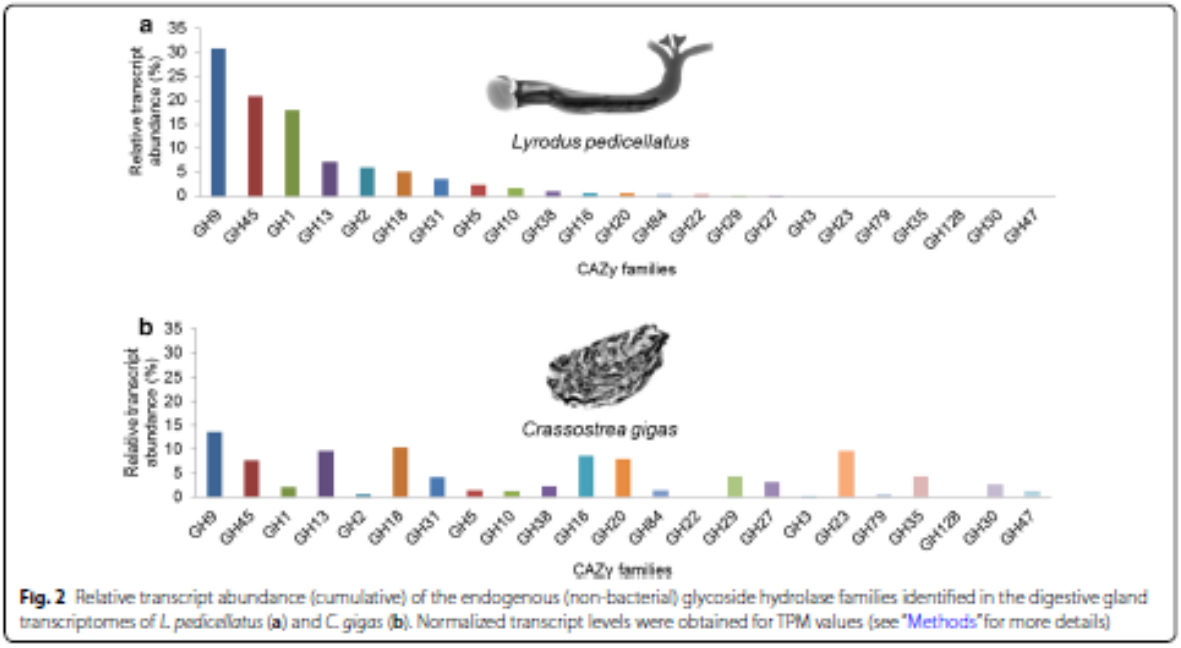


Fig. 2 Relative transcript abundance (cumulative) of the endogenous (non-bacterial) glycoside hydrolase families identified in the digestive gland transcriptomes of *L. pedicellatus* (**a**) and *C. gigas* (**b**). Normalized transcript levels were obtained for TPM values (see "Methods" for more details)

Proteomic analysis of the shipworm cecum content

Previous work on *B. setacea* concluded that most lignocellulolytic enzymes in the shipworm digestive system were of bacterial origin [6]. The authors, however, did not take into account the contribution of endogenous enzymes by the animal itself. By carrying out shotgun proteomics on the total protein extract from the cecum content, we found that CAZymes represent 25% of the total cecum proteome (Fig. 3), while the remaining 75% includes abundant proteases, immunity-related and structural proteins (data not shown). Our analysis shows that less than 15% of the CAZymes detected in the cecum of *L. pedicellatus* are bacterial, while over 85% are endogenously produced by the animal (Fig. 3). Interestingly, abundant GHs identified in the cecum proteome usually correspond to the most highly transcribed genes in the digestive gland (Fig. 1d), suggesting that the mature proteins are secreted and transported by ciliary tracts to the cecum. GH1 represents the dominant enzyme family in the cecum proteome, and mostly occurs as multi-modular proteins, with domains connected by short peptide linkers. The largest multi-domain GH1 (~300 kDa), identified from both transcriptome and proteome, appears as the predominant band in SDS-PAGE analysis of the crude cecum extract, and its identity was confirmed by tryptic digestion and MALDI-MS/MS analysis (Additional file 1: Figure S2). This sequence is specifically expressed in the digestive gland and bears similarity to lactase phlorizin

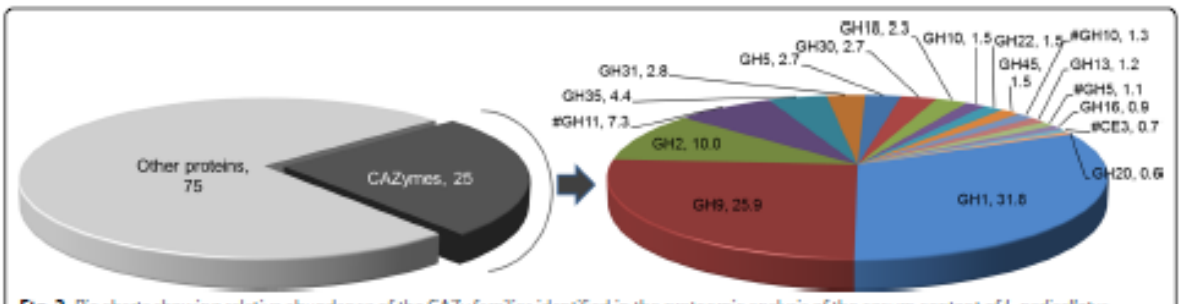


Fig. 3 Pie charts showing relative abundance of the CAZy families identified in the proteomic analysis of the cecum content of *L. pedicellatus*. Enzyme families where all members are of bacterial origin are marked with a pound sign (#). Numbers indicate the percentage of molar abundance derived from cumulative emPAI values

hydrolase (LPH), an enzyme that is localized at the intestinal brush border membrane in mammals. LPH comprises four distinct GH1 domains and mainly exhibits lactase activity [11].

Interestingly, although GH9s and GH1s are abundant in both transcriptome and proteome, the relative abundance of GH45s is high in the transcriptomic data but lower in the proteome, where it accounts for only 1.5% in molar percentage of total CAZymes identified. The major bacterial contributions to the proteome are provided by GH11, 10 and 5 proteins, which typically function as xylanases, mannanases and endo-glucanases. In *B. setacea*, bacterial GH5s and GH6 were reported to account for over 30% of the total protein content of the cecum [6], but here only make up 0.5% of the total proteome (less than 2% of all CAZymes).

Isolation and characterization of the multi-domain GH1 from *L. pedicellatus* (*Lp*MDGH1)

Our combined transcriptomic and proteomic analyses show that a novel multi-domain GH1 (*Lp*MDGH1) is among the most highly expressed sequences in the digestive gland and represents the most abundant CAZyme in the shipworm digestive system (over 3% by molar abundance and 20% by mass) and likely plays an important role in wood digestion. In order to verify that the identified sequence is not an artifact of the de novo transcriptome assembly, we cloned the full-length cDNA of *Lp*MDGH1 and confirmed that it comprises a single open reading frame coding for a polypeptide of 2752 amino acids (Additional file 1: Text S1). While the mammalian lactase phlorizin hydrolase (LPH) has four GH1 domains in the immature protein followed by a transmembrane sequence [11], the shipworm gene encodes an N-terminal signal peptide, followed by six GH1 domains and no transmembrane sequence (Fig. 4a), confirming our proteomic observations of a soluble extracellular protein. In mammals, the first half of the LPH protein has been shown to act as a chaperone that facilitates the folding of the second half [12]. The mammalian LPH undergoes several post-translational modifications, and only two domains (3 and 4) are found in the mature protein. In contrast, the shipworm MDGH1 mature protein, found in the cecum, retains all six GH1 domains, as confirmed by MALDI-MS/MS analysis (Additional file 1: Figure S2) and size on SDS-PAGE gels (Additional file 1: Figure S2).

A modular protein (*Cj*Cel1A) comprising 2 sequential GH1s, reminiscent of the mammalian LPH, was shown to be produced in the digestive gland of the clam *Corbicula japonica* [13]. Based on amino acid sequence, it was hypothesized that the anterior part (first GH1 domain) of the protein from *Corbicula* is not active as a glycoside hydrolase and might instead work as a chaperone, in a similar fashion to the mammalian LPH [12]. Alignment of the six putative GH1 modules of the *L. pedicellatus* protein reveals that domains 2, 4, 5 and 6 possess the required amino acids for hydrolytic activity (regions "NE" and "TENG" in the protein alignment), while domains 1 and 3 lack these residues, are unlikely to have GH activity, and might thus be involved in protein folding or perhaps substrate interactions (Additional file 1: Figure S3).

A BlastP search of the full-length *Lp*MDGH1 against non-redundant (nr) databases finds best matches among molluscs (e.g., *Lottia gigantea*) and reveals that LPH-like sequences are also present in insects, reptiles, birds, amphibians, mammals and fish (Fig. 4b), but not in bacteria, fungi and some animal taxa (e.g., crustaceans). The multi-domain GH1s from vertebrates typically feature a putative C-terminal transmembrane region (Additional file 1: Figure S4), likely involved in anchoring the mature protein to the outer face of the cell plasma membrane. Although our analysis indicates that the lack of this transmembrane region is a common trait among multi-domain GH1s from invertebrates (Additional file 1: Figure S4), the six-domain architecture appears unique to the shipworm protein and might, therefore, represent a specific adaptation towards wood digestion.

Despite our best efforts, we could not obtain soluble *Lp*MDGH1 (nor any of its GH1 domains separately) in the heterologous expression systems we tested. However, we carried out size-exclusion chromatography of soluble cecum extracts and successfully isolated the mature *Lp*MDGH1 to high purity. Zymograms and in vitro assays with the purified enzyme showed that it is the major β-glucosidase in *L. pedicellatus* cecum (Fig. 4c). Interestingly, the enzyme showed activity towards both short-chain gluco-oligosaccharides and long-chain glucans (glucomannan, β-glucan, and lichenan), with preference towards β-1,4 linkages, suggesting roles in the digestion of cellulose and hemicelluloses (Fig. 4d). Such a release of sugars from complex glucans was not seen in assays with a commercially sourced single-module GH1 from *Agrobacteria* (data not shown) and might be an unusual feature of the multi-modular protein. MALDI-TOF MS analysis of the reaction products revealed that *Lp*MDGH1 releases medium- and long-chain oligosaccharides from glucans, suggesting that this enzyme might also have endo-glucanase activity (Additional file 1: Figure S5). Interestingly, a soluble 210 kDa enzyme isolated from the digestive fluids of the sea hare *Aplysia kurodai* was shown to share high similarity with the human lactase phlorizin hydrolase (based on N-terminal protein sequencing) [14]. Although the authors did not manage to isolate the coding sequence nor localize the organ where it was produced, they showed that the purified enzyme can hydrolyze gluco-oligosaccharides as well

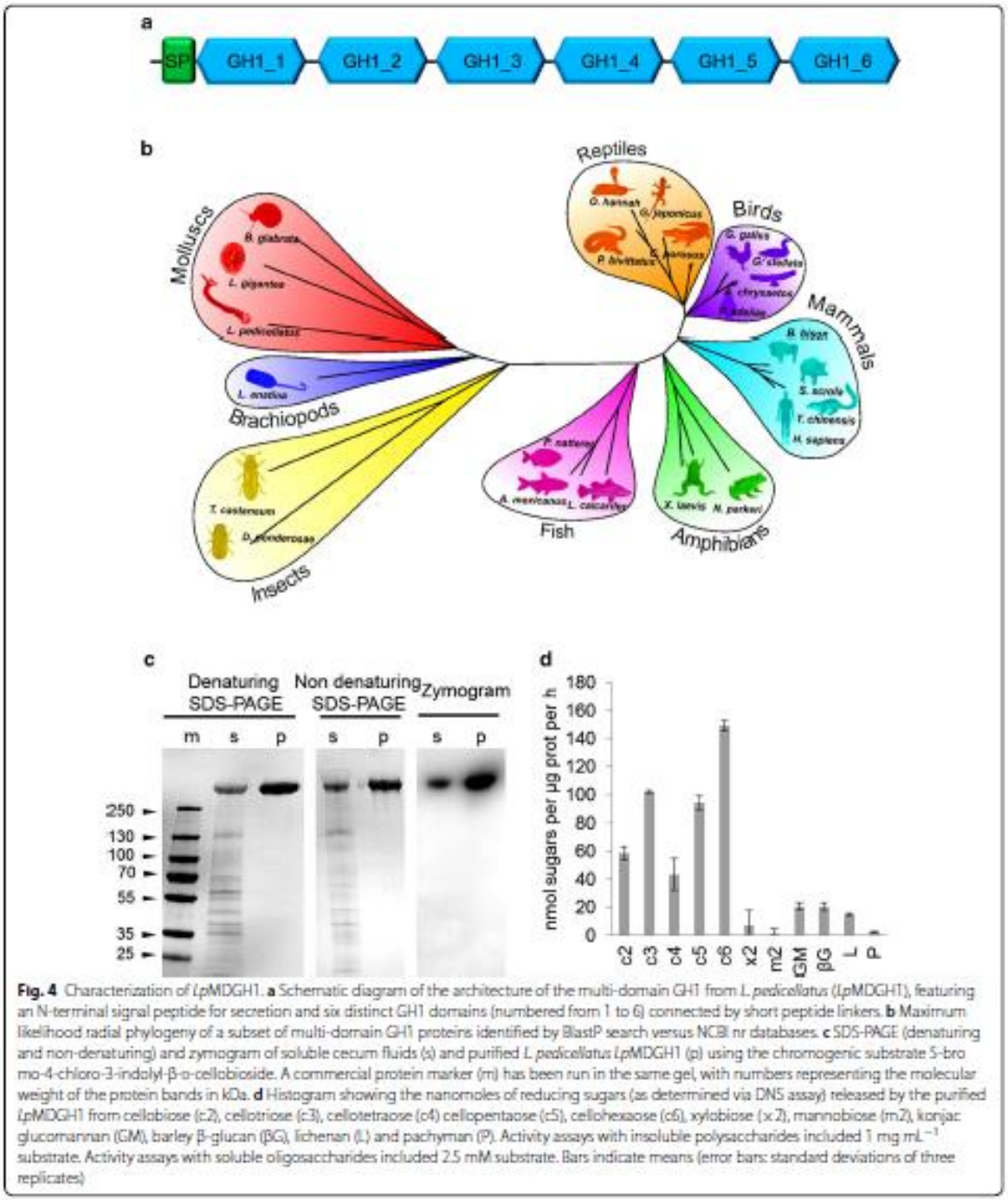


Fig. 4 Characterization of *Lp*MDGH1. **a** Schematic diagram of the architecture of the multi-domain GH1 from *L. pedicellatus* (*Lp*MDGH1), featuring an N-terminal signal peptide for secretion and six distinct GH1 domains (numbered from 1 to 6) connected by short peptide linkers. **b** Maximum likelihood radial phylogeny of a subset of multi-domain GH1 proteins identified by BlastP search versus NCBI nr databases. **c** SDS-PAGE (denaturing and non-denaturing) and zymogram of soluble cecum fluids (s) and purified *L. pedicellatus Lp*MDGH1 (p) using the chromogenic substrate 5-bromo-4-chloro-3-indolyl-β-D-cellobioside. A commercial protein marker (m) has been run in the same gel, with numbers representing the molecular weight of the protein bands in kDa. **d** Histogram showing the nanomoles of reducing sugars (as determined via DNS assay) released by the purified *Lp*MDGH1 from cellobiose (c2), cellotriose (c3), cellotetraose (c4) cellopentaose (c5), cellohexaose (c6), xylobiose (x2), mannobiose (m2), konjac glucomannan (GM), barley β-glucan (βG), lichenan (L) and pachyman (P). Activity assays with insoluble polysaccharides included 1 mg mL^{-1} substrate. Activity assays with soluble oligosaccharides included 2.5 mM substrate. Bars indicate means (error bars: standard deviations of three replicates)

as complex polysaccharides (lichenan, laminarin, and cardran), thus suggesting a key role in the digestion of sea lettuce by the sea hare [14]. The protein size, in vitro activity, and sequence similarity to mammalian LPH suggest that the enzyme from the sea hare has four GH1 domains and plays a similar role to the MDGH1 from *L. pedicellatus*.

Anatomy of the digestive system

Visible light and electron microscopy analysis of sections of the shipworm's digestive system show that wood particles coming from the grinding action of the valves accumulate in the cecum, where most lignocellulose breakdown is thought to occur [2, 5] . Examination of the cecum luminal walls via electron microscopy revealed high abundance of microvilli and cilia at the apical surface of the cells (Fig. 5a, b), confirming the previously hypothesized role of the cecum in agitating food particles and in the absorption of breakdown products (sugars) from wood digestion [2]. This absorptive function is supported by the abundant expression of putative glucose transporters (solute carrier family 2 transporters and sodium-dependent glucose transporters) in the cecum tissues (Additional file 1: Figure S6).

Although the cecum contains most of the ingested wood particles, the digestive gland has long been hypothesized to be involved in production of digestive enzymes in shipworms and other molluscs [15–22], and our meta-transcriptomic and proteomics data confirm this is the case for *Lyrodus*. The gland has a lobular structure

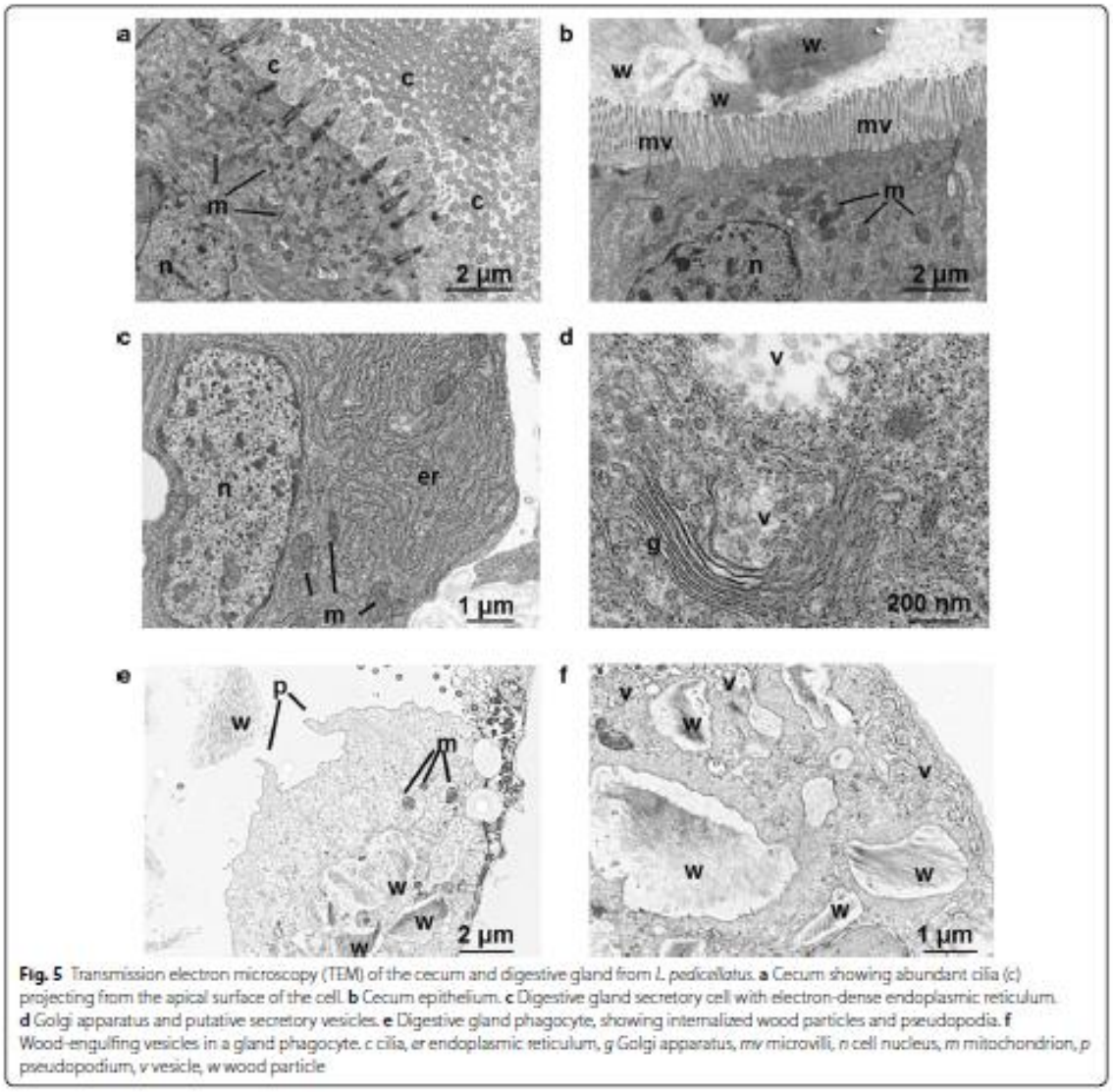


Fig. 5 Transmission electron microscopy (TEM) of the cecum and digestive gland from *L. pedicellatus*. **a** Cecum showing abundant cilia (c) projecting from the apical surface of the cell. **b** Cecum epithelium. **c** Digestive gland secretory cell with electron-dense endoplasmic reticulum. **d** Golgi apparatus and putative secretory vesicles. **e** Digestive gland phagocyte, showing internalized wood particles and pseudopodia. **f** Wood-engulfing vesicles in a gland phagocyte. *c* cilia, *er* endoplasmic reticulum, *g* Golgi apparatus, *mv* microvilli, *n* cell nucleus, *m* mitochondrion, *p* pseudopodium, *v* vesicle, *w* wood particle

reminiscent of secretory organs in other animals and is directly connected to the stomach by ducts [4, 22]. Our high-resolution microscopic analyses reveal that the gland contains secretory cells, as previous observed in other molluscs [19–21]. These specialized cells occupy the crypts of the tubule, are mostly pyramidal, and have a well-developed granular endoplasmic reticulum (Fig. 5c) and an extensive Golgi apparatus, producing numerous vesicle-like structures (Fig. 5d) of variable sizes potentially containing glycoside hydrolases and other secretory enzymes identified in our studies. The digestive gland also features abundant amoeboid cells (phagocytes, ~20 µm in diameter) with pseudopodia and internalized wood particles of variable dimensions (Fig. 5e, f). EM images show numerous vesicles budding from the Golgi apparatus and apparently fusing with the cell wall of the cavities containing the internalized wood fragments (Fig. 5f), suggesting the formation of lysosomes that might be responsible for intracellular wood breakdown. Although wood phagocytes have been previously reported in shipworms with hand drawings [22], this is the first high-resolution image of these cells and provides new evidence of their role in lignocellulose digestion. This suggests that, while the cecum provides the major site of wood digestion, it may be supplemented by intracellular digestion. Indeed, the cecum appears to be a specific adaptation of shipworms and has not been reported in other bivalves, where most digestive processes are restricted to the gland and intestinal systems.

Discussion

Lignocellulose represents the most abundant and ubiquitous organic material in nature. The breakdown of this recalcitrant polymer plays a critical role in the global carbon cycle, and is attracting growing interest from a biotechnological perspective. As society moves away from the use of net greenhouse gas-emitting fossil resources, the use of surplus woody biomass to provision fuels, chemicals, and materials is becoming imperative. Understanding the digestive systems of major wood-digesting animals provides insights and enzymes that could help towards the cost-effective breakdown of lignocellulosic biomass into simple sugars and other building blocks. We have undertaken a detailed and multifaceted study of the digestive system of *L. pedicellatus*, and revealed the complex molecular mechanisms of lignocellulose digestion in shipworms. Our work shows that the digestive gland of *L. pedicellatus* produces a complex enzymatic mixture containing most of the activities required for the digestion of the plant cell wall, including cellulases (endo-β-1,4-glucanases, β-glucosidases) and hemicellulases (β-mannanases/mannosidases, β-xylanases/xylosidases). Although the meta-transcriptome of *L. pedicellatus* lacks endogenous cellobiohydrolases (CBHs), it contains bacterial GH6s (which can act as CBHs) expressed in the gills, as previously observed in *B. setacea* [6]. Similarly, we also detected expression of bacterial lytic polysaccharide monooxygenases (LPMOs), which likely synergize endogenous as well as bacterial glucanases. Indeed, previous work has shown that LPMOs can insert breaks into highly crystalline polysaccharides and, by doing so, boost the activity of glycoside hydrolases by several orders of magnitude [23–25]. The surprisingly low levels of both GH6 and LPMO mature proteins in the shipworm cecum, however, suggest that the corresponding transcripts might not be translated efficiently, or that the mature enzymes are unstable in the shipworm digestive tract. In support of this, our cecum proteomics analysis revealed abundant endogenous proteases, which might reduce the half-life of some bacterial enzymes.

Our work in *L. pedicellatus*, and previous studies on wood-feeding cockroaches, beetles, and termites [26–28], suggest that a combination of endogenous and symbiotic enzymes is optimal for efficient plant cell wall digestion in invertebrates. Interestingly, the genomes of insects, crustaceans, annelids and molluscs encode numerous enzymes involved in plant cell wall digestion, implying that some of these genes were present in the last common ancestor of bilaterian animals [29], before bacterial symbioses developed, and likely represent an ancestral mechanism for lignocellulose digestion. Our analysis of the digestive gland transcriptome from shipworm and oyster confirms that molluscs share a complex array of endogenous lignocellulolytic enzymes, and that their expression levels are adapted to their specific diets. *C. gigas* (oyster) is characterized by high expression of glycoside hydrolases and sulfatases involved in the deconstruction of sulfated polysaccharides that are abundant in marine algae (as an adaptation to highly ionic environment) [30] but not in fresh water algae and terrestrial plants. Our data show that *L. pedicellatus* relies mostly on GH9s, GH45s, and GH1s to breakdown terrestrial woody plants, and the same is likely true for most shipworm species, which typically feed on submerged wood. There are, however, some notable exceptions. For example, *Zachsia zenkewitschi* feeds on the rhizomes of seagrasses such as *Zostera*, one of the few examples of marine angiosperms to have regained the ability to produce sulfated polysaccharides [31]. Even more puzzling is the giant mud-boring teredinid *Kuphus polythalamia*, where sulfur-oxidizing bacteria have replaced the ancestral cellulolytic symbionts in the gills [32]. Further work is needed to elucidate the function of the digestive gland in this enigmatic chemoautotrophic bivalve, which represents a unique example of a shipworm that has entirely lost the ability to digest plant biomass.

Previous studies in the shipworm *B. setacea* suggested that glycoside hydrolases produced by endosymbiotic bacteria located in the gills dominate the shipworm digestive system [6]. The apparent discrepancies in results between our study and that by O'Connor et al. [6] may be in part due to differences between species. Both belong to the family Teredinidae, rely on a diet of wood, and share some key anatomical features, including digestive glands, extended cecum, and the presence of analogous bacterial populations in the gill bacteriocytes (mostly gammaproteobacteria related to *Saccharophagus degradans*). Therefore, it would be of interest to determine what factors underpin the differences between our results and those reported by O'Connor et al. [6]. It is worth noting that in the study by O'Connor et al. [6] the cecum proteomic data were reported to be searched specifically against bacterial DNA extracted from the gills, which would have precluded identification of proteins produced by the animal itself. Indeed, our proteomics studies reveal that the bulk of the CAZymes found in the cecum of *L. pedicellatus* is secreted by a specialized digestive gland, while bacterial enzymes coming from the gills play a supporting role. The size of the cecum and abundance of GHs found there suggest that this is the major site of wood digestion in shipworms. Yet, our ultrastructural studies indicate a potential role in wood digestion for ameboid cells in the digestive gland. Although intracellular wood digestion is unusual, it has also been inferred from microscopic analysis of commensal ciliates found in the digestive system of lower termites, which appear to engulf wood particles [33].

By investigating the cecum meta-proteome, we have discovered that the most abundant enzyme in the *L. pedicellatus* digestive tract is an unusual multi-modular GH1 with sequence similarity to the lactase phlorizin hydrolase found in mammals. In contrast to the mature peptide in the mammalian LPH, which has two GH1 domains linked by a short peptide and is mostly active as a β-glucosidase, the *L. pedicellatus* MDGH1 has six domains, and our studies have revealed its specificity against β-linked glucosides as well as complex glucans, suggesting the ability to cleave these linkages in cellulose and glucomannans during wood digestion. Based on the presence of key amino acid residues predicted to be involved in substrate attack, we would expect four of the six GH1 domains in the *Lp*MDGH1 to be active glycoside hydrolases. Future work is needed to clarify if the endo- and exo-activities observed in *Lp*MDGH1 depend on the distinct GH1 domains or rather on the fusion nature of the polyprotein, and whether the two putatively inactive GH1 domains play any role in the two mechanisms of action.

The sustainability of biorefineries hinges on the identification of the most effective enzymatic cocktails for the saccharification of plant biomass. Our investigation into the digestive system of *L. pedicellatus* uncovers a wide range of new glycoside hydrolases attacking major fractions of lignocellulose (cellulose, xylans, and mannans) and unprecedented information regarding their relative abundance, which could help engineer an optimal enzymatic cocktail for the breakdown of lignocellulose. In this context, the identification of *Lp*MDGH1 as the major enzyme in the shipworm's cecum is particularly interesting, as one of the major bottlenecks in the industrial breakdown of plant biomass is the ability to prevent the accumulation of cellobiose, a potent inhibitor of endoglucanases and cellobiohydrolases [34]. Shipworms seem to have overcome this issue by mass producing a unique multi-domain hydrolase with dual activity towards long-chain glucans and cellobiose, an elegant evolutionary solution that may help simplify the enzymatic cocktails used in cellulosic biorefineries.

Conclusions

This work is the first comprehensive investigation of the complex molecular and physiological processes by which shipworms extract nutrients from wood and play a fundamental role in the global carbon recycling. The identification of the key enzymes produced by the shipworm *L. pedicellatus*, and the in vitro characterization of the most abundant glycoside hydrolase found in its digestive system, may open up new opportunities in the biotechnological deconstruction of lignocellulose in support of a sustainable bio-economy.

Methods

Substrates

Phosphoric acid swollen cellulose (PASC) was prepared as follows. 5 g of Avicel® PH-101 was moistened with water and treated with 150 mL ice cold 85% phosphoric acid, stirred on an ice bath for 1 h. Then 500 mL cold acetone was added while stirring. The swollen cellulose was filtered on a glass-filter funnel and washed 3 times with 100 mL ice cold acetone and subsequently twice with 500 mL water. PASC was then suspended in 500 mL water and blended to homogeneity.

High-purity pachyman (β-D-1,3-glucan), barley β-glucan (β-D-1,3-1,4-glucan), lichenan (from Icelandic moss, β-D-1,3-1,4-glucan), mannan (borohydride reduced), konjac glucomannan (β-D-1,4), carob galactomannan, larch arabinogalactan, wheat arabinoxylan, cellotriose, cellotetraose, cellopentaose, cellohexaose, mannobiose, and xylobiose were purchased from Megazyme. Locust bean gum, carboxymethyl cellulose (CMC),

beechwood xylan, and cellobiose were purchased from Sigma. 5-bromo-4-chloro-3-indolyl-β-d-cellobioside was purchased from Santa Cruz Biotechnology.

Specimen collection

Adult *Lyrodus pedicellatus*, matching the sequences of cytochrome c oxidase subunit I and small subunit rRNA 18S of the Atlantic population of *Lyrodus* as per Borges et al. [35], were obtained from an infested pier composed of greenheart wood (*Chlorocardium rodiei*) at Portsmouth, UK. Adults were harvested and dissected in order to yield the larvae. Subsequent cultures were reared in aquaria at the Institute of Marine Sciences, University of Portsmouth. Seawater was taken directly from Langstone Harbour, maintained at a temperature between 15–18 °C and a salinity of 33 PSU, and kept aerated throughout. Tanks were regularly provided with small panels of Scot pine wood for larval settlement.

mRNA extraction, preparation, and sequencing

Digestive gland, cecum and gills were dissected from three healthy adult *L. pedicellatus* and total RNA was extracted using TRIzol® Reagent (Thermo Fisher Scientific). Samples were DNase treated using Turbo DNA-free (Ambion) before quantification using a Qubit Fluorometer (Thermo Fisher Scientific). Ribosomal RNA depletion was carried out with a RiboZero™ Magnetic Gold Kit (Epidemiology) (Epicentre). mRNA was then concentrated using RNA Clean & Concentrator™-5 (Zymo Research). RNA-Seq libraries were prepared from each mRNA sample as per the Ion Total RNA-Seq kit v2 (Thermo Fisher Scientific), using an RNaseIII treatment time of 2.5–3 min. Samples were barcoded using the Ion Xpress RNA-Seq Barcode kit (Thermo Fisher Scientific). Yields and library sizes were assessed using the High Sensitivity D1K screentapes and reagents on a 2200 TapeStation Nucleic Acids System (Agilent Technologies). Appropriately diluted library aliquots were combined in pairs in equimolar amounts and used for template preparation using the Ion OneTouch 200 Template Kit v2 DL on a OneTouch system (Thermo Fisher Scientific) prior to loading onto a 318 chip and sequenced on an Ion Torrent PGM™ prepared as per the manufacturer's instructions (IonPGM200Kit; Thermo Fisher Scientific). All raw sequence data were deposited in NCBI under BioProject PRJNA412369 (SRA files: SRR6106265, SRR6106266, SRR6106267, SRR6106268, SRR6106269, SRR6106270, SRR6106271, SRR6106272, SRR6106273). Assembled contigs are available from the authors upon request.

Transcriptome assembly, sequence annotation, and identification of putative CAZymes

After removing the primer sequences and low-quality reads from raw EST sequencing reads, the EST sequences from three tissues (digestive gland, cecum, and gills) from three healthy adults were assembled into unigene contigs using Trinity [36]. The contigs from the three animals were then assembled into supercontigs with the CAP3 DNA Sequence Assembly Program [37]. Raw reads were mapped onto the transcriptomes of the three molluscs and normalized expression values (TPM = Transcripts per kilobase Million) were calculated for each transcript using Salmon (part of the Galaxy toolshed) [38]. Although all three animals provided raw reads with high quality and could be used to assemble a reference transcriptome, TPM values from one animal were found to be poorly correlated with the other two (possibly as a result of illness or distress) and were therefore excluded from the following analysis. Average TPM values for each contig in the three tissues (digestive gland, cecum, and gills) were thus calculated from the two healthy animals. The top 10,000 assembled contigs (ranked according to the cumulative number of mapped ESTs from all organs) were annotated with the BLASTx algorithm [39] to search against non-redundant (nr) peptide database downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). CAZy annotation was carried out using the CAZYmes Analysis Toolkit (CAT) on the BioEnergy Science Center website (http://mothra.ornl.gov/cgi-bin/cat/cat.cgi). Sequences annotated as glycosyltransferases (GTs) and carboxyl esterases, mostly involved in intracellular processes not relevant to lignocellulose digestion, were excluded from the analysis. Among auxiliary activity (AA) families, only those clearly involved in polysaccharide degradation were considered (LPMO families AA9, AA10, AA11, and AA13).

Contigs were converted to putative ORFs using the online tool Emboss (http://www.bioinformatics.nl/cgi-bin/emboss/getorf), putative N-terminal signal peptides were predicted with SingalP (http://www.cbs.dtu.dk/services/SignalP/), and putative transmembrane regions were predicted using TMHMM (http://www.cbs.dtu.dk/services/TMHMM/).

Raw transcriptome sequencing data for the digestive gland of a wildly caught *C. gigas* (Japanese oyster) were retrieved from the EBI portal (run Accession SRR334213) [40]. The published transcriptome of *C. gigas* (based on the annotated genome) was retrieved from NCBI (Accession PRJNA276446). Raw reads were mapped onto the transcriptome of the mollusc and normalized expression values were calculated for each transcript using Salmon (part of the Galaxy toolshed) [38]. The identity and

relative abundance of the CAZyme families were then compared to those obtained from the digestive gland of *L. pedicellatus*. CAZy annotation was carried out using the CAZYmes Analysis Toolkit (CAT) on the BioEnergy Science Center website (http://mothra.ornl.gov/cgi-bin/cat/cat.cgi).

Proteomics analysis

The ceca from five animals grown on Scots pine were dissected in 50 mM sodium phosphate buffer pH 7 and the content (food particles and enzymes) was collected, pooled together, added with 1% SDS, beta-mercapto ethanol, DTT, boiled for 10 min, centrifuged, and the supernatant run into a 10% polyacrylamide gel to a depth of 1 cm, before staining with Coomassie.

In-gel tryptic digestion was performed post reduction with DTE and *S*-carbamidomethylation with iodoacetamide. Resulting peptides were analyzed 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 [41]. Protein identification was performed by searching tandem mass spectra against the assembled transcriptome of *L. pedicellatus* using the Mascot search program. Matches were passed through Mascot percolator to achieve a false discovery rate of < 1% and further 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 [42]. Proteins identified in the proteomics analysis were annotated via Blastx versus non-redundant NCBI databases. CAZy annotation was carried out using the CAZYmes Analysis Toolkit (CAT) on the BioEnergy Science Center website (http://mothra.ornl.gov/cgi-bin/cat/cat.cgi).

One aliquot of the cecum extract (content only) and the purified *Lp*MDGH1 was added with 1% SDS, beta-mercapto ethanol, DTT, boiled for 10 min, centrifuged, and the supernatant run in a 4–20% gradient polyacrylamide gel. After Coomassie staining, the most abundant band (with an approximate molecular weight of 300 kDa) was excised and in-gel digested as described for LC-MS/MS samples.

A 1 µL aliquot of peptide mixture was applied directly to a ground steel MALDI target plate and overlaid with an equal volume of a 5 mg/mL 4-hydroxy-α-cyano-cinnamic acid in 50% aqueous (v:v) acetonitrile containing 0.1% trifluoroacetic acid (v:v). Positive-ion MALDI mass spectra were obtained using a Bruker Ultraflex III in reflectron mode, equipped with a Nd:YAG smart beam laser. MS spectra were acquired over a range of m/z 800–4000. The ten most intense precursors with S/N greater than 30 were selected for MS/MS fragmentation in LIFT mode without collision gas. The default calibration was used for MS/MS spectra, which were baseline-subtracted and smoothed (Savitsky–Golay, width 0.15 m/z, cycles 4); monoisotopic peak detection used a SNAP averaging algorithm (C 4.9384, N 1.3577, O 1.4773, S 0.0417, H 7.7583) with a minimum S/N of 6. Bruker flexAnalysis software (version 3.3) was used for spectral processing and peak list generation. Tandem mass spectral data were submitted to database searching using a locally running copy of the Mascot program (Matrix Science Ltd., version 2.5), through the Bruker ProteinScape interface (version 2.1). Search criteria included enzyme, trypsin; fixed modifications, carbamidomethyl (C); variable modifications, oxidation (M), deamidated (N, Q); peptide tolerance, 100 ppm; MS/MS tolerance, 0.5 Da; Instrument, MALDI-TOF–TOF. Peptide matches were filtered to require expect scores of 0.05 or better.

Cloning the *Lp*MDGH1 cDNA

The native sequence (from start to stop codon) for *Lp*MDGH1 was cloned from cDNA generated from RNA extracted from the digestive gland of *L. pedicellatus* using external oligonucleotide primers designed on the assembled contig from the transcriptome. Total RNA was extracted from one animal using the TRIzol® Reagent (Thermo Fisher Scientific) and cDNA was generated with an oligodT primer using SuperScript® II reverse transcriptase (Thermo Fisher Scientific). PCR reactions were then set up using Phusion® High-Fidelity DNA Polymerase (Thermo Fisher Scientific) and the amplicon was cloned into an auxiliary plasmid using the StrataClone Blunt PCR Cloning Kit (Stratagene) and the correct sequence was verified with the Sanger method using internal primers. Open reading frames (ORFs) were calculated using the online EXPASY tool "Translate" and confirmed that the cloned sequence codes for a unique polypeptide of 2752 amino acid residues (without internal stop codons). The sequence has been deposited in GenBank with Accession No. MG013499.

Phylogeny and sequence analysis of *Lp*MDGH1

A protein sequence alignment of the single GH1 domains from *Lp*MDGH1 was obtained using T-Coffee [43] and visualized using JalView [44].

The *Lp*MDGH1 protein sequence was searched via BlastP against NCBI non-redundant databases and orthologues from molluscs, insects, fish, amphibians, reptiles, birds, and mammals were retrieved. The resulting amino acid sequences were aligned using Muscle [45], operating with default parameters. A distance matrix was made with Mega6 [46] using a Jones–Taylor–Thornton matrix and a phylogenetic tree was then calculated by the

maximum likelihood algorithm and standard parameters. The resulting tree was visualized using Dendroscope [47].

Purification of *Lp*MDGH1

The ceca of twenty *L. pedicellatus* specimens were dissected in 50 mM sodium phosphate buffer pH 7 and the content (wood particles and enzymes) was collected and centrifuged. The supernatant was then filtered with 0.22-µm syringe filters and applied to a Superose™ 6 Increase 10/300 GL size-exclusion chromatography column (GE Healthcare) pre-equilibrated with 20 mM Tris-HCl pH 8 plus 100 mM NaCl. Eluted fractions were analyzed by denaturing SDS-PAGE and those corresponding to the *Lp*MDGH1 were pooled, concentrated using Microsep™ Advance Centrifugal Filters (Pall Laboratory, 100 kDa cut-off), and re-applied to the same column pre-equilibrated with 20 mM sodium phosphate buffer pH 7 plus 100 mM NaCl. Protein purity was again assessed via SDS-PAGE analysis.

Enzymatic assays and zymograms

Activity of the cecum fluids on a panel of polysaccharides and oligosaccharides was determined by DNS reducing sugar assay [8]. Briefly, ten ceca were dissected in 50 mM sodium phosphate buffer pH 7 and the content fully re-suspended by pipetting. After centrifugation, the soluble portion (supernatant) was filtered through 0.22-µm porous membranes, quantified with the Bradford [48] reagent and used for assays. 50 µL reactions were carried out in 96-well plates in 50 mM sodium phosphate buffer pH 6 with either 1187 ng of total soluble cecum protein or 237 ng of purified GH1, and 1 mg mL^{-1} polysaccharide or 2.5 mM oligosaccharide. All reactions, including controls, were performed in triplicate. The micro-plate was incubated at 28 °C shaking at 320 rpm for 24 h, and then 100 µL of DNS reagent was added to each reaction before heating at 100 °C for 5 min. Absorbance at 540 nm was measured with a micro-plate reader and nanomoles of reducing sugars released were determined based on absorbance obtained with glucose standards. The DNS reagent was prepared by mixing 0.75 g of dinitrosalicylic acid, 1.4 g NaOH, 21.6 g sodium potassium tartrate tetrahydrate, 0.53 mL phenol, and 0.59 g sodium metabisulfite in 100 mL pure water.

Zymograms were performed as follows. 1.9 µg of total protein from the cecum fluids and purified *Lp*MDGH1 were run in a non-denaturing SDS-PAGE gel (4–20% Mini-PROTEAN® TGX™ Precast Protein Gel, Biorad). The gel was then incubated in 20 mL of 20 mM sodium phosphate buffer pH 6 plus 2.5% Triton X100 for 30 min, and then washed twice in 20 mL of 20 mM sodium phosphate buffer pH 6 for 10 min. The gel was then incubated for 4 h in 20 mL of 20 mM sodium phosphate buffer pH 6 with 0.01% (w/v) of 5-bromo-4-chloro-3-indolyl-β-D-cellobioside to allow formation of the insoluble dye.

Compositional analysis of wood and frass

Untreated Scots Pine (powdered) and dried frass (obtained from *L. pedicellatus* grown on submerged panels of Scots Pine) were analyzed for cellulose, hemicellulose, and lignin content in five technical replicates each, using the methods reported in Marriot et al. [49].

Transmission electron microscopy

Dissected shipworm tissue (cecum, digestive gland) was fixed for 1–2 h at ambient temperature in primary fixative (4% formaldehyde (w/v), 2.5% (w/v) glutaraldehyde in 100 mM sodium phosphate buffer pH 7.2), and then washed (3 × 10 min) in 100 mM sodium phosphate buffer pH 7.2 before incubation in secondary fixative for 1 h on ice (1% osmium tetroxide in 100 mM sodium phosphate buffer pH 7.2). Samples were dehydrated through a graded ethanol series (15 min each), followed by two washes (5 min each) in epoxy propane. Samples were infiltrated with a series of epoxy propane/Epon araldite (25%, 50%, 75% Epon Araldite with a minimum of 1 h at each stage, all at 30°C) concluding with a minimum of two changes of Epon araldite resin over 24 h at 30 °C, and polymerized at 60 °C for 48 h in flat embedding molds. Pale gold (70–90 nm) ultra-thin sections were cut with a Diatome diamond knife, using a Leica Ultracut UCT microtome, and mounted on hexagonal 200-mesh nickel grids. Sections were post-stained with 2% (w/v) aqueous uranyl acetate (10 min), then with lead citrate (5 min) [50] in a carbon dioxide-free chamber and viewed using a FEI Tecnai 12 BioTWIN G2 TEM operating at 120 kV. Images were captured using AnalySIS software and a Megaview III CCD camera.

Additional file

Additional file 1. Additional tables, figures, and additional text.

Authors' contributions

FS, GP, JRS, SMC, NCB, and SJM designed research. FS, GP, LE, and KB performed research. CS and MS contributed to TEM sample preparations and imaging; AAD and RB contributed to proteomic sample preparation and data analysis; FS, DAR, and YL analyzed transcriptomic data; FS, NCB, and SJM wrote the paper. All authors reviewed and commented on the manuscript. All authors read and approved the final manuscript.

Author details

[1] Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5DD, UK. [2] Bioscience Technology Facility, Department of Biology, University of York, Heslington, York YO10 5DD, UK. [3] Biorenewables Development Centre, 1 Hassacarr Close, Chessingham Park, Dunnington,

York YO19 5SN, UK. [4] Marine Science Center, Northeastern University, Nahant, MA 01908, USA. [5] School of Biological Sciences, University of Portsmouth, King Henry Building, King Henry 1st St, Portsmouth PO1 2DY, UK.

Acknowledgements
This work is funded by the UK Biotechnology and Biological Sciences Research Council (Grant Number BB/L001926/1).

Competing interests
The authors declare that they have no competing interests.

Data availability
The datasets supporting the conclusions of this article are available in NCBI repository, under BioProject PRJNA412369 (SRA files: SRR6106265, SRR6106266, SRR6106267, SRR6106268, SRR6106269, SRR6106270, SRR6106271, SRR6106272, SRR6106273). The *Lp*MDGH1 sequence can be accessed through GenBank Accession No. MG013499.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 2 December 2017 Accepted: 21 February 2018
Published online: 07 March 2018

References

1. Feller IC, Lovelock CE, Berger U, McKee KL, Joye SB, Ball MC. Biocomplexity in mangrove ecosystems. Ann Rev Mar Sci. 2010;2:395–417.
2. Turner RD. A survey and illustrated catalogue of the Teredinidae (Mollusca: Bivalvia). Cambridge: The Museum of Comparative Zoology, Harvard University; 1966.
3. Cragg SM. Marine wood boring invertebrates of New Guinea and its surrounding waters. In: Beehler BM, Marshall AJ, editors. The ecology of Papua. Singapore: Periplus; 2007. p. 539–63.
4. Morton B. The functional anatomy of the organs of feeding and digestion of *Teredo navalis* Linnaeus and *Lyrodus pedicellatus* (Quatrefages). Proc Malac Soc Lond. 1970;39:151–67.
5. Betcher MA, Fung JM, Han AW, O'Connor R, Seronay R, Concepcion GP, Distel DL, Haygood MG. Microbial distribution and abundance in the digestive system of five shipworm species (Bivalvia: Teredinidae). PLoS ONE. 2012;7:e45309.
6. O'Connor RM, Fung JM, Sharp KH, Benner JS, McClung C, Cushing S, Lamkin LR, Fomenkov AI, Henrissat B, Londer YY, Scholz MB, Posfai J, Malfatti S, Tringe SG, Woyke T, Malmstrom RR, Coleman-Derr D, Altamia MA, Dedrick S, Kaluziak ST, Haygood MG, Distel DL. Gill bacteria enable a novel digestive strategy in a wood-feeding mollusk. Proc Natl Acad Sci USA. 2014;111:E5096–104.
7. Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biol Evol. 2016;8:1785–801.
8. Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem. 1959;31:426–8.
9. Pomin VH, Mourão PA. Structure, biology, evolution, and medical importance of sulfated fucans and galactans. Glycobiology. 2008;18:1016–27.
10. Ponis E, Probert I, Veron B, Mathieu M, Robert R. New microalgae for the Pacific oyster *Crassostrea gigas* larvae. Aquaculture. 2006;253:618–27.
11. Naim HY, Naim H. Dimerization of lactase-phlorizin hydrolase occurs in the endoplasmic reticulum, involves the putative membrane spanning domain and is required for an efficient transport of the enzyme to the cell surface. Eur J Cell Biol. 1996;70:198–208.
12. Jacob R, Peters K, Naim HY. The prosequence of human lactase-phlorizin hydrolase modulates the folding of the mature enzyme. J Biol Chem. 2002;277:8217–25.
13. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Molecular cloning of endogenous beta-glucosidase from common Japanese brackish water clam *Corbicula japonica*. Gene. 2009;435:72–9.
14. Tsuji A, Tominaga K, Nishiyama N, Yuasa K. Comprehensive enzymatic analysis of the cellulolytic system in digestive fluid of the Sea Hare *Aplysia kurodai*. Efficient glucose release from sea lettuce by synergistic action of 45 kDa endoglucanase and 210 kDa ß-glucosidase. PLoS ONE. 2013;8:e65418.
15. Sakamoto K, Touhata K, Yamashita M, Kasai A, Toyohara H. Cellulose digestion by common Japanese freshwater clam *Corbicula japonica*. Fish Sci. 2007;73:675–83.
16. Sakamoto K, Toyohara H. Putative endogenous xylanase from brackish-water clam *Corbicula japonica*. Comp Biochem Physiol Part B Biochem Mol Biol. 2009;154:85–92.
17. Sakamoto K, Uji S, Kurokawa T, Toyohara H. Immunohistochemical, in situ hybridization and biochemical studies on endogenous cellulase of *Corbicula japonica*. Comp Biochem Physiol Part B Biochem Mol Biol. 2008;150:216–21.
18. Owen G. The fine structure of the digestive tubules of the marine bivalve *Cardium edule*. Phil Trans Royal Soc Lond. 1970;258:245–60.
19. Pal SG. The fine structure of the digestive tubules of *Mya arenaria L.*: I. Basiphil cell. Proc Malac Soc Lond. 1971;39:303–9.
20. Pal SG. The fine structure of the digestive tubules of *Mya arenaria L.* II. Digestive cell. Proc Malac Soc Lond. 1972;40:161–70.
21. Taleb N. Distribution of digestive tubules and fine structure of digestive cells of *Aplysia punctata* (Cuvier, 1803). J Molluscan Stud. 2000;67:169–82.
22. Potts FA. The structure and function of the liver of *Teredo*, the shipworm (biological sciences). Proc Camb Phil Soc. 1923;1:1–17.
23. Vaaje-Kolstad G, Westereng B, Horn SJ, Liu Z, Zhai H, Sørlie M, Eijsink VG. An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. Science. 2010;330:219–22.
24. Quinlan RJ, Sweeney MD, Lo Leggio L, Otten H, Poulsen JC, Johansen KS, Krogh KB, Jørgensen CI, Tovborg M, Anthonsen A, Tryfona T, Walter CP, Dupree P, Xu F, Davies GJ, Walton PH. Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components. Proc Natl Acad Sci USA. 2011;108:15079–84.
25. Leggio LL, Simmons TJ, Poulsen JC, Frandsen KE, Hemsworth GR, Stringer MA, von Freiesleben P, Tovborg M, Johansen KS, De Maria L, Harris PV, Soong CL, Dupree P, Tryfona T, Lenfant N, Henrissat B, Davies GJ, Walton PH. Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. Nat Commun. 2015;6:5961. https://doi.org/10.1038/ncomms6961.
26. Watanabe H, Tokuda G. Cellulolytic systems in insects. Annu Rev Entomol. 2010;55:609–32.
27. Brune A. Symbiotic digestion of lignocellulose in termite guts. Nat Rev Microbiol. 2014;12:168–80.
28. Franco Cairo JPL, Carazzolle MF, Leonardo FC, Mofatto LS, Brenelli LB, Gonçalves TA, Uchima CA, Domingues RR, Alvarez TM, Tramontina R, Vidal RO, Costa FF, Costa-Leonardo AM, Paes Leme AF, Pereira GAG, Squina FM. Expanding the knowledge on lignocellulolytic and redox enzymes of worker and soldier castes from the lower termite *Coptotermes gestroi*. Front Microbiol. 2016;7:1518. https://doi.org/10.3389/fmicb.2016.01518.
29. Calderón-Cortés N, Quesada M, Watanabe H, Cano-Camacho H, Oyama K. Endogenous plant cell wall digestion: a key mechanism in insect evolution. Annu Rev Ecol Evol Syst. 2012;43:45–71.
30. Popper ZA, Michel G, Hervé C, Domozych DS, Willats WG, Tuohy MG, Kloareg B, Stengel DB. Evolution and diversity of plant cell walls: from algae to flowering plants. Annu Rev Plant Biol. 2011;62:567–90.
31. Shipway JR, O'Connor R, Stein D, Cragg SM, Korshunova T, Martynov A, Haga T, Distel DL. *Zachsia zenkewitschi* (Teredinidae), a rare and unusual seagrass boring bivalve revisited and redescribed. PLoS ONE. 2016. https://doi.org/10.1371/journal.pone.0155269.
32. Distel DL, Altamia MA, Lin Z, Shipway JR, Han A, Fortoza J, Antemano R, Limbaco MGJP, Tabo AG, Dechavez R, Albano J, Rosenberg G, Concepcion GP2, Schmidt EW, Haygood MG. Discovery of chemoautotrophic symbiosis in the giant shipworm *Kuphus polythalamia* extends wooden-steps theory. Proc Natl Acad Sci USA. 2017;114:E3652–8.
33. Kiuchi I, Moriya S, Kudo T. Two different size-distributions of engulfment-related vesicles among symbiotic protists of the lower termite, *Reticulitermes speratus*. Microbes Environ. 2004;19211–4.

34. Sørensen A, Lübeck M, Lübeck PS, Ahring BK. Fungal beta-glucosidases: a bottleneck in industrial use of lignocellulosic materials. Biomolecules. 2013;3:612–31.
35. Borges LMS, Sivrikaya H, le Roux CA, Shipway JR, Cragg SM, Costa FO. Investigating the taxonomy and systematics of marine wood borers (Bivalvia: Teredinidae) combining evidence from morphology, DNA barcodes and nuclear locus sequences. Invertebr Syst. 2012;26:572–82.
36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
37. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9:868–77.
38. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
39. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
40. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z, Que H, Xie Y, Holland PW, Paps J, Zhu Y, Wu F, Chen Y, Wang J, Peng C, Meng J, Yang L, Liu J, Wen B, Zhang N, Huang Z, Zhu Q, Feng Y, Mount A, Hedgecock D, Xu Z, Liu Y, Domazet-Lošo T, Du Y, Sun X, Zhang S, Liu B, Cheng P, Jiang X, Li J, Fan D, Wang W, Fu W, Wang T, Wang B, Zhang J, Peng Z, Li Y, Li N, Wang J, Chen M, He Y, Tan F, Song X, Zheng Q, Huang R, Yang H, Du X, Chen L, Yang M, Gaffney PM, Wang S, Luo L, She Z, Ming Y, Huang W, Zhang S, Huang B, Zhang Y, Qu T, Ni P, Miao G, Wang J, Wang Q, Steinberg CE, Wang H, Li N, Qian L, Zhang G, Li Y, Yang H, Liu X, Wang J, Yin Y, Wang J. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490:49–54.
41. Dowle AA, Wilson J, Thomas JR. Comparing the diagnostic classification accuracy of ITRAQ, peak-area, spectral-counting, and emPAI methods for relative quantification in expression proteomics. J Proteome Res. 2016;10:3550–62.
42. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteom. 2005;4:1265–72.
43. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302:205–17.
44. Clamp M, Cuff J, Searle SM, Barton GJ. The jalview java alignment editor. Bioinformatics. 2004;20:426–7.
45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
46. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–9.
47. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61:1061–7.
48. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976;72:248–54.
49. Marriott PE, Sibout R, Lapierre C, Fangel JU, Willats WG, Hofte H, Gómez LD, McQueen-Mason SJ. Range of cell-wall alterations enhance saccharification in *Brachypodium distachyon* mutants. Proc Natl Acad Sci USA. 2014;111:14601–6.
50. Reynolds ES. The use of lead citrate at high pH as an electron opaque stain in electron microscopy. J Cell Biol. 1963;17:208–12.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central