Design, characterisation and *in planta* expression of a chimeric xylanase

by

Jonathan Botha

Submitted in partial fulfilment of the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences
Department of Biochemistry, Genetics and Microbiology
University of Pretoria
Pretoria
October 2018

Under the supervision of Prof. Donald A. Cowan and
Co-supervision of Prof. Alexander A. Myburg and
Dr Eshchar Mizrachi

DECLARATION

I, the undersigned, hereby declare that the thesis submitted herewith for the degree PhD to the University of Pretoria, contains my own independent work and has not been submitted for a degree at any other university.

____

Jonathan Botha

TABLE OF CONTENTS

CHAPTER 1

PREFACE

Second generation feedstocks, such as lignocellulosic biomass, are promising candidates for the sustainable and economically viable synthesis of bioproducts. However, lignocellulosic biomass is highly recalcitrant to enzymic digestion, making it difficult to extract and process the energy-rich biopolymers contained within the biomass. Normally, industrial pre-treatments (milling, grinding, steam explosion, ammonia fibre explosion etc.) are used to reduce recalcitrance of the material, but these are harsh, expensive to perform and produce degradation products that inhibit downstream processes. One strategy for overcoming recalcitrance and reducing the need for industrial pre-treatment is through the heterologous expression of Carbohydrate Active enZymes (CAZymes) directly in the biomass. CAZymes can target and degrade oligosaccharides, polysaccharides and glycoconjugates in lignocellulosic biomass, reducing the need for industrial pre-treatments and external enzyme loading. However, most CAZymes do not operate optimally at the extreme temperatures and pHs required for most industrial processes. A solution to this problem is found in extremely thermophilic organisms, which grow optimally at temperatures exceeding 70°C. These organisms provide a pool of thermostable CAZymes which can be used in industrial processes. Full thermostable CAZymes have been examined in detail, but comparatively little research has been performed on thermostable CAZyme domains. Thermostable CAZyme domains could be used for the design and synthesis of chimeric enzymes for lignocellulose degradation. By combining thermostable catalytic domains (such as a glycoside hydrolases, or GHs) with plant-derived protein domains (such as carbohydrate-binding modules, or CBMs) it may be possible to design a chimeric enzyme that is targeted to a specific location on a biopolymer in the plant secondary cell wall (SCW), while remaining inactive at mesophilic temperatures. This allows for accumulation in the biomass without negatively affecting growth and development of the plant. The biomass can then be harvested and heated, activating the catalytic domain and hydrolysing the biomass. This would be an important step for synthetic biology, and would allow for the production of synthetic enzymes, tailored to the specific needs of a given industrial process.

The aim of the thesis is to design, synthesise and characterise a chimeric enzyme consisting of one or more *E. grandis* CBMs and an extremely thermophilic catalytic domain that will degrade xylan in the SCW when exposed to high temperature in pre-treatment of woody biomass, and determine the effect that heterologous expression will have on the growth and development of the plant.

Chapter 1 is a review of recent literature surrounding the use of extremely thermophilic CAZyme domains in protein engineering, which serves as an introduction to the thesis. In this chapter, the discovery and characterisation of extremely thermophilic CAZyme domains is discussed. The engineering of known extremely thermophilic domains is addressed, as well as how they may be used as modules to construct synthetic thermostable enzymes. Finally, a list of predicted CAZyme domains from the proteomes of extremely thermophilic organisms is provided, and the capacity for degradation of lignocellulose is highlighted in the dataset.

Chapter 2 is an in depth analysis of the dataset obtained from Chapter 1. In this chapter, the composition and abundance of CAZyme domains between Archaea and Bacteria is compared and contrasted. The capacity for lignocellulose degradation is also closely interrogated, with a focus on domains that could target and degrade cellulose and xylan. Lastly, an estimation of CAZyme representation in the dataset is performed, and the likelihood of discovery of additional thermostable CAZyme domains with the sequencing of more extremely thermophilic organism genomes is assessed.

In **Chapter 3*,*** a chimeric enzyme is designed, synthesised and characterised. The enzyme consists of a thermostable GH11 domain, obtained from a compost-soil metagenomic library, and xylan-targeting CBM22 domains from *E.* grandis. The enzyme is heterologously expressed in *A. thaliana* and the effect on growth and development of the plant is assessed. Additionally, the accumulation of the enzyme in the plant biomass is investigated, as well as its localisation to the SCW. Finally, the effect of heterologous expression of the enzyme in the plant biomass on recalcitrance to enzymic digestion is studied.

Concluding remarks are included at the end of the thesis in **Chapter 4**. In this chapter, the results of the thesis are put into context of current studies and literature and their value to both the academe and industry are discussed. Finally, shortcomings of the study and possible improvements are addressed, and avenues for future research are highlighted.

List of publications from the thesis work

Refereed Publications

Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

Posters and conference presentations

Botha J, Pinard DS, Myburg AA, Mizrachi E, Cowan DA (2014) Towards creating a catalogue of hyperthermophile carbohydrate active enzyme (CAZyme) domains. The joint South African Society for Bioinformatics and South African Genetics Society (SASBi-SAGS) conference in Pretoria, South Africa (Poster presentation)

Botha J, Myburg AA, Mizrachi E, Cowan DA (2015). Hyperthermophiles: A source of CAZymes for industrial lignocellulosic degradation. International Union of Forest Research Organizations (IUFRO) Tree Biotechnology conference in Florence, Italy (Poster presentation)

Botha J, Myburg AA, Mizrachi E, Cowan DA (2015). Chimeric thermostable CAZymes for in planta degradation of lignocellulose. The Genomics Research Institute (GRI) annual symposium (Oral presentation)

ACKNOWLEDGEMENTS

I would like to acknowledge the following people, organisations and institutes for the assistance rendered to me in the completion of this study:

- The National Research Foundation (NRF) for funding supplied by the Innovation scholarship.
- The Human Resources and Technology for Industry Programme (THRIP), for funding the project.
- The Department of Biochemistry, Genetics and Microbiology (BGM), the Centre for Microbial Ecology and Genomics (CMEG), and the Forestry and Molecular Genetics (FMG) group at the University of Pretoria for providing the top notch support, facilities and infrastructure required to successfully complete a thesis.
- Prof. Don Cowan for all the work needed to get a PhD student to produce a half-decent thesis. Your support, guidance and mentorship will always be appreciated.
- Prof. Zander Myburg and Dr Eshchar Mizrachi for providing insight and guidance, especially with respect to the plant-based work. I’m sure that FMG will continue to flourish under your leadership.
- The staff of both CMEG and FMG. You guys keep everything running. I’m not sure most people realise how much of our work depends on you and the great job you all do.
- Dr Bilal Mir, for providing plant material and teaching the value of perseverance through adversity.
- (Soon to be Dr) Desre Pinard for all the help with the bioinformatic work and proofreading, but more importantly, cheeky beers.

- (Soon to be Dr) Marja O'neill, for the help with proofreading, but more importantly, (also) cheeky beers.
- (Soon to be Dr) Caryn Oates for helping me build a microscopy protocol from scratch, but more importantly, coffee and procrastination.
- Dr Steven Hussey for all the help with protein expression. You were so busy, yet still found time to help me out when I asked for it.
- Dr Mark Maistry and Mr Donovin Coles for keeping an eye on my plants when I couldn't.
- Dr Louise Shuey for all the enthusiasm and emotional support.
- My friends and labmates, but specifically Marja, Des, Caryn, Martin, Donno, Vicki and Lazarus. It was great to work with you all. I'm sure I'll be hearing good things about you all in the future.
- My good friends Andre, Terence, Stew, Marcio and Tristan. Good friendships can make even the hardest times bearable. I am eternally grateful.
- My family. My mother for all the support she provided, the least of which was financial. I'm not sure I could have done this without having someone so understanding and enthusiastic propping me up. Also, my brother and his wonderful family. You guys are great together, and were always a breath of fresh air when the thesis was dragging me down.
- And finally, Nora. Sometimes, language fails. Sometimes words cannot adequately capture the extent of love and support that someone may provide, or just how much that person means to another. Words can fall utterly short, and so to you I simply say thank you, for everything.

THESIS SUMMARY

Design, characterisation and *in planta* expression of a chimeric xylanase

Jonathan Botha

Supervised by: ***Prof. Donald A. Cowan***

Co-supervised by: ***Prof. Alexander A. Myburg*** *and* ***Dr Eshchar Mizrachi***

Submitted in partial fulfilment of the requirements for the degree ***Philosiphiae Doctor***

Department of Biochemistry, Genetics and Microbiology

University of Pretoria

The recalcitrance of lignocellulosic biomass to enzymic digestion remains a significant obstacle to the adoption of an environmentally and economically sustainable strategy for the synthesis of biomaterials. Traditional industrial pre-treatments are harsh, require significant investments of energy and money, and tend to produce degradation products which inhibit downstream processes. Carbohydrate Active enZymes (CAZymes) may reduce recalcitrance, through heterologous expression directly in the lignocellulosic biomass. CAZymes from extremely thermophilic organisms are not normally active at the mesophilic temperatures, allowing for accumulation in the biomass without negatively affecting the growth and development of the plant. Harvested biomass could then be heat-treated, activating the CAZymes and inducing hydrolysis of the biomass. Additionally, chimeric thermostable enzymes could be constructed from extremely thermophilic CAZyme domains, tailored to target specific biopolymers and perform directed modifications. However, while full-length CAZymes have been investigated, the extent of lignocellulose degrading capacity of extremely thermophilic CAZyme domains has not been assessed and the ability to produce and express chimeric CAZymes *in planta* has not been determined.

In this thesis, the CAZyme domain content of extremely thermophilic organisms was surveyed and capacity for degradation of lignocellulose was assessed. A list of CAZyme domains from extremely thermophilic organisms was produced via HMMER analysis. There were differences in CAZyme composition between extremely thermophilic archaea and bacteria, which could be mainly attributed to differences in nutritional strategy as well as synthesis, composition and structure of the cell walls in the organisms. Many putative lignocellulose degrading and targeting domains were present in the dataset, identified mostly in bacteria, though some were found only in archaea. It was also seen that more CAZyme domain variants and CAZyme domain classes are likely to be identified as more genomes of extremely thermophilic organisms are sequenced.

Additionally, a chimeric CAZyme consisting of a thermostable GH11 domain and plant-derived CBM22 domains designated Xyl22L was designed, synthesised and heterologously expressed in *Arabidopsis thaliana*. The effect on growth and development of the plant as well as recalcitrance to enzymic digestion of the biomass was determined. Xyl22L did not retain catalytic xylanase activity but was able to accumulate in transgenic plant biomass, and expression of *Xyl22L* was strongly correlated with an increase in transgenic plant biomass. Fluorescent confocal microscopy showed that Xyl22L was associated with the secondary cell wall (SCW) in transgenic plants, indicating that the CBM22 domains retained function. Finally, transgenic plant lines showed increased recalcitrance to enzymic digestion, possibly through Xyl22L adhering to the SCW and preventing access of hydrolytic enzymes.

This work provides a list of extremely thermophilic CAZyme domains, providing insight into the survival and evolution of extremely thermophilic organisms as well as a toolbox of thermostable domains for the synthesis of custom chimeric enzymes. Additionally, this work provides an example of such an enzyme, and provides proof of concept that plant-based CBMs may be used to target enzymes to specific biopolymers or locations in plant biomass. Together, these findings could be applied to white biotechnological processes, allowing for cheaper and more energy efficient bioproduct synthesis, enabling a transition away from a petrochemical-based products.

LIST OF TABLES

CHAPTER 1

CHAPTER 2

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

SUPPLEMENTARY MATERIALS

Electronic copies of all supplementary materials are located in Appendix A, which can be found on the disc secured to the back cover of this thesis.

SUPPLEMENTARY TABLES AND FIGURES

CHAPTER 3

SUPPLEMENTARY TABLES

SUPPLEMENTARY FIGURES

APPENDIX A: SUPPLEMENTARY MATERIAL (see attached disk)

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 1:
LITERATURE REVIEW

Carbohydrate active enzyme domains from extreme thermophiles – components of a modular toolbox for lignocellulose degradation

Jonathan Botha[1,2,4], Eshchar Mizrachi[2,3], Alexander A. Myburg[2,3], Don A. Cowan[1,2,4]*

*Corresponding author: Don.Cowan@up.ac.za

[1]Centre for Microbial Ecology and Genomics, Department Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

[2]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

[3]Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

[4]Genomics Research Institute (GRI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

This chapter was submitted to the scientific journal Extremophiles as a review paper. I conceived of and drafted the manuscript, as well as prepared all figures. Prof. D.A. Cowan, Prof. A.A. Myburg and Dr E. Mizrachi provided guidance and oversight, as well as critical review of the work. This chapter was published under the following title: Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

1.1 Abstract

Lignocellulosic biomass is a promising feedstock for the manufacture of biodegradable and renewable bioproducts. However, the complex lignocellulosic polymeric structure of woody tissue is difficult to access without extensive industrial pre-treatment. Enzyme processing of partly depolymerised biomass is an established technology, and there is evidence that high temperature (extremely thermophilic) lignocellulose degrading enzymes (CAZymes) may enhance processing efficiency. However, wild-type thermophilic CAZymes will not necessarily be functionally optimal under industrial pre-treatment conditions. With recent advances in synthetic biology, it is now potentially possible to build CAZyme constructs from individual protein domains, tailored to the conditions of specific industrial processes. In this review, we identify a 'toolbox' of thermostable CAZyme domains from extremely thermophilic organisms and highlight recent advances in CAZyme engineering which will allow for the rational design of CAZymes tailored to specific aspects of lignocellulose digestion.

1.2 Keywords

Lignocellulose, CAZyme, Extreme thermophiles, Synthetic biology, Protein domains

1.3 Introduction

Biomaterials (materials derived from renewable and sustainable biological substrates) could provide an economically viable strategy to mitigate or ameliorate environmental challenges and potentially replace products derived from petrochemical feedstocks (Naik et al. 2010). However, to maximise efficiency of bioproduct synthesis, large sources of simple polysaccharides are required. Second generation feedstocks are a potential source of such material, as they are typically not food crops (such as many first generation bioproduct feedstocks) and, therefore, do not impact on human food security. Some plant species produce large amounts of dense lignocellulosic biomass and are able to grow in a wide range of conditions and environments (Hendriks and Zeeman 2009; Himmel et al. 2007). However, lignocellulosic biomass is highly recalcitrant (Himmel et al. 2007), requiring large investments of energy (with associated financial and waste costs) to access (Alvira et al. 2010).

The integration of Carbohydrate Active Enzymes (CAZymes) in the processing of lignocellulosic biomass is a promising strategy for reducing the difficulty and cost of biopolymer extraction (Mir et al. 2014; Turumtay 2015). CAZymes are active on oligosaccharides, polysaccharides and glycoconjugates. They consist of protein domains that are classified into a hierarchy of families based on their structure and function, including Glycoside Hydrolases (GHs), GlycosylTransferases (GTs), Carbohydrate-Binding Module (CBMs), Carbohydrate Esterases (CEs), Polysaccharide Lyases (PLs) and Auxiliary Activity families (AAs) (Cantarel et al. 2009; Lombard et al. 2014).

However, while CAZymes represent a useful tool for breaking down lignocellulosic biomass, they are generally not suited to the harsh conditions (especially extreme temperature) that form the basis of many industrial pre-treatments of lignocellulosic biomass (Blumer-Schuette et al. 2014). As one approach to overcoming the problem of enzyme functional stability, thermostable CAZymes have been identified in and isolated from hyperthermophilic and extremely thermophilic organisms [i.e., organisms that grow optimally at temperatures exceeding 70^{o}C (Leuschner and Antranikian 1995; Gerday and Glansdorff 2007)]. By mining genomes of extremely thermophilic organisms, it is possible to identify CAZymes that may operate effectively under the extreme conditions characteristic of industrial pre-treatments and other white biotechnological (i.e. industrial) applications. An added benefit of using extremely thermophilic CAZymes is that they may be expressed *in planta* with little fear of cytotoxicity, due to their very low activity at mesophilic temperatures (Mir et al. 2014). This approach can potentially reduce the need for exogenous enzyme loading, where accumulation of the expressed enzyme in the plant tissues produces 'self-processing' plants, potentially increasing the efficiency of lignocellulose degradation at high temperatures [without impacting the normal growth and development of the plant at mesophilic temperatures (Mir et al. 2014)].

The field of synthetic biology has a long history, and through innovations such as the iGEM competition and Biobricks foundation (Smolke 2009; Vilanova and Porcar 2014) has established a registry of 'components' (including promoters, ribosome binding sites, protein coding sequences and terminators, among others; http://parts.igem.org) for the rational design and programming of systems in living cells (Endy 2005; Hartwell et al. 1999; Purnick and Weiss 2009). However, characterisation and incorporation of new individual components into the system remains an issue, and is a bottleneck to progress (Cameron et al. 2014). Most components are derived from wild-type systems and are limited in function. For example, a xylanase which binds to cellulose may be desirable, in order to digest xylan closely associated with cellulose to make it more accessible to enzymic degradation, but xylanase CBMs do not typically bind well to cellulose (McCartney et al. 2006). Nevertheless, the concept of engineering rationally designed multi-component CAZymes, potentially capable of targeting lignocellulose biopolymer backbones and accessory structures would appear to offer considerable promise for lignocellulosic biomass processing.

There have been a number of recent reviews summarising the industrial applications of extremely thermophilic CAZymes, either of specific enzyme classes (Atomi et al. 2011; Elleuche et al. 2015; Fatima and Hussain ; Nisha and Satyanarayana 2016) or with a focus on lignocellulose processing (Blumer-Schuette et al. 2014; Elleuche et al. 2014; Guerriero et al. 2015; Mir et al. 2014; Turumtay 2015; Urbieta et al. 2015). However, these reviews do not highlight the modular nature of hyperthermophilic CAZymes or how they might be used to tailor enzymes for lignocellulose deconstruction. In this review, we provide a survey of CAZyme modules with known lignocellulose degrading abilities derived from extremely thermophilic organisms, and explore how they might be applied in white biotechnology and enzyme engineering. We identify a list of predicted CAZyme domains (using publicly available proteomes from Ensemble Bacteria and HMMER protein domain prediction; Online Resource 1) from the proteomes of a selection of extremely thermophilic organisms (Figure 1.1, Online Resource 2). These domains cover a significant proportion of existing CAZyme families (Figure 1.2) and comprise a wide range of activities and substrate specificities, as summarised in the CAZy database [www.cazy.org (Lombard et al. 2014)].

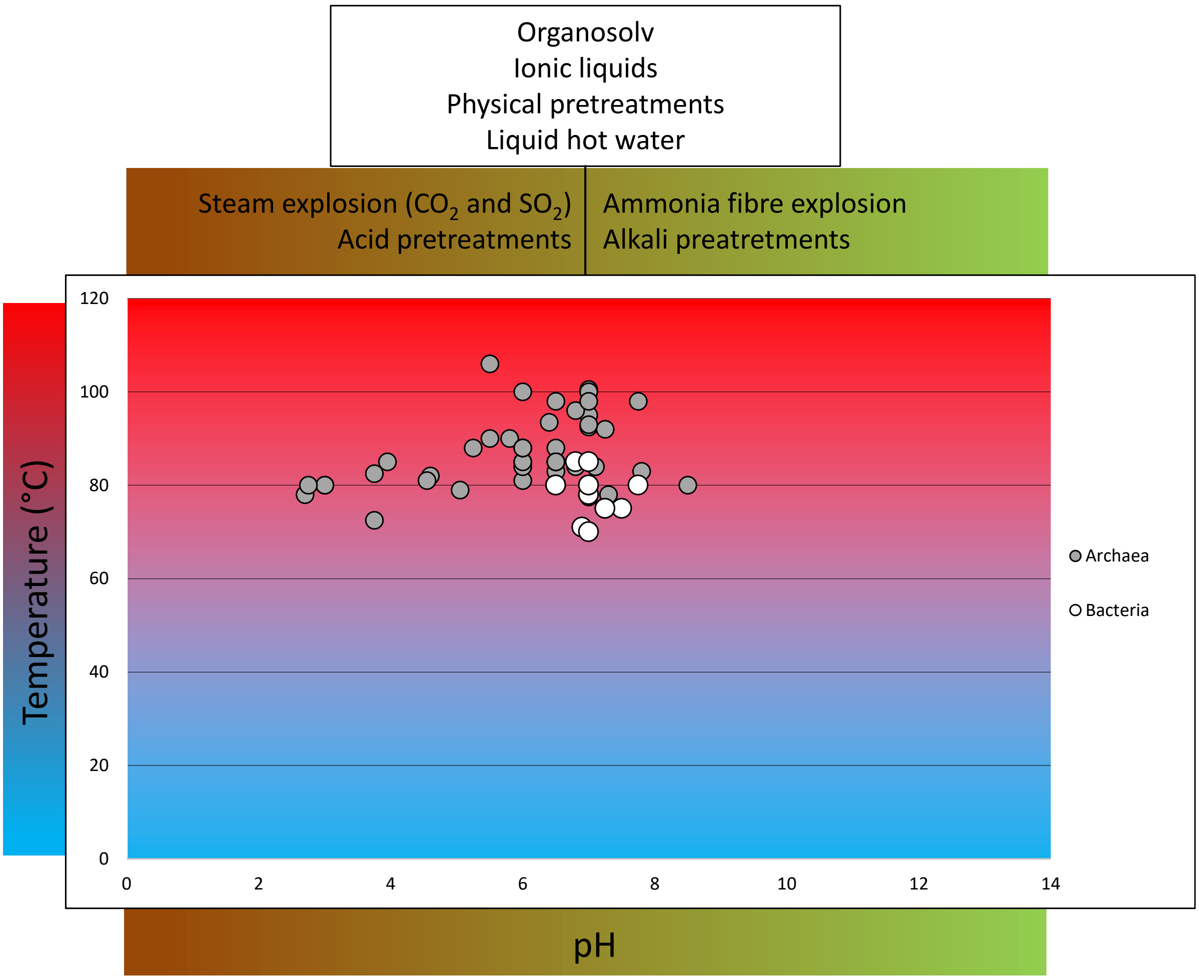


Figure 1.1 The spread of optimum growth temperature and pH of extremely thermophilic organisms covered in this review, as well as some of the main pre-treatments (Alvira et al. 2010) associated with lignocellulose deconstruction.

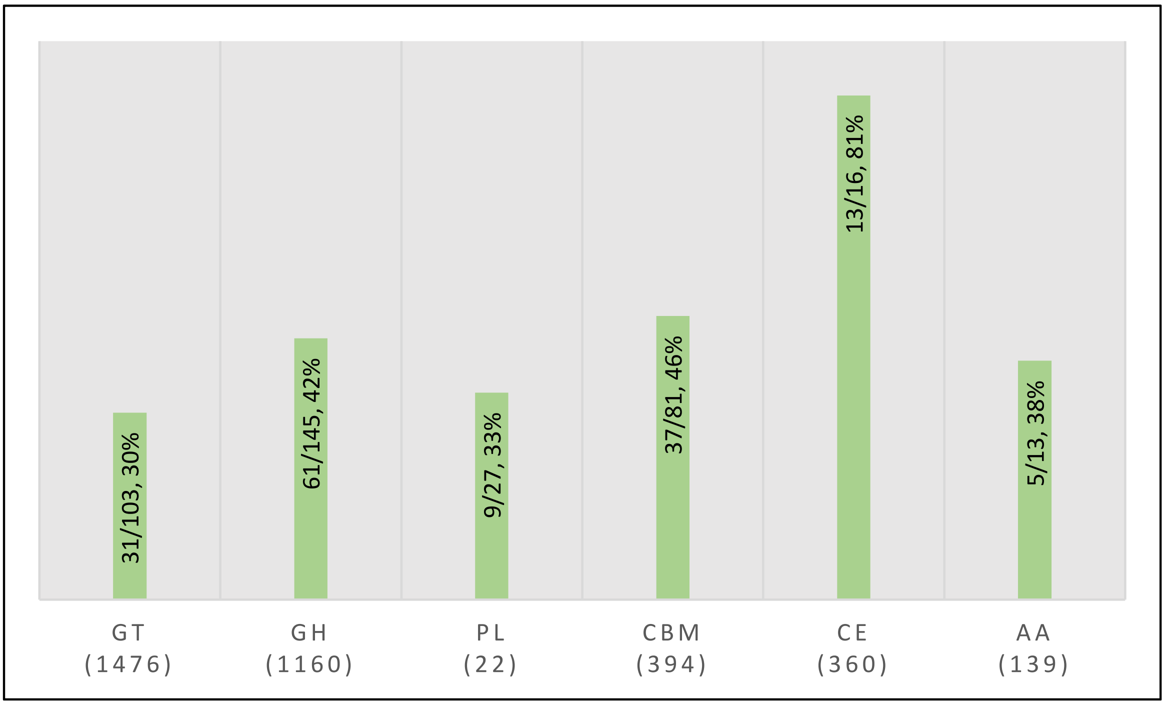


Figure 1.2 The overall coverage of CAZyme families identified from extremely thermophilic organisms, expressed as a percentage. The CAZyme class is listed on the X-axis. GT: Glycosyl Transferase, GH: Glycoside hydrolase, PL: Polysaccharide Lyase, CBM: Carbohydrate-Binding Module, CE: Carbohydrate Esterase and AA: Auxilliary Activity. The number listed below each class is the absolute count of domains identified in each class for extremely thermophilic organisms. The first number in the bars indicates the absolute proportion of all known CAZyme domain families present in the extremely thermophilic organisms and the second number is this proportion expressed as a percentage.

1.4 CAZyme categories

CAZyme domains are divided into families based on their structures (Cantarel et al. 2009; Lombard et al. 2014). This classification implies that two CAZyme domains that are members of the same family are likely to have similar activities and substrate specificities, but in reality this is not always the case. While broad trends are often seen within a CAZyme family, the diversity of activities, specificities and thermostabilities is high. This is evident in the GH3 domain family, where four GH3 enzymes from *Cellulomonas fimi* showed a range of activities on many xylo- and oligosaccharides (Gao and Wakarchuk 2014) even though no other catalytic domains were apparent. The GH5 domain family also displays considerable functional diversity, with recent studies highlighting specificities for substrates such as cellulose (Huy et al. 2016; Valadares et al. 2016; Wang et al. 2016b), xylans and xyloglucans (dos Santos et al. 2015; Ghatge et al. 2014), mannans (Tóth et al. 2016; Zang et al. 2015) and glycoceramides (Han et al. 2017b). Occasionally, new CAZyme domain families or rare variants of existing families are also discovered, such as a xylan degrading (Corrêa et al. 2012) or multifunctional (Morrison et al. 2016) GH39 enzyme, or the recently defined GH116 family, which was shown to act on glucosylceramide, N-acetylglucosaminides and xylosides (Cobucci-Ponzano et al. 2010; Ferrara et al. 2014). If each CAZyme domain is viewed as a potential building-block for the rational design and synthesis of enzymes for a range of applications, then identification and characterisation of new CAZyme domains from both known and novel families will further supplement our toolset for enzyme design.

'Omics' technologies are excellent methods for expanding the CAZyme repertoire. By investigating the (meta)genomes, (meta)transcriptomes, (meta)proteomes and (meta)secretomes of organisms and communities which process lignocellulose (Kuuskeri et al. 2016; López-Mondéjar et al. 2016; Montella et al. 2017; Schneider et al. 2016; Solomon et al. 2016; Wang et al. 2016a), a full complement of lignocellulose degrading CAZymes may be identified. Additionally, CAZyme genes from organisms which synthesise lignocellulose [such as plants (Geisler-Lee et al. 2006; Pinard et al. 2015) and some bacteria (Zhang et al. 2017)] could allow for diversified modification of biopolymers [such as adding side chains or chemical groups, and altering the structure of xylan (Abramson et al. 2010)].

Cataloguing each domain individually could provide a comprehensive toolbox for enzyme design, but the capability of such a toolbox could be drastically increased by determining the underlying mechanisms of domain variety and implementing them in protein design. This would allow for fine-tuning of synthetic enzymes to specific processes. A key technique for investigating differences in mechanisms of action is protein crystallography, which uses high quality structural data to identify mechanisms of binding or stability between and within CAZyme domains (Czjzek and Ficko-Blean 2017). Relatively few CAZyme domains have been structurally resolved at high resolution (www.cazy.org), but every CAZyme class is based on at least one resolved structure.

Recent examples of mechanistic insights from structural analyses include new variants of rare domain structures (Godoy et al. 2016), general mechanisms of substrate binding [as seen in studies on CBM35 xylanases (Sainz-Polo et al. 2014a; Sainz-Polo et al. 2014b; Valenzuela et al. 2012)], GH30 xylanases (Sainz-Polo et al. 2014a; Sainz-Polo et al. 2014b; Verma and Goyal 2014; Verma et al. 2013) and a GH52 β-xylosidase (Espina et al. 2014) as well as interactions with specific substrates such as xyloglucan (Attia et al. 2016; dos Santos et al. 2015) and cellulose (Pires et al. 2017).

The value of structural data is not limited to understanding interactions of domains with substrates. In some cases it can help to elucidate how enzymes behave under certain conditions [such as the mechanistic basis for glucose tolerant and intolerant GH1 domains (Yang et al. 2015)]. Structural data can reveal interactions between domains in a single enzyme. The resolved structure of Xyn10C from *Paenibacillus barcinonensis* (Sainz-Polo et al. 2015) is the first structure of an enzyme with two tandem CBMs, showing the architecture and interaction of multiple domains in an enzyme.

1.5 Engineering CAZyme domains

Studying CAZyme domains and the mechanisms by which they perform their functions is the first step towards informed rational CAZyme design. While it is possible for a protein to have multiple domains with different substrate targets (Sainz-Polo et al. 2015), enzyme design constitutes more than simply choosing domains with desirable functions and combining them in an arbitrary manner (André et al. 2014; Elleuche 2015). For example, the addition of CBM3, CBM4 or CBM22 CAZyme domains to a GH7 cellobiohydrolase (Voutilainen et al. 2014), GH9 cellulase (Duan et al. 2017), gluco-oligosaccharide oxidase (Foumani et al. 2015) and CE1 acetyl xylan esterase (Liu and Ding 2016) increased binding affinity, thermostability and enzyme activity. The removal of CBM domains also had a deleterious effect on enzyme performance, with reduced activity and thermostability observed in a GH5 endoglucanase (Ghatge et al. 2014), PL7 algenate lyase (Li et al. 2015) and GH9/GH48 cellulase (Yi et al. 2013). The relationship between domain functions is not necessarily additive. Whole CAZymes can behave in a synergistic manner (Chung et al. 2015; Liu and Ding 2016) and CAZyme domains have also been shown to functionally synergise (Diogo et al. 2015; Liu and Ding 2016). The effects of domain addition or deletion are not necessarily predictable. For example, separating two domains that co-occur in a wild-type enzyme (such as the CBM46 and GH5 domains from *Bacillus halodurans Bh*Cel5b) can reduce or abolish the function of both (Venditto et al. 2015). Conversely, some domains operate more efficiently when separated from each other, such as the GH10 domain from *Clostridium thermocellum* XynZ, which displayed higher thermostability and catalytic activity in truncated variants lacking the native CBM6 domain (Sajjad et al. 2010). The addition of a known thermostabilizing domain such as CBM22 (Khan et al. 2013; Lee et al. 1993) can reduce the thermostability of the protein construct while increasing catalytic efficiency (Araki et al. 2006). A summary of these interactions, as well as some additional examples can be found in Table 1. These examples emphasise the complex interactions that occur between CAZyme domains within an enzyme.

The most obvious and immediate use for a toolbox of CAZyme domains would be to provide a catalogue of parts from which custom enzymes may be assembled. However, as the toolbox expands, it may begin to fill a substantially more significant function—providing a set of protein scaffolds for rational engineering of new functionalities. To date, reports of rationally engineered CAZymes are limited, but two main strategies are being employed: directed evolution (DE) and rational design (Davids et al. 2013).

Using DE, thermostability of GH10 Xyn III from *Trichoderma reesei* was enhanced (Matsuzawa et al. 2016). Similarly, GH51 α-L-arabinofuranosidases have been engineered for higher transglycosylating activity (Arab-Jaziri et al. 2013; Arab-Jaziri et al. 2015) and reduced secondary hydrolysis of transglycosylation products (Bissaro et al. 2014). While DE is an attractive option for protein engineering, widespread use of the technique is impeded by difficulties with screening and selection of variants with desirable traits (Turner 2009).

Table 1.1 Summary of the effects of addition or removal of CAZyme domains to enzymes

Base domain(s)[a]	WT protein[b]	Modification type[c]	Added/removed domain[d]	Added/removed domain origin[e]	Effect[f]	Reference
CBM22, CBM22, GH10, CBM9	*Cs*Xyl10B	-N	CBM22, CBM22	*Cs*Xyl10B	↓ Ts	Araki et al. 2006
CBM22, GH10	*Ct*XynC	-N	CBM22	*Ct*XynC	↓ Ts	Sajjad et al. 2010
CBM6, GH10	*Ct*XynZ	-N	esterase, dockerin, CBM6	*Ct*XynZ	↑ As, ↑ Ts	
CBM6, GH10	*Ct*XynZ	-N; +C	CBM6; CBM22	*Ct*XynZ	↑ As	Khan et al. 2013
CE1, CBM1	Acetyl Xylan Esterase, *V. volvacea*	=CBM1	CBM4-2	NA	↓ Ba, ↑ Ts, ↑ As	Liu and Ding 2016
		=CBM1	CBM6	NA	↓ Ba, ↑ As	
		=CBM1	CBM22-2	NA	↓ Ba, ↑ As	
		-C	CBM1	Acetyl Xylan Esterase, *V. volvacea*	↓ Ba, ↓ As	
GH10	*Tr*XynIII	+C	Xylan Binding Domain	XBD, *S. olivaceoviridis*	↑ Ba, ↑ As	Matsuzawa et al. 2016
GH10, CBM3b, CBM3b, GH48	*Cb*Xyn10C/Cel48B	-C	CBM3b, CBM3b, GH48	*Cb*Xyn10C/Cel48B	↓ As, ↑ Sr	Xue et al. 2015
GH11	*Bs*XynA	+C	GH43	*Bs*XynB	↑ Ts, ↑ As	Diogo et al. 2015
GH11	*Bs*Xyl11	+C	CBM6	*C*the_1963	↑ As	Hoffmam et al. 2016
GH5	*Fm*EG	+N	CBM1	EG1, *V. vovacea*	↑ Ts, ↑ As	Pan et al. 2016
GH5, CBM6-2, CBM6-2	*Hc*Cel5	-C	CBM6-2	*Hc*Cel5	↓ Ba, ↓ As	Ghatge et al. 2014
GH5_4, CBM46	*Bh*CBM46	-C	CBM46	*Bh*CBM46	↓ Ba, ↓ As	Venditto et al. 2015
		-N	GH5_4		↓ Ba, ↓ As, ↓ Ts	
GH7	*Te*Cel7A	+C	CBM1	*Tr*Cel7A	↑ Ts, ↑ As	Voutilainen et al. 2014
		+C	CBM2	*Cf*Xyn10A	↑ Ts, ↑ As	
		+C	CBM3	*Ct*CipA	↑ Ts, ↑ As	
GH9	*Um*Cel9A	+C	CBM1	*Tr*Cel7A	↑ Ba, ↑ As	Duan et al. 2017
		+C	CBM2	GH9 endoglucanase, *C. flavigena*	↑ Ba, ↑ As	
		+C	CBM3	GH9 endoglucanase, *R. thermocellum*	↑ Ba, ↑ As	
		+C	CBM4	GH9 endoglucanase, *C. cellulolyticum*	↑ Ba, ↑ As	
		+C	CBM10	GH9 endoglucanase, *C. japonicus*	↑ Ba, ↑ As	
		+C	CBM72	GH5 endoglucanase, Uncultured organism	↑ Ba, ↑ As	

Table 1.1 (continued) Summary of the effects of addition or removal of CAZyme domains to enzymes

Base domain(s)[a]	WT protein[b]	Modification type[c]	Added/removed domain[d]	Added/removed domain origin[e]	Effect[f]	Reference
GH9, CBM3c, CBM3b, CBM3b, GH48	*Cb*Cel9A/Cel48A	-C	GH48	*Cb*Cel9A/Cel48A	↓ As	Yi et al. 2013
		-N	GH9		↓ As, ↑ Ts	
		-C	CBM3b, CBM3b, GH48		↓ As	
		-C	CBM3b, GH48		↓ As	
GOOX	Gluco-oligosaccharide oxidase, *S. strictum*	+C	CBM3	*Ct*CipA	↑ Ba, ↓ As	Foumani et al. 2015
		+C	CBM11	*Ct*Cel5E	↑ Ba, ↑ As	
		+C	CBM44	*Ct*Cel44A	↑ Ba, ↑ As	
		+N	CBM3	*Ct*CipA	↑ Ba, ↑ As	
		+N	CBM11	*Ct*Cel5E	↑ Ba, ↑ As	
		+N	CBM44	*Ct*Cel44A	↑ Ba, ↑ As	
PL7, CBM13	Alginate lyase, *Agarivorans* sp. L11	-N	CBM13	Alginate lyase, *Agarivorans* sp. L11	↓ As, ↓ Ts	Li et al. 2015

[a] The domains present in the native protein before modification

[b] The protein that is subject to modification

[c] The type of modification performed on the protein. – and + indicate removal and addition of a domain, respectively. The terminal at which the modification is made is indicated by N and C for the N-terminal and C-terminal, respectively. = indicates a substitution, with the domain which is being substituted noted immediately after the symbol.

[d] The domain which is added, removed or substituted into the protein, as described in the previous column.

[e] The protein from which the domains which are added, removed or substituted originate.

[f] The effect of the described modification on the protein. The up and down arrows indicates increase and decrease, respectively. As: Activity on substrate, Ba: Binding affinity, Sr: Substrate range, Ts: Thermostability.

To overcome this problem, a number of strategies have been developed. By limiting analysis to mutations in specific residues (such as those in the active sites of proteins), it is possible to reduce the number of variants to be screened. CASTing, or Combinatorial Active Site Testing is one method by which this can be achieved (Reetz et al. 2006). This approach requires knowledge of the crystal structure of the protein to accurately identify the sites for modification. If the crystal structure is not available, high throughput screens are used. Techniques such as ribosome display (Gan and Jewett 2016), mRNA display (Horiya et al. 2017), yeast display (Traxlmayr and Shusta 2017) and phage-based techniques (Brödel et al. 2017) may be used in combination with *in vitro* compartmentalization-based, fluorescence-activated cell sorting (IVC-FACS) to identify and pool variants with desirable properties (Ma et al. 2016). Finally, colorimetric assays based on enzyme reaction mechanisms can be adapted to high throughput applications (Smart et al. 2017). Recently, a kit was developed that contains chromogenic substrates that can be used to test the activities of carbohydrate degrading enzymes that can be multiplexed in a 96-well format (Schückel et al. 2016). Although substantial progress has been made, selection and screening of DE clone libraries remains a challenge and continues to be an area of active research and innovation (Klenk et al. 2016; Lin et al. 2017; Ma et al. 2016; Reetz 2017).

The limitations of DE (Turner 2009) may be overcome through rational design (directly targeting and mutating specific residues of a protein to obtain a desired effect), although this approach requires in-depth knowledge of the structure and mechanisms of the protein. Using this method, various CAZyme properties have been altered; e.g., increasing both the thermal stability and optimal catalytic temperature of GH10 (de Souza et al. 2016) and GH11 (Han et al. 2017a) xylanases by the introduction of a mutations (by site-directed mutagenesis) to influence disulphide bond formation, salt bridges and the ratio of acidic to basic amino acids (de Souza et al. 2016; Han et al. 2017a).

Enzyme engineering of CAZymes in order to confer new functional characteristics has been applied with some success. Altering active site architecture through mutation (e.g. W22Y) changed the binding properties of *T. reesei* GH12 enzyme *Tr*Cel12A, resulting in expanded substrate specificity (Zhang et al. 2015). Similarly, a GH1 β-glucosidase from *Thermus thermophilus* was converted to a trans-β-acetylglucosaminidase by mutation of the N163 and E338 residues in the active site to remove steric conflicts with the N-acetyl-D-glucosamine substrate (André-Miral et al. 2015). This is an especially significant result, considering that no native GH1 family protein exhibits trans-β-acetylglucosaminidase activity. Additionally, a single predicted amino acid change in a *Geobacillus stearothermophilus* β-xylosidase (Y509E) was enough to confer new exo-xylanase functionality to the enzyme (Huang et al. 2014).

1.6 Applying the toolbox: Extremely thermophilic CAZymes for *in planta* lignocellulose degradation

The efficient breakdown of lignocellulose can be achieved using extremes of temperature, pressure and pH, which differ depending on which biopolymer is the target of extraction (Alvira et al. 2010). While supplementing physical and chemical pre-treatment processes with enzymes can potentially reduce the economic and energy investment required for biopolymer extraction (Blumer-Schuette et al. 2014), *in planta* expression of these enzymes (as opposed to external enzyme loading) may also be beneficial (Mir et al. 2014; Mir et al. 2017). Expressing thermostable enzymes directly in the plant tissue may not disrupt normal growth and development of the biomass, due to inactivity of the enzymes at lower temperature (Mir et al. 2014; Mir et al 2017; Montalvo-Rodriguez et al. 2000; Ziegler et al. 2000). On heating, the harvested biomass undergoes some autohydrolysis (Bhatia et al. 2017; Mir et al. 2014). This strategy has been shown to be effective in first generation feedstocks (Kim et al. 2016).

However, studies on the heterologous *in planta* expression of thermostable CAZymes have shown that while this strategy does improve digestibility, localisation of the product is important (Castiglia et al. 2016; Kim et al. 2016). Expressing a protein in the wrong cellular compartment can lead to deleterious effects. For example, plastid-targeted expression of a thermostable endoglucanase in tobacco resulted in binding of the recombinant gene product to thylakoid membranes and a negative impact on plastid development (Castiglia et al. 2016). Conversely, enzymes may perform better, or accumulate to a higher level if expressed in the correct locale [such as Xyl10b from *Thermotoga maritima MSB8,* which had higher yield and specific activity when targeted to the apoplast vs the chloroplast (Kim et al. 2016)]. Fusion to a CBM is one strategy through which translocation of a protein product to a desired location may be achieved (Oliveira et al. 2015).

1.7 Conclusions

Hyperthermophilic and extremely thermophilic organisms are a potentially valuable source of thermostable CAZyme domains for industrial use (Blumer-Schuette et al. 2014). The ability to degrade lignocellulosic substrates is surprisingly common in hyperthermophilic and extremely thermophilic organisms, despite the oligotrophic nature the ecological niches these organisms typically inhabit (but see Chaban et al. 2006; Rothschild and Mancinelli 2001). There are now numerous reports of highly stable CAZymes, across most of the CAZyme families and we have provided a list of putative thermostable CAZyme domains that can break down cellulose and xylan (Table 2, Figure 1.3), as well as other important lignocellulosic biopolymers (Online Resource 3).

However, many issues relating to the function and engineering of these enzymes remain unresolved. Modifications to CAZymes can have unpredictable consequences and the achievable limits for engineered catalytic ability are not known. Additionally, considering the functional diversity in some CAZyme families, it is unclear whether molecular modelling and docking studies on some enzymes in a family may be applied to others. Finally, while producing self-processing lignocellulosic biomass is an attractive prospect for industry, the impact of *in planta* production of enzymes on other performance factors (such as disease resistance) of plant growth need to be determined.

CAZyme domains from extremely thermophilic organisms provide a toolbox which can be used to design enzymes suited to a range of industrial processes. As our understanding of protein domains and enzyme engineering increases, so will the value of such a toolbox. These domains provide a way to modify existing enzymes and are also highly thermostable scaffolds for further modifications. Studying these domains and their interactions within proteins could eventually facilitate *de novo* design and synthesis of new highly thermostable and highly catalytic proteins.

Table 1.2 List of CAZyme families identified by this review and the main substrates they are known to target, as listed in the CAZy database (www.cazy.org).

CAZyme family[a]	Number[b]	Main substrates[c]
CBM2	3	Cellulose, chitin and xylan.
CBM3	18	Cellulose and chitin
CBM4	18	Xylan, β-1,3-glucan, β-1,3-1,4-glucan, β-1,6-glucan and amorphous cellulose
CBM6	1	Amorphous cellulose, xylan, β-1,3-glucan, β-1,3-1,4-glucan, and β-1,4-glucan
CBM9	19	Xylan and cellulose
CBM13	3	Mannose and xylan
CBM16	3	Cellulose and glucomannan
CBM22	37	Xylan and mixed β-1,3/β-1,4-glucans
CBM28	2	Non-crystalline cellulose, cellooligosaccharides, and β-(1,3)(1,4)-glucans
CBM35	15	Xylan, decorated soluble mannans, mannooligosaccharides and β-galactan.
CBM36	1	Xylans and xylooligosaccharides
CBM37	8	Xylan, chitin, microcrystalline and phosphoric-acid swollen cellulose, alfalfa cell walls, banana stem and wheat straw
CBM44	28	Cellulose and xyloglucan
CBM46	1	Cellulose
CBM54	9	Xylan, yeast cell wall glucan and chitin
CBM60	1	Xylan
CBM63	1	Cellulose
GH10	24	Cellulose and xylan
GH11	1	Xylan
GH1	54	Cellulose, xylan, mannan and xyloglucan
GH3	30	Cellulose and xylan
GH5	31	Cellulose, xylan, mannan, lichenin, chitosan, xyloglucan and arabinoxylan
GH8	1	Chitosan, cellulose, lichenin and xylan
GH9	3	Cellulose, lichenin, xyloglucan
GH12	28	Cellulose, xylan and xyloglucan
GH16	10	Cellulose, xylan, xyloglucan, lichenin and chitin
GH26	5	Mannan, xylan and cellulose

[a]The CAZyme family designation
[b]The number of domains identified from extremely thermophilic proteins using HMMER
[c]The main substrates on which these domains act

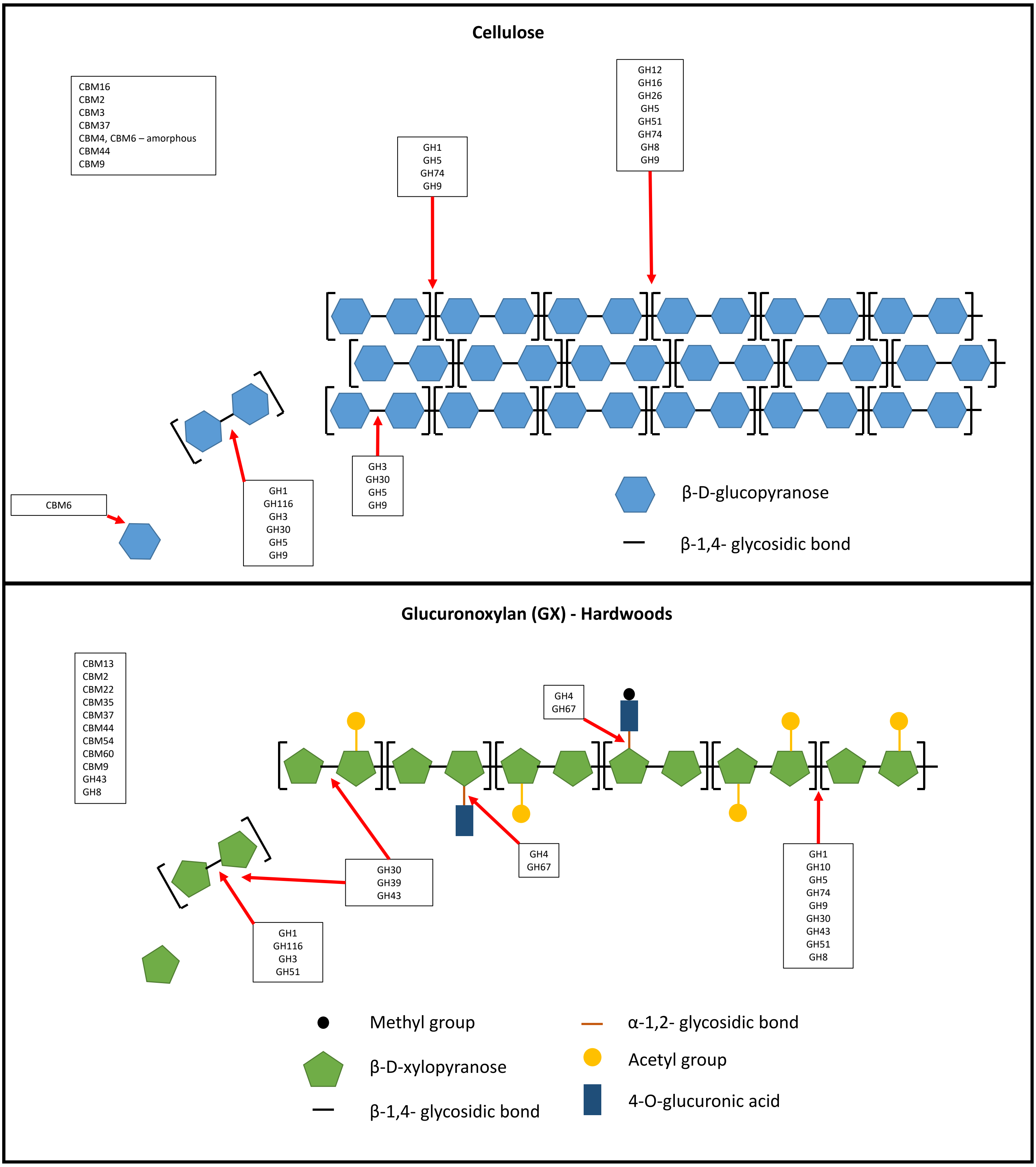


Figure 1.3 Schematic representation of important plant secondary cell wall biopolymers, as well as the CAZyme domains which can degrade them covered by this review. The name of the biopolymer is listed at the top of the figure. The box in the top left of the diagram indicates CAZyme domain families which have activities on the biopolymer, but have no record of specific interactions. The red arrows indicate specific areas of activity. The key is located at the bottom of each diagram.

1.8 Acknowledgements

The authors wish to thank the National Research Foundation (South Africa) for financial support.

1.9 Conflict of interest

The authors declare no conflict of interest.

1.10 Electronic Supplementary Material captions

Please note that all electronic supplementary material may be accessed via the article online:

Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

Online Resource 1 Extremely thermophilic organisms from which the domain list was constructed

Online Resource 2 CAZyme domains identified from extremely thermophilic organisms proteomes using HMMER analysis

Online Resource 3 Schematic representation of additional plant secondary cell wall biopolymers, and all putative CAZyme domains which can degrade them. The name of the biopolymer is listed at the top of the figure. The box in the top left of the diagram indicates CAZyme domain families which have activities on the biopolymer, but have no record of specific interactions. The red arrows indicate specific areas of activity. The key is located at the bottom of each diagram.

Electronic supplementary material is also located in Appendix A.

1.11 References

Abramson M, Shoseyov O, Shani Z (2010) Plant cell wall reconstruction toward improved lignocellulosic production and processability. Plant Science 178:61-72

Alvira P, Tomás-Pejó E, Ballesteros M, Negro M (2010) Pre-treatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: a review. Bioresource technology 101:4851-4861

André-Miral C, Koné FM, Solleux C, Grandjean C, Dion M, Tran V, Tellier C (2015) *De novo* design of a trans-β-N-acetylglucosaminidase activity from a GH1 β-glycosidase by mechanism engineering. Glycobiology 25:394-402

André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M (2014) CAZyme discovery and design for sweet dreams. Current Opinion in Chemical Biology 19:17-24

Arab-Jaziri F et al. (2013) Engineering transglycosidase activity into a GH51 α-l-arabinofuranosidase. New biotechnology 30:536-544

Arab-Jaziri F, Bissaro B, Tellier C, Dion M, Fauré R, O'Donohue MJ (2015) Enhancing the chemoenzymatic synthesis of arabinosylated xylo-oligosaccharides by GH51 α-l-arabinofuranosidase. Carbohydrate Research 401:64-72

Araki R, Karita S, Tanaka A, Kimura T, Sakka K (2006) Effect of family 22 carbohydrate-binding module on the thermostability of Xyn10B catalytic module from *Clostridium stercorarium.* Bioscience Biotechnology and Biochemistry 70:3039

Atomi H, Sato T, Kanai T (2011) Application of hyperthermophiles and their enzymes. Current Opinion in Biotechnology 22:618-626

Attia M, Stepper J, Davies GJ, Brumer H (2016) Functional and structural characterization of a potent GH74 endo-xyloglucanase from the soil saprophyte *Cellvibrio japonicus* unravels the first step of xyloglucan degradation. FEBS Journal 283:1701-1719

Bhatia R, Gallagher JA, Gomez LD, Bosch M (2017) Genetic engineering of grass cell wall polysaccharides for biorefining. Plant Biotechnology Journal 15:1071–1092

Bissaro B et al. (2014) Mutation of a pH-modulating residue in a GH51 α-l-arabinofuranosidase leads to a severe reduction of the secondary hydrolysis of transfuranosylation products. Biochimica et Biophysica Acta (BBA)-General Subjects 1840:626-636

Blumer-Schuette SE et al. (2014) Thermophilic lignocellulose deconstruction. FEMS microbiology reviews 38:393-448

Brödel AK, Jaramillo A, Isalan M (2017) Intracellular directed evolution of proteins from combinatorial libraries based on conditional phage replication. Nature Protocols 12:1830-1843

Cameron DE, Bashor CJ, Collins JJ (2014) A brief history of synthetic biology. Nature Reviews Microbiology 12:381-390

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic acids research 37:D233-D238

Castiglia D et al. (2016) High-level expression of thermostable cellulolytic enzymes in tobacco transplastomic plants and their use in hydrolysis of an industrially pretreated *Arundo donax L.* biomass. Biotechnology for Biofuels 9:1

Chaban B, Ng SY, Jarrell KF (2006) Archaeal habitats-from the extreme to the ordinary. Canadian Journal of Microbiology 52:73-116

Chung D, Young J, Cha M, Brunecky R, Bomble YJ, Himmel ME, Westpheling J (2015) Expression of the *Acidothermus cellulolyticus* E1 endoglucanase in *Caldicellulosiruptor bescii* enhances its ability to deconstruct crystalline cellulose. Biotechnology for biofuels 8:1

Cobucci-Ponzano B et al. (2010) A New Archaeal β-Glycosidase from *Sulfolobus solfataricus*: seeding a novel retaining β-glycan-specific glycoside hydrolase family along with the human non-lysosomal glucosylceramidase GBA2. Journal of Biological Chemistry 285:20691-20703

Corrêa JM et al. (2012) Expression and characterization of a GH39 β-xylosidase II from *Caulobacter crescentus.* Applied Biochemistry and Biotechnology 168:2218-2229

Czjzek M, Ficko-Blean E (2017) Probing the complex architecture of multimodular carbohydrate-active enzymes using a combination of small angle X-ray scattering and X-ray crystallography protein-carbohydrate interactions. Methods and Protocols:239-253

Davids T, Schmidt M, Böttcher D, Bornscheuer UT (2013) Strategies for the discovery and engineering of enzymes for biocatalysis. Current Opinion in Chemical Biology 17:215-220

de Souza AR et al. (2016) Engineering increased thermostability in the GH-10 endo-1,4-β-xylanase from *Thermoascus aurantiacus* CBMAI 756. International Journal of Biological Macromolecules 93:20-26

Diogo JA et al. (2015) Development of a chimeric hemicellulase to enhance the xylose production and thermotolerance. Enzyme and Microbial Technology 69:31-37

dos Santos CR, Cordeiro RL, Wong DW, Murakami MT (2015) Structural basis for xyloglucan specificity and α-d-Xyl p (1→ 6)-d-Glc p recognition at the− 1 subsite within the GH5 family. Biochemistry 54:1930-1942

Duan C-J, Huang M-Y, Pang H, Zhao J, Wu C-X, Feng J-X (2017) Characterization of a novel theme C glycoside hydrolase family 9 cellulase and its CBM-chimeric enzymes. Applied Microbiology and Biotechnology:1-15

Elleuche S (2015) Bringing functions together with fusion enzymes—from nature's inventions to biotechnological applications. Applied Microbiology and Biotechnology 99:1545-1556

Elleuche S, Schäfers C, Blank S, Schröder C, Antranikian G (2015) Exploration of extremophiles for high temperature biotechnological processes. Current Opinion in Microbiology 25:113-119

Elleuche S, Schröder C, Sahm K, Antranikian G (2014) Extremozymes—biocatalysts with unique properties from extremophilic microorganisms. Current Opinion in Biotechnology 29:116-123

Endy D (2005) Foundations for engineering biology. Nature 438:449-453

Espina G, Eley K, Pompidor G, Schneider TR, Crennell SJ, Danson MJ (2014) A novel β-xylosidase structure from *Geobacillus thermoglucosidasius*: the first crystal structure of a glycoside hydrolase family GH52 enzyme reveals unpredicted similarity to other glycoside hydrolase folds. Acta Crystallographica Section D: Biological Crystallography 70:1366-1374

Fatima B, Hussain Z (2015) Xylose isomerases from thermotogales. Journal of Animal and Plant Sciences 25.1: 10-18.

Ferrara MC, Cobucci-Ponzano B, Carpentieri A, Henrissat B, Rossi M, Amoresano A, Moracci M (2014) The identification and molecular characterization of the first archaeal bifunctional exo-β-glucosidase/N-acetyl-β-glucosaminidase demonstrate that family GH116 is made of three functionally distinct subfamilies. Biochimica et Biophysica Acta (BBA)-General Subjects 1840:367-377

Foumani M, Vuong TV, MacCormick B, Master ER (2015) Enhanced polysaccharide binding and activity on linear β-glucans through addition of carbohydrate-binding modules to either terminus of a glucooligosaccharide oxidase. PloS One 10:e0125398

Gao J, Wakarchuk W (2014) Characterization of five β-glycoside hydrolases from *Cellulomonas fimi* ATCC 484. Journal of Bacteriology 196:4103-4110

Gan R, Jewett MC (2016) Evolution of translation initiation sequences using in vitro yeast ribosome display Biotechnology and Bioengineering 113:1777-1786

Geisler-Lee J et al. (2006) Poplar carbohydrate-active enzymes. Gene identification and expression analyses. Plant Physiology 140:946-962

Gerday C, Glansdorff N (2007) Physiology and biochemistry of extremophiles. ASM Press, Washington

Ghatge SS et al. (2014) Characterization of modular bifunctional processive endoglucanase Cel5 from *Hahella chejuensis* KCTC 2396. Applied Microbiology and Biotechnology 98:4421-4435

Godoy AS, de Lima MZ, Camilo CM, Polikarpov I (2016) Crystal structure of a putative exo-β-1, 3-galactanase from *Bifidobacterium bifidum* S17. Acta Crystallographica Section F: Structural Biology Communications 72:288-293

Guerriero G, Hausman J-F, Strauss J, Ertan H, Siddiqui KS (2015) Destructuring plant biomass: Focus on fungal and extremophilic cell wall hydrolases. Plant Science 234:180-193

Han N, Miao H, Ding J, Li J, Mu Y, Zhou J, Huang Z (2017a) Improving the thermostability of a fungal GH11 xylanase via site-directed mutagenesis guided by sequence and structural analysis. Biotechnology for Biofuels 10:133

Han Y-B, Chen L-Q, Li Z, Tan Y-M, Feng Y, Yang G-Y (2017b) Structural insights into the broad substrate specificity of a novel endoglycoceramidase I belonging to a new subfamily of GH5 glycosidases. Journal of Biological Chemistry 292:4789-4800

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47-C52

Hendriks A, Zeeman G (2009) Pre-treatments to enhance the digestibility of lignocellulosic biomass. Bioresource technology 100:10-18

Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315:804-807

Hoffmam ZB et al. (2016) Xylan-specific carbohydrate-binding module belonging to family 6 enhances the catalytic performance of a GH11 endo-xylanase. New Biotechnology 33:467-472

Horiya S, Bailey JK, Krauss IJ (2017) Directed evolution of glycopeptides using mRNA display. Methods in Enzymology 597:83-141

Huang Z, Liu X, Zhang S, Liu Z (2014) GH52 xylosidase from *Geobacillus stearothermophilus*: characterization and introduction of xylanase activity by site-directed mutagenesis of Tyr509. Journal of Industrial Microbiology & Biotechnology 41:65-74

Huy ND et al. (2016) Characterization of a novel manganese dependent endoglucanase belongs in GH family 5 from *Phanerochaete chrysosporium*. Journal of Bioscience and Bioengineering 121:154-159

Khan MIM, Sajjad M, Sadaf S, Zafar R, Niazi UH, Akhtar MW (2013) The nature of the carbohydrate binding module determines the catalytic efficiency of xylanase Z of *Clostridium thermocellum*. Journal of Biotechnology 168:403-408

Kim JY, Nong G, Rice JD, Gallo M, Preston JF, Altpeter F (2016) *In planta* production and characterization of a hyperthermostable GH10 xylanase in transgenic sugarcane. Plant Molecular Biology 93:465-478

Klenk C, Ehrenmann J, Schütz M, Plückthun A (2016) A generic selection system for improved expression and thermostability of G protein-coupled receptors by directed evolution. Scientific Reports 6:21294

Kuuskeri J et al. (2016) Time-scale dynamics of proteome and transcriptome of the white-rot fungus *Phlebia radiata*: growth on spruce wood and decay effect on lignocellulose. Biotechnology for Biofuels 9:192

Lee Y-E, Lowe S, Henrissat B, Zeikus JG (1993) Characterization of the active site and thermostability regions of endoxylanase from *Thermoanaerobacterium saccharolyticum* B6A-RI. Journal of Bacteriology 175:5890-5898

Leuschner C, Antranikian G (1995) Heat-stable enzymes from extremely thermophilic and hyperthermophilic microorganisms. World Journal of Microbiology and Biotechnology 11:95-114

Li S, Yang X, Bao M, Wu Y, Yu W, Han F (2015) Family 13 carbohydrate-binding module of alginate lyase from *Agarivorans sp*. L11 enhances its catalytic efficiency and thermostability, and alters its substrate preference and product distribution. FEMS microbiology letters 362:10

Lin J-L, Wagner JM, Alper HS (2017) Enabling tools for high-throughput detection of metabolites: Metabolic engineering and directed evolution applications. Biotechnology Advances https://doi.org/10.1016/j.biotechadv.2017.07.005

Liu S, Ding S (2016) Replacement of carbohydrate binding modules improves acetyl xylan esterase activity and its synergistic hydrolysis of different substrates with xylanase. BMC Biotechnology 16:73

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495

López-Mondéjar R, Zühlke D, Větrovský T, Becher D, Riedel K, Baldrian P (2016) Decoding the complete arsenal for cellulose and hemicellulose deconstruction in the highly efficient cellulose decomposer *Paenibacillus* O199. Biotechnology for Biofuels 9:104

Ma F, Fischer M, Han Y, Withers SG, Feng Y, Yang G-Y (2016) Substrate engineering enabling fluorescence droplet entrapment for IVC-FACS-based ultrahigh-throughput screening. Analytical Chemistry 88:8587-8595

Matsuzawa T, Kaneko S, Yaoi K (2016) Improvement of thermostability and activity of *Trichoderma reesei* endo-xylanase Xyn III on insoluble substrates. Applied Microbiology and Biotechnology 100:8043-8051

McCartney L, Blake AW, Flint J, Bolam DN, Boraston AB, Gilbert HJ, Knox JP (2006) Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. Proceedings of the National Academy of Sciences of the United States of America 103:4765-4770

Mir BA, Mewalal R, Mizrachi E, Myburg AA, Cowan DA (2014) Recombinant hyperthermophilic enzyme expression in plants: a novel approach for lignocellulose digestion. Trends in Biotechnology 32:281-289

Mir B, Myburg, A, Mizrachi E, Cowan DA (2017) *In planta* expression of hyperthermophilic enzymes as a strategy for accelerated lignocellulosic digestion. Scientic Reports 7:11462

Montalvo-Rodriguez R, Haseltine C, Huess-LaRossa K, Clemente T, Soto J, Staswick P, Blum P (2000) Autohydrolysis of plant polysaccharides using transgenic hyperthermophilic enzymes Biotechnology and bioengineering 70:151-159

Montella S, Ventorino V, Lombard V, Henrissat B, Pepe O, Faraco V (2017) Discovery of genes coding for carbohydrate-active enzyme by metagenomic analysis of lignocellulosic biomasses. Scientific Reports 7:42623

Morrison JM, Elshahed MS, Youssef N (2016) A multifunctional GH39 glycoside hydrolase from the anaerobic gut fungus *Orpinomyces sp*. strain C1A. PeerJ 4:e2289

Naik SN, Goud VV, Rout PK, Dalai AK (2010) Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews 14:578-597

Nisha M, Satyanarayana T (2016) Characteristics, protein engineering and applications of microbial thermostable pullulanases and pullulan hydrolases. Applied Microbiology and Biotechnology 100:5661-5679

Oliveira C, Carvalho V, Domingues L, Gama FM (2015) Recombinant CBM-fusion technology—applications overview. Biotechnology Advances 33:358-369

Pan R, Hu Y, Long L, Wang J, Ding S (2016) Extra carbohydrate binding module contributes to the processivity and catalytic activity of a non-modular hydrolase family 5 endoglucanase from *Fomitiporia mediterranea* MF3/22. Enzyme and Microbial Technology 91:42-51

Pinard D et al. (2015) Comparative analysis of plant carbohydrate active enZymes and their role in xylogenesis. BMC Genomics 16:402

Pires VM et al. (2017) Stability and ligand promiscuity of type A carbohydrate-binding modules are illustrated by the structure of *Spirochaeta thermophila* StCBM64C. Journal of Biological Chemistry. M116:767541

Purnick PE, Weiss R (2009) The second wave of synthetic biology: from modules to systems. Nature Reviews Molecular Cell Biology 10:410-422

Reetz MT (2017) Recent advances in directed evolution of stereoselective enzymes. Directed Enzyme Evolution: Advances and Applications. Springer International Publishing, pp 69-99

Reetz MT, Carballeira JD, Peyralans J, Höbenreich H, Maichele A, Vogel A (2006) Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. Chemistry-A European Journal 12:6031-6038

Rothschild LJ, Mancinelli RL (2001) Life in extreme environments. Nature 409:1092-1101

Sainz-Polo MA, González B, Menéndez M, Pastor FJ, Sanz-Aparicio J (2015) Exploring multimodularity in plant cell wall deconstruction: structural and functional analysis of Xyn10C containing the CBM22-1-CBM22-2 tandem. Journal of Biological Chemistry. M115:659300

Sainz-Polo MA, Valenzuela SV, González B, Pastor FJ, Sanz-Aparicio J (2014a) Structural analysis of glucuronoxylan-specific Xyn30D and its attached CBM35 domain gives insights into the role of modularity in specificity. Journal of Biological Chemistry 289:31088-31101

Sainz-Polo MÁ, Valenzuela SV, Pastor FJ, Sanz-Aparicio J (2014b) Crystallization and preliminary X-ray diffraction analysis of Xyn30D from *Paenibacillus barcinonensis.* Acta Crystallographica Section F: Structural Biology Communications 70:963-966

Sajjad M, Khan MIM, Akbar NS, Ahmad S, Ali I, Akhtar MW (2010) Enhanced expression and activity yields of *Clostridium thermocellum* xylanases without non-catalytic domains. Journal of Biotechnology 145:38-42

Schneider WDH et al. (2016) *Penicillium echinulatum* secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. Biotechnology for Biofuels 9:66

Schückel J, Kračun SK, Willats WG (2016) High-throughput screening of carbohydrate-degrading enzymes using novel insoluble chromogenic substrate assay kits. Journal of Visualized Experiments 115

Smart M, Huddy RJ, Cowan DA, Trindade M (2017) Liquid phase multiplex high-throughput Screening of metagenomic libraries using p-nitrophenyl-linked substrates for accessory lignocellulosic enzymes metagenomics. Methods and Protocols 1539:219-228

Smolke CD (2009) Building outside of the box: iGEM and the BioBricks Foundation. Nature Biotechnology 27:1099-1102

Solomon KV et al. (2016) Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. Science 351:1192-1195

Tóth Á et al. (2016) Cloning, Expression and biochemical characterization of endomannanases from *Thermobifida* species isolated from different niches. PloS One 11:e0155769

Traxlmayr MW, Shusta EV (2017) Directed evolution of protein thermal stability using yeast surface display. Synthetic Antibodies: Methods and Protocols. 1575:45-65

Turner NJ (2009) Directed evolution drives the next generation of biocatalysts. Nature Chemical Biology 5:567-573

Turumtay H (2015) Cell wall engineering by heterologous expression of cell wall-degrading enzymes for better conversion of lignocellulosic biomass into biofuels. BioEnergy Research 8:1574-1588

Urbieta MS, Donati ER, Chan K-G, Shahar S, Sin LL, Goh KM (2015) Thermophiles in the genomic era: biodiversity, science, and applications. Biotechnology Advances 33:633-647

Valadares F et al. (2016) Exploring glycoside hydrolases and accessory proteins from wood decay fungi to enhance sugarcane bagasse saccharification. Biotechnology for Biofuels 9:110

Valenzuela SV, Diaz P, Pastor FJ (2012) Modular glucuronoxylan-specific xylanase with a family CBM35 carbohydrate-binding module. Applied and Environmental Microbiology 78:3923-3931

Venditto I et al. (2015) Family 46 carbohydrate-binding modules contribute to the enzymatic hydrolysis of xyloglucan and β-1,3–1,4-glucans through distinct mechanisms. Journal of Biological Chemistry 290:10572-10586

Verma AK, Goyal A (2014) *In silico* structural characterization and molecular docking studies of first glucuronoxylan-xylanohydrolase (Xyn30A) from family 30 glycosyl hydrolase (GH30) from *Clostridium thermocellum*. Molecular Biology 48:278-286

Verma AK et al. (2013) Overexpression, crystallization and preliminary X-ray crystallographic analysis of glucuronoxylan xylanohydrolase (Xyn30A) from *Clostridium thermocellum*. Acta Crystallographica Section F: Structural Biology and Crystallization Communications 69:1440-1442

Vilanova C, Porcar M (2014) iGEM 2.0 - refoundations for engineering biology. Nature Biotechnology 32:420-424

Voutilainen SP, Nurmi-Rantala S, Penttilä M, Koivula A (2014) Engineering chimeric thermostable GH7 cellobiohydrolases in *Saccharomyces cerevisiae.* Applied Microbiology and Biotechnology 98:2991-3001

Walker JA et al. (2015) Multifunctional cellulase catalysis targeted by fusion to different carbohydrate-binding modules. Biotechnology for Biofuels 8:220

Wang C, Dong D, Wang H, Müller K, Qin Y, Wang H, Wu W (2016a) Metagenomic analysis of microbial consortia enriched from compost: new insights into the role of *Actinobacteria* in lignocellulose decomposition. Biotechnology for Biofuels 9:22

Wang Y, Yu W, Han F (2016b) Expression and characterization of a cold-adapted, thermotolerant and denaturant-stable GH5 endoglucanase Celal_2753 that withstands boiling from the psychrophilic bacterium *Cellulophaga algicola* IC166T. Biotechnology Letters 38:285-290

Xue X et al. (2015) The N-terminal GH10 domain of a multimodular protein from *Caldicellulosiruptor bescii* is a versatile xylanase/β-glucanase that can degrade crystalline cellulose. Applied and Environmental Microbiology 81:3823-3833

Yang Y et al. (2015) A mechanism of glucose tolerance and stimulation of GH1 β-glucosidases. Scientific Reports 5:17296

Yi Z, Su X, Revindran V, Mackie RI, Cann I (2013) Molecular and biochemical analyses of CbCel9A/Cel48A, a highly secreted multi-modular cellulase by *Caldicellulosiruptor bescii* during growth on crystalline cellulose. PloS One 8:e84172

Zang H et al. (2015) A novel thermostable GH5_7 β-mannanase from *Bacillus pumilus* GBSW19 and its application in manno-oligosaccharides (MOS) production. Enzyme and Microbial Technology 78:1-9

Zhang H et al. (2017) Complete genome sequence of the cellulose-producing strain *Komagataeibacter nataicola* RZS01. Scientific Reports 7:4431

Zhang X et al. (2015) Subsite-specific contributions of different aromatic residues in the active site architecture of glycoside hydrolase family 12. Scientific reports 5:18357

Ziegler MT, Thomas SR, Danna KJ (2000) Accumulation of a thermostable endo-1,4-β-D-glucanase in the apoplast of *Arabidopsis thaliana* leaves. Molecular Breeding 6:37-46

CHAPTER 2:

Comparative analyses of hyperthermophile genomes reveals multiple lignocellulose degrading CAZyme domains

Jonathan Botha[1,2,4], Eshchar Mizrachi[2,3], Alexander A. Myburg[2,3], Don A. Cowan[1,2,4]

[1]Centre for Microbial Ecology and Genomics, Department Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[2]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[3]Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[4]Genomics Research Institute (GRI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

In this chapter, I generated all CAZyme domain datasets using bioinformatics scripts developed in collaboration with Ms Desre Pinard. I performed all analyses, generated all figures and prepared this manuscript. Prof. D.A. Cowan, Prof. A.A. Myburg and Dr E. Mizrachi provided advice, direction and supervision during the planning of the study, as well as guidance during the interpretation of the results. They also performed critical revision of the manuscript.

2.1 Abstract

Biomaterials may be able to replace current petrochemically derived products, but require polysaccharide biopolymers for synthesis. Lignocellulosic biomass could provide the required biopolymers, but is recalcitrant to enzymic digestion. Currently, industrial pre-treatments are used to reduce recalcitrance and extract biopolymers, but are costly, difficult to apply and produce waste products, resulting in lower yields. Thermostable Carbohydrate Active Enzymes (CAZymes) from extremely thermophilic organisms, which grow at temperatures exceeding 70°C, may be used in industrial processes which often take place under extreme conditions, such as high temperature. In an attempt to identify a list of thermostable CAZyme domains, we performed a scan using the HMMER 3.2 tool suite on a number of proteomes from extremely thermophilic organisms, using CAZyme family HMMs obtained from the dbCAN database. A number of predicted CAZyme domains were identified from organisms belonging to the domains of archaea and bacteria. Differences in CAZyme composition and abundance were investigated between and within archaea and bacteria, and the overall coverage of currently known CAZyme domains within the dataset was assessed. In total, 3773 CAZyme domains were identified, which represented 30%, 45%, 33%, 42%, 88% and 38% of currently known GTs, GHs, PLs, CBMs, CEs and AAs, respectively. Differences in CAZyme composition between archaea and bacteria were apparent, possibly due to the different lifestyles and ecological niches the organisms inhabit. Domains for the degradation of the main lignocellulosic biopolymers, cellulose and xylan, were prevalent in both archaea and bacteria.

2.2 Introduction

There is currently a need for sustainable alternatives to petrochemically derived fuels, plastics and adhesives. Products synthesised from biological materials (i.e., bioproducts) are a promising substitute, and could help to ameliorate the environmental and economic challenges associated with petrochemical products (Naik et al. 2010). Lignocellulosic biomass is a convenient feedstock for biomaterial synthesis, due to a lack of competition with food crops, an ability to grow in diverse environments and high cellulose, hemicellulose and lignin content (Hendriks and Zeeman 2009; Himmel et al. 2007). However, it is highly recalcitrant to digestion, often requiring harsh pre-treatments (Alvira et al. 2010; Hendriks and Zeeman 2009) before enzyme loading in order to obtain a hydrolysed product. This is due to the cross-linking and structural complexity of the biopolymers that make up the secondary cell wall (SCW) of lignocellulosic biomass, namely cellulose, hemicellulose and lignin (Cosgrove and Jarvis 2012).

A key strategy for the degradation of lignocellulosic biomass is through the use of Carbohydrate Active enZymes (CAZymes), enzymes with activities on oligosaccharides, polysaccharides and glycoconjugates (Cantarel et al. 2009; Lombard et al. 2014). CAZymes are classified into families based on their domain structures and the functions they perform. Glycoside Hydrolases (GHs) and Polysaccharide Lyases (PLs) hydrolyse glycosidic bonds, Carbohydrate Esterases (CEs) break ester bonds between biopolymers and chemical side chains, Glycosyl Transferases (GTs) synthesise glycosidic bonds, Carbohydrate Binding Modules (CBMs) recognise and bind to specific carbohydrates and Auxilliary Activity enzymes (AAs) act on carbohydrates, but do not fall into any of the other classes (Cantarel et al. 2009; Lombard et al. 2014). CAZymes are modular in nature, consisting of single domains, or multiple domains in different configurations that allow for variety in specificity and function (André et al. 2014).

Most CAZymes are not suited to the extreme temperatures, pressures and pHs required of most industrial pre-treatments of lignocellulosic biomass. Extremely thermophilic organisms grow optimally at temperatures exceeding 70°C (Gerday and Glansdorff 2007; Leuschner and Antranikian 1995) and provide a pool of thermostable CAZymes for possible use in industrial pre-treatments of lignocellulose that require extreme temperatures and pHs (Blumer-Schuette et al. 2014). Additionally, the efficiency of lignocellulose degradation could be further increased through rational engineering of thermostable CAZymes and CAZyme domains (Botha et al. 2017).

In order to catalogue the potential of CAZymes for industrial processing applications using synthetic biology and rational design, we identified and compared the predicted CAZyme complement of 64 extremely thermophilic proteomes, spanning the domains of archaea and bacteria. We provide a list of CAZyme domains using the latest Hidden Markov Model (HMM) profiles for each CAZyme family, updating the previously published dataset in Botha et al. (2017). The differences and distributions of CAZyme domain families across these organisms were investigated, with special attention applied to CAZyme domains with previously described activities on lignocellulosic biopolymers. This study provides insight into extremely thermophilic CAZyme domains as a resource for industrial processing of lignocellulosic biomass.

2.3 Materials and Methods

2.3.1 Acquisition of organism proteomes

In order to obtain the proteomes of extremely thermophilic organisms, the Genome OnLine Database website (https://gold.jgi-psf.org/) was interrogated for completely sequenced and published genomes, filtered with the descriptive field of "hyperthermophile". The optimal growth temperature (T_{opt}) of every organism identified was investigated, and those with T_{opt} values lower than 70^{o}C were removed from the dataset. A total of 64 organisms were identified, with 15 and 49 organisms belonging to the domains of bacteria and archaea, respectively. The predicted proteome for each organism was downloaded from the Ensembl Bacteria website (http://bacteria.ensembl.org/).

2.3.2 CAZyme domain identification and analysis

HMMER analyses were performed using HMMER 3.2 windows binary files available from the HMMER website [http://hmmer.org/ (Eddy 1998)] in order to identify putative CAZyme domains in the selected proteomes. The HMMs for all CAZyme domain families, the lengths of the HMMs and the parsing script were downloaded from the dbCAN web database [http://csbl.bmb.uga.edu/dbCAN/download.php (Yin et al. 2012)]. The dbCAN HMM profiles were prepared for HMMER with the hmmpress tool. Each proteome was interrogated for sequence homology to known CAZyme domains using the hmmscan tool. Results were submitted to a custom dbCAN parsing script in order to convert them to a human readable format. Abundance and diversity of CAZyme domains within and across all characterised organisms were determined using custom R language scripts (Supplementary File 2.1). Heatmaps were generated using the iheatmapr package for R [(Team 2013) Supplementary File 2.1].

2.3.3 Rarefaction curve

A rarefaction curve was plotted from $S_{(est)}$-values derived from the EstimateS program v 9.1.0 (http://viceroy.eeb.uconn.edu/estimates/) in order to estimate the likelihood that more CAZyme domains would be identified if more proteomes were examined. The abundance data for each CAZyme domain per organism was formatted as tab-separated values and uploaded into the program. Diversity statistics (including the $S_{(est)}$-value) were calculated using default parameters. The $S_{(est)}$-values were then plotted to achieve a rarefaction curve.

2.4 Results

2.4.1 Extremely thermophilic organisms have abundant GT and GH CAZyme domains

In this study, proteomes were identified from 64 extremely thermophilic organisms that grow optimally across a substantial temperature and pH range (Supplementary Data 2.1). Bacterial phyla included Aquificae, Thermotogae and Firmicutes, while archaeal phyla included Crenarchaeota, Nanoarchaeota and Euryarchaeota. A total of 3773 CAZyme domains were identified from these proteomes using HMMER 3.2 (http://hmmer.org/; Eddy 1998), including 27 PL, 130 AA, 360 CE, 422 CBM, 1227 GH and 1445 GT CAZyme domains. The remaining 162 domains fell within structural dockerin, scaffoldin and cohesin families (Figure 2.1, Figure 2.2, Supplementary Data 2.2). GT and GH domains were by far the most abundant CAZyme domain classes, making up 38% and 33% of all identified domains, respectively (Figure 2.1, Figure 2.2). The most common CAZyme domains identified in the study per class were GT4 (520 domains) and GT2 (435 domains) for GT domains, GH57 (142 domains), GH109 (128 domains) and GH13 (120 domains) for GH domains, CBM50 (84 domains), CBM48 (44 domains) and CBM22 (37 domains) for CBM domains, and CE4 (71 domains) and CE1 (59 domains) for CE domains (Figure 2.1, Figure 2.2).

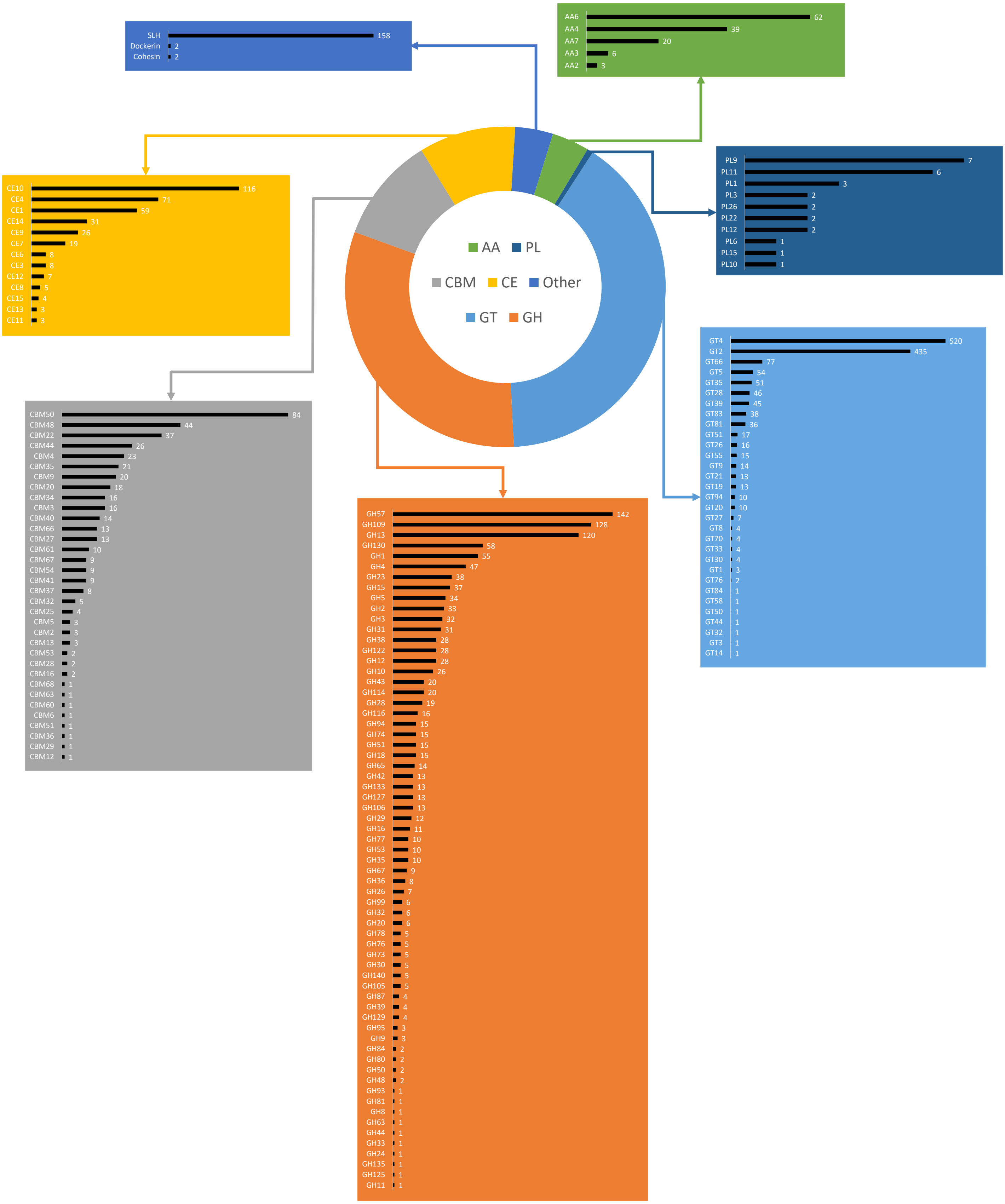


Figure 2.1 Summary of CAZyme domain dataset. The ring in the centre indicates the proportions of each CAZyme class of which the dataset is comprised. The coloured boxes show CAZyme families present in each class, as well as the absolute count of each family in the given class. GH: Glycoside Hydrolases, CBM: Carbohydrate-Binding Modules, GT: Glycosyl Transferases, CE: Carbohydrate Esterases, PL: Polysaccharide Lyases and AA: Auxilliary Activities. Other: Category comprising structural and other non-CAZyme domains

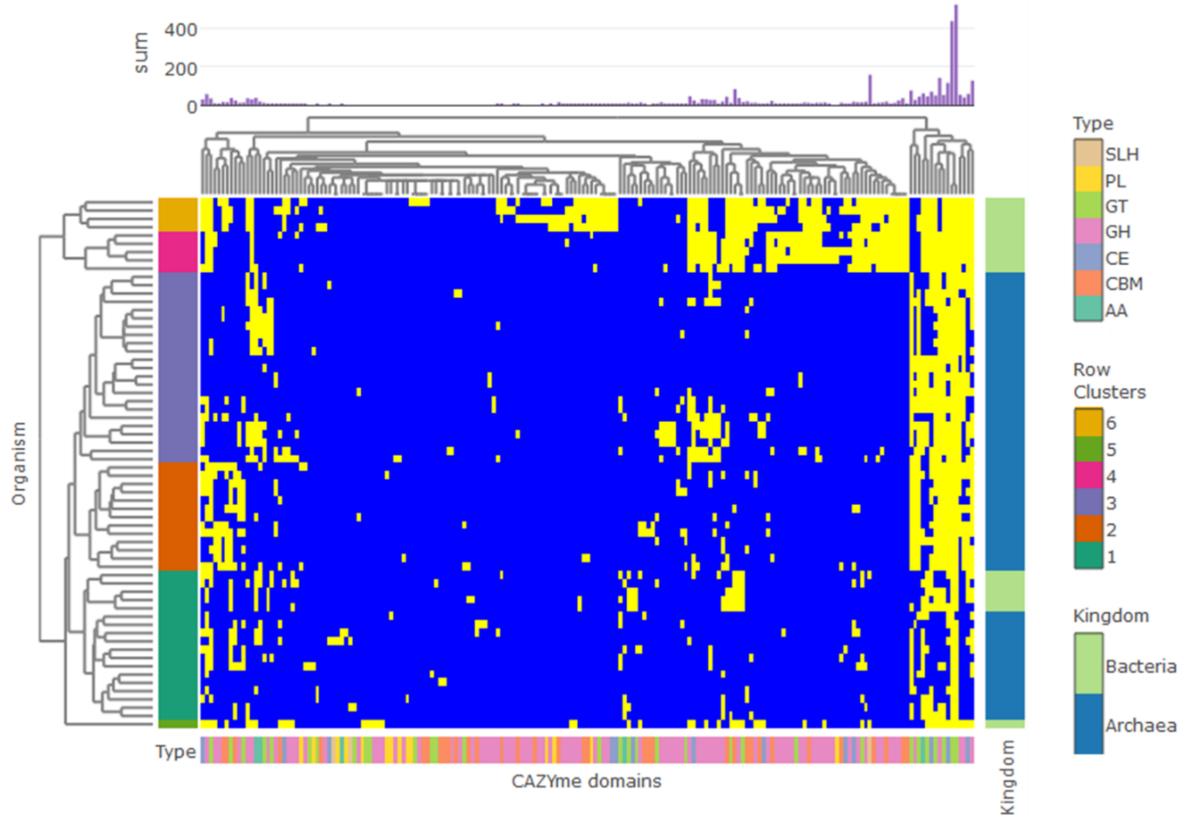


Figure 2.2 Heatmap of CAZyme domains in all organisms studied. The CAZyme domain types (GH, GT, CBM, etc.) are represented by the bar below the heatmap. Organisms are clustered according to presence (yellow) or absence (blue) of each CAZyme domain. Clusters are represented by the bar to the left of the heatmap. The kingdom to which each organism belongs (bacteria or archaea) is indicated by the bar to the right of the heatmap. The phylogram above the heatmap indicates clustering of the CAZyme domain families. The absolute counts (sum) of each CAZyme domain is indicated by the histogram above the phylogram at the top of the heatmap. The key is on the far right of the heatmap.

2.4.2 Extremely thermophilic bacteria and archaea have unique CAZyme domain composition

Differences were observed in the content and abundance of CAZyme domains between bacterial and archaeal proteomes. GH domains were more abundant than GT domains in bacterial proteomes, with 10 of the 15 bacterial proteomes containing more predicted GH than GT domains, compared with only 4 of the 49 archaeal proteomes investigated (Figure 2.3, Supplementary Data 2.2). Even though fewer extremely thermophilic and extremely thermophilic bacteria were available for the study (15 bacteria as opposed to 50 archaea), 21 CAZyme domain families were identified as unique to bacteria, as well as 21 unique CAZyme families in archaea. When normalised for genome size (0.022 and 0.051 CAZy families per kb in bacteria vs archaea, *p* < 0.001, Figure 2.4) and number of coding genes (0.021 and 0.053 CAZy families per coding gene in bacteria vs archaea, *p* < 0.001, Figure 2.4, Supplementary Data 2.3), bacterial proteomes appear to have more predicted CAZyme domains than archaeal proteomes. CBMs and CEs are also more abundant in bacterial proteomes (Figure 2.2, Figure 2.3).

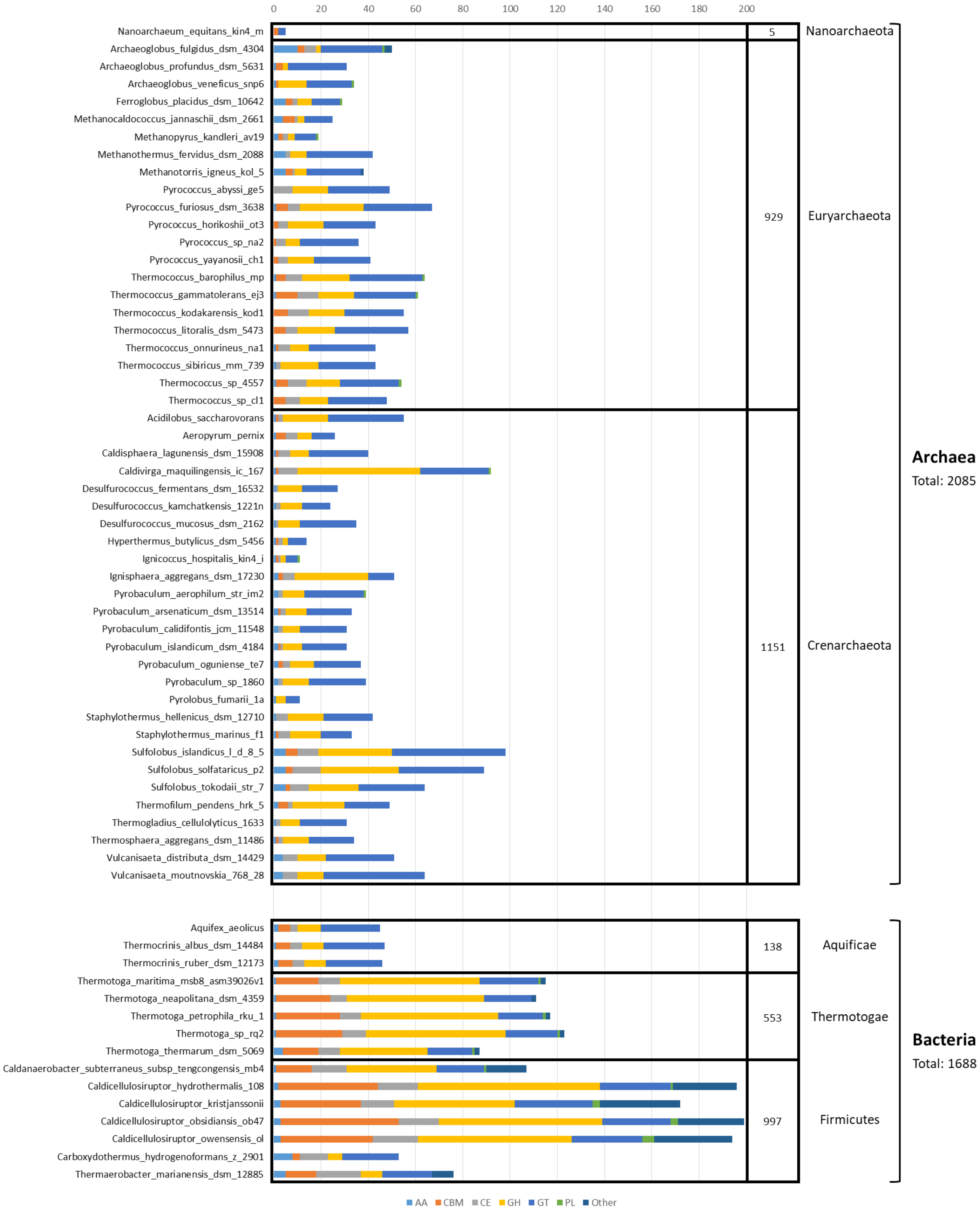


Figure 2.3 Absolute abundance of predicted CAZyme classes in each organisms studied. The column on the left contains the name of the organism. The bar graph shows the absolute number of each class of CAZyme, represented by a colour. AA: Auxilliary Activities, GT: Glycosyl Transferases, CE: Carbohydrate Esterases, CBM: Carbohydrate-Binding Modules, Other: Structural proteins such as dockerin, scaffoldin and cohesin, PL: Polysaccharide Lyases and GH: Glycoside hydrolases. On the right, the phylum and domain of the organisms are indicated, as well as the total number of predicted CAZymes in each.

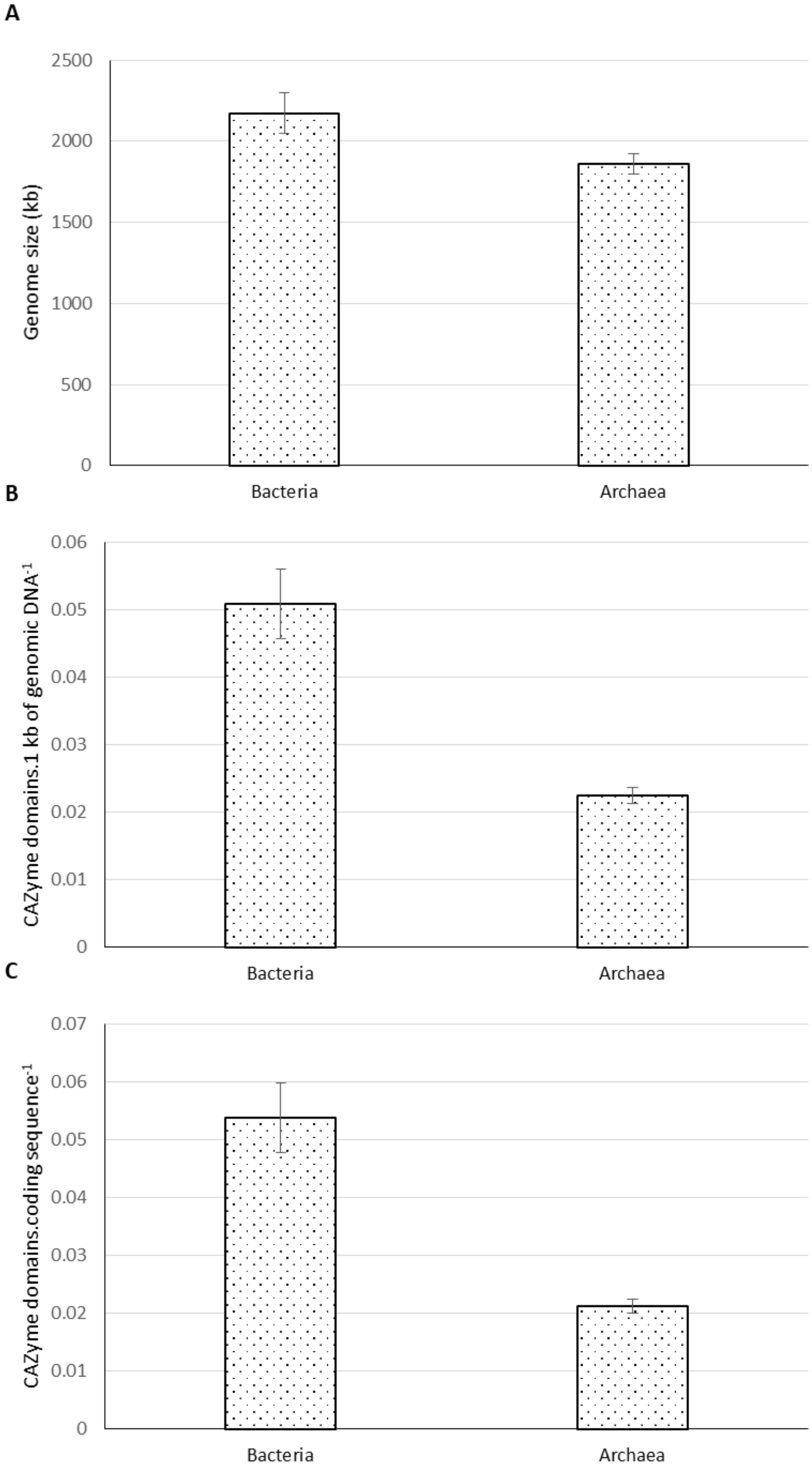


Figure 2.4 Average CAZyme domains per organism in bacteria and archaea. A: Average genome size in kb. B: Average CAZyme domains per 1 kb of genomic DNA. C: Average CAZyme domains per coding gene for all organisms studied.

2.4.3 Lignocellulose binding and degrading CAZyme domains are prevalent in hyperthermophilic proteomes

A number of CAZyme domain families with xylan binding function (CBM13, CBM22, CBM35, CBM36, CBM54 and CBM60), cellulose binding function (CBM3, CBM16, CBM28 and CBM44), or both (CBM2, CBM4, CBM6, CBM9, CBM37) were identified (Table 2.1, Supplementary Data 2.4). Some of these domains, such as CBM4, CBM9, CBM13, CBM37 and CBM54, were well represented in bacterial but not in archaeal proteomes. Others, such as CBM3, CBM9, CBM22 and CBM28, were only present in bacterial proteomes, while CBM44 was only identified in archaeal proteomes. CBM35 domains were abundant in both archaeal and bacterial proteomes. The remainder of the cellulose and xylan binding CBM domains, namely CBM2, CBM6, CBM16, CBM36 and CBM60 were present in very few, or single proteomes.

Xylan degrading domains (GH10, GH11, GH39, GH43, GH67 and GH116), cellulose degrading domains (GH1, GH9, GH12, GH16, GH26, GH44, GH48, GH74 and GH94), and domains which hydrolyse both cellulose and xylan (GH3, GH5, GH8, GH30 and GH51) were identified in the dataset (Table 2.1, Supplementary Data 2.4). Both of the well-characterised and exclusively xylan-targeting domains, GH10 and GH11, were detected only in bacterial proteomes, with GH11 only being observed in a single Firmicute. GH9, GH26, GH30, GH44, GH67, GH74, GH81 and GH87 were only observed in bacterial proteomes, while GH51 and GH94 were mainly observed in bacterial proteomes, but were also identified at lower numbers in archaeal proteomes. GH1, GH3, GH5, GH12, GH16, GH39 and GH43 were relatively well represented in archaeal and bacterial proteomes, while GH8, GH9, GH39, GH44, GH48 and GH116 were observed in relatively few proteomes. GH116 appears to be the only identified lignocellulose targeting GH domain exclusive to the archaeal proteomes examined in this study.

Additionally, a number of CE domain families (CE1, CE3, CE4, CE6, CE7 and CE12) with acetyl-xylan esterase activities were observed (Table 2.1, Supplementary Data 2.4). Most CEs were well represented in both archaeal and bacterial proteomes, with the exception of the CE6 domain, which was only observed in bacterial proteomes, in the Firmicutes.

Table 2.1 CAZyme domain families identified in the study that have previously reported activities on cellulose and xylan.

CAZyme domain family	Activity on cellulose	Activity on xylan	Number of domains identified
CBM13		x	3
CBM16	x		2
CBM2	x	x	3
CBM22		x	37
CBM28	x		2
CBM3	x		16
CBM35		x	21
CBM36		x	1
CBM37	x	x	8
CBM4	x	x	23
CBM44	x		26
CBM54		x	9
CBM6	x	x	1
CBM60		x	1
CBM9	x	x	20
CE1		x	59
CE12		x	7
CE3		x	8
CE4		x	71
CE6		x	8
CE7		x	19
GH1	x		55
GH10		x	26
GH11		x	1
GH116		x	16
GH12	x		28
GH16	x		11
GH26	x	x	7
GH3	x	x	32
GH30	x	x	5
GH39		x	4
GH43		x	20
GH44	x		1
GH48	x		2
GH5	x	x	34
GH51	x	x	15
GH67		x	9
GH74	x		15
GH8	x	x	1
GH9	x		3
GH94	x		15

2.4.4 Low frequency CAZyme domains

Eight CBM domain families were present only in single proteomes (Table 2.2, Supplementary Data 2.5). They were CBM6, CBM12, CBM29, CBM36, CBM51, CBM60, CBM63 and CBM68. Twelve CBM domain families were phylum specific, namely CBM2, CBM3, CBM6, CBM12, CBM28, CBM29, CBM32, CBM36, CBM51, CBM60, CBM63 and CBM68 (Table 2.2). Ten GH domains were only present in one studied proteome, specifically GH8, GH11, GH24, GH33, GH44, GH63, GH81, GH93, GH125 and GH135 (Table 2.2). Furthermore, 20 GH and 11 GT domain families were phylum specific (Table 2.3. Supplementary Data 2.6), while seven GT domain families were found in single proteomes, namely GT3, GT14, GT32, GT44, GT50, GT58 and GT84 (Table 2.2). Six PL domain families were phylum specific and three PL domain families (PL6, PL10 and PL15) were found in proteomes of single organisms (Table 2.3). CE domain families did not have members detected in single proteomes, but three phylum specific CE domain families were identified, specifically CE11 (Aquificae), CE6 and CE15 (Firmicutes, Table 2.3). A number of domains present in single organism proteomes had putative lignocellulose associated activity (Table 2.2, Supplementary Data 2.4).

Table 2.2 CAZyme domains present in single organisms

CAZyme domain family	Substrate[a]
CBM12	
CBM29	
CBM36	X
CBM51	
CBM6	C, X
CBM60	X
CBM63	
CBM68	
GH11	X
GH125	
GH135	
GH24	
GH33	
GH44	C
GH48	C
GH63	
GH8	C, X
GH81	
GH87	
GH93	
GT14	
GT3	
GT32	
GT44	
GT50	
GT58	
GT84	
PL10	
PL15	
PL6	

[a]The known substrate target of the domain (C: Cellulose, X: Xylan)

Table 2.3 Phylum specific CAZyme domain families. Numbers indicate the absolute count of a given domain in each phylum. N: Nanoarchaeota, E: Euryarchaeota, C: Crenarchaeota, A: Aquificae, T: Thermotogae and F: Firmicutes.

	Archaea			Bacteria		
CAZyme family	**N**	**E**	**C**	**A**	**T**	**F**
AA2	0	3	0	0	0	0
CBM12	0	1	0	0	0	0
CBM2	0	3	0	0	0	0
CBM28	0	0	0	0	0	2
CBM29	1	0	0	0	0	0
CBM3	0	0	0	0	0	16
CBM32	0	0	0	0	0	5
CBM36	0	0	0	0	0	1
CBM51	0	0	0	0	0	1
CBM6	0	0	0	0	0	1
CBM60	0	0	1	0	0	0
CBM63	0	0	1	0	0	0
CBM68	1	0	0	0	0	0
CE11	0	0	0	3	0	0
CE15	0	0	0	0	0	4
CE6	0	0	0	0	0	8
cohesin	0	2	0	0	0	0
dockerin	0	2	0	0	0	0
GH11	0	0	0	0	0	1
GH116	0	0	16	0	0	0
GH125	0	0	0	0	0	1
GH129	0	0	0	0	0	4
GH135	0	1	0	0	0	0
GH24	0	0	0	0	0	1
GH30	0	0	0	0	0	5
GH33	0	0	1	0	0	0
GH44	0	0	0	0	0	1
GH48	0	0	0	0	0	2
GH5	0	0	2	0	0	0
GH63	0	0	1	0	0	0
GH8	0	0	0	1	0	0
GH80	0	0	2	0	0	0
GH81	0	0	0	0	0	1
GH84	0	0	2	0	0	0
GH87	0	0	0	0	0	4
GH9	0	0	0	0	0	3
GH93	0	0	1	0	0	0
GH95	0	0	0	0	0	3
GT14	0	0	0	0	0	1
GT3	0	1	0	0	0	0
GT32	0	0	0	0	0	1
GT33	0	4	0	0	0	0
GT44	0	0	0	0	0	1
GT50	0	0	1	0	0	0
GT58	0	1	0	0	0	0
GT70	0	0	0	0	0	4
GT76	0	2	0	0	0	0

Table 2.3 (continued) Phylum specific CAZyme domain families. Numbers indicate the absolute count of a given domain in each phylum. N: Nanoarchaeota, E: Euryarchaeota, C: Crenarchaeota, A: Aquificae, T: Thermotogae and F: Firmicutes.

	Archaea			Bacteria		
CAZyme family	N	E	C	A	T	F
GT8	0	0	0	0	4	0
GT84	0	0	0	0	0	1
PL10	0	1	0	0	0	0
PL12	0	0	2	0	0	0
PL15	0	1	0	0	0	0
PL26	0	0	0	0	0	2
PL3	0	0	0	0	0	2
PL6	0	1	0	0	0	0

2.4.5 Overall coverage of CAZymes.

The CAZyme domains identified in this study represent a significant proportion of all CAZyme domain families identified. In most cases, a third to half of all CAZyme domain families in a particular class are present in the dataset [with the exception of CE domains, of which approximately 80% of CE domain classes are represented (Figure 2.5)]. A rarefaction curve of CAZyme domains identified per analysed proteome shows that a plateau had not been reached (Figure 2.6); therefore, if more proteomes were examined, more CAZyme domain classes would likely be identified.

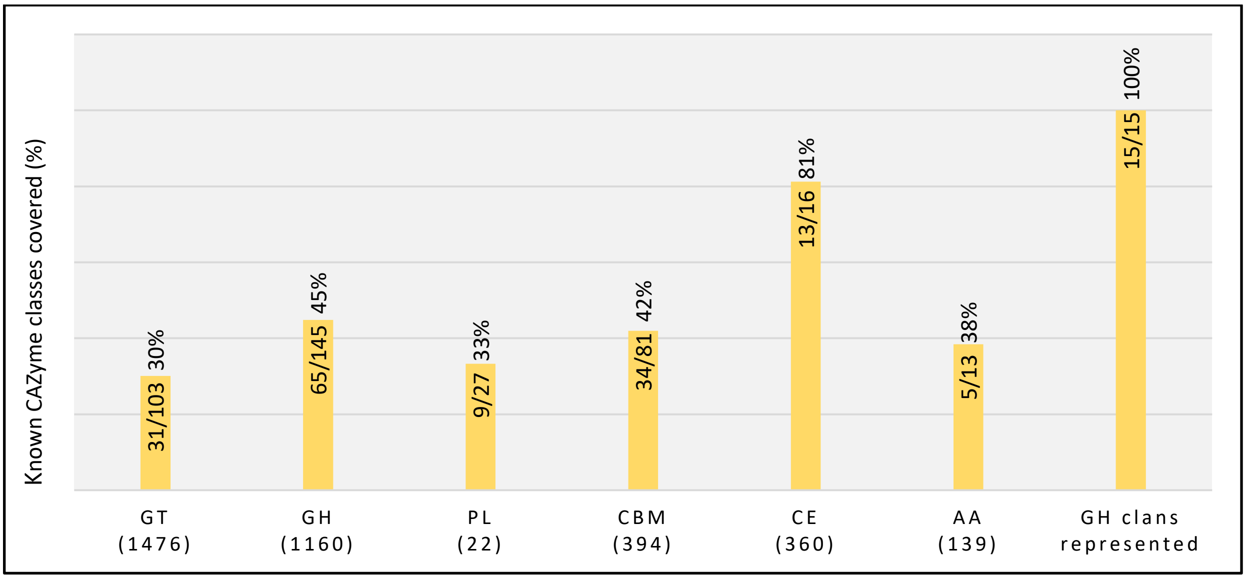


Figure 2.5 The overall coverage of CAZyme families identified from extremely thermophilic organisms, expressed as a percentage, updated from Botha et al. (2017). The CAZyme class is listed on the X-axis. GT: Glycosyl Transferase, GH: Glycoside hydrolase, PL: Polysaccharide Lyase, CBM: Carbohydrate-Binding Module, CE: Carbohydrate Esterase and AA: Auxilliary Activity. The number listed below each class is the absolute count of domains identified in each class for extremely thermophilic organisms. The numbers in the bars indicate the absolute proportion of all currently known CAZyme domain families present in the extremely thermophilic organisms.

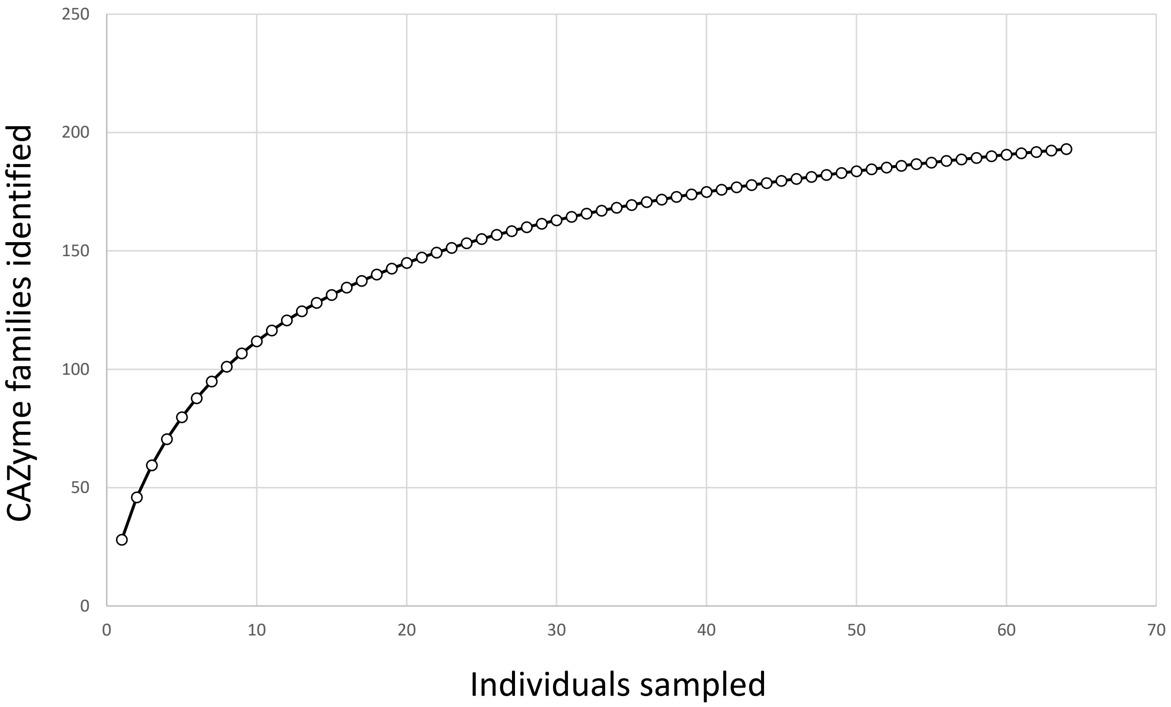


Figure 2.6: Rarefaction curve showing the number of CAZyme families identified in the dataset, per genome sampled. N=64

2.5 Discussion

2.5.1 CAZyme abundance and diversity in archaea and bacteria

The proteomes used in this study were derived from organisms that inhabit a range of ecological niches, exhibit different nutritional strategies and growing conditions (Supplementary Data 2.1) and represent a range of phyla, including Nanoarchaeota, Euryarchaeota, Crenarchaeota, Aquificae, Thermotogae and Firmicutes. It is, therefore, expected that the different proteomes would exhibit considerable diversity in content and abundance of CAZyme domains, and that such CAZymes should exhibit diverse temperature and pH optima. Industrial pre-treatments of lignocellulose require a range of pHs and temperatures (Alvira et al. 2010), and it may be possible to identify a CAZyme or CAZyme domain that will operate efficiently in these conditions (Blumer-Schuette et al. 2014).

The CAZyme domain content of the organisms in this study (Supplementary Data 2.1) varied between individuals, phyla and domains of life (Figure 2.2, Supplementary Data 2.2). Scaffoldin (SLH) domains were detected in most bacterial proteomes, and cohesin and dockerin were detected only in the *A. fulgidus* proteome, indicating the possible presence of extremely thermophilic cellulosome complexes (Artzi et al. 2017; Bayer et al. 2004). In general, bacterial proteomes had more GH, CBM, CE and PL domains than archaeal proteomes, while GT domains where roughly similar in number in archaeal and bacterial proteomes (Figure 2.3). Most bacterial proteomes had more GH than GT domains, while the opposite was true for archaeal proteomes. Differences in nutritional strategies may account for the differing CAZyme domain content between the organisms in this study, with the majority of bacteria able to utilise lignocellulosic biomass as a carbon source (Supplementary Data 2.1). In order to do so, lignocellulosic biopolymers need to be degraded to monomers (or short oligosaccharides), which would require numerous CBM, PL, GH and CE domain activities. Many archaea use other carbon sources (Supplementary Data 2.1) and can generate energy from inorganic substrates (Berg et al. 2010), reducing the need for complex carbohydrate biopolymer deconstruction.

Cell envelope (specifically cell wall and S-layer) composition may also contribute to the difference in CAZyme domain content between bacterial and archaeal proteomes. Bacterial cell walls are typically constructed from peptidoglycan, consisting of repeating N-acetylglucosamine and N-acetylmuramic acid sugars linked in a β-1,4 configuration (Schleifer and Kandler 1972), while most archaeal cell walls consist of an S-layer composed of glycoproteins (Rodrigues-Oliveira et al. 2017), supplemented with linked glycans, as well as polysaccharides biopolymers such as pseudomurien and methanochondroitin (Albers and Meyer 2011). Since carbohydrate polymers comprise the main portion of bacterial cell walls, bacteria would need an expanded CAZyme domain repertoire to synthesize, maintain and modify them (Cabeen and Jacobs-Wagner 2005), accounting for the relatively larger pool of CBM, GH, PL and CE domains identified from bacterial proteomes. While carbohydrates are also present in archaeal cell walls, they are often not the primary structural component, and are included as components of glycoproteins or other biopolymers (Albers and Meyer

2011; Rodrigues-Oliveira et al. 2017). Both glycosylation of glycoproteins and synthesis of carbohydrate biopolymers are mediated by GT domains (Lombard et al. 2014), so high representation of GT domains in bacterial and archaeal proteomes is expected. Additionally, differences in synthesis and composition of carbohydrate-rich biofilms may contribute to variable CAZyme content in bacterial and archaeal proteomes (Orell et al. 2013; Orell et al. 2017).

While CAZyme content differs between archaeal and bacterial proteomes, some CAZyme domain families are highly represented across both. The most abundant CAZyme domain class across all organisms was GT domains, followed by GH, CBM, CE, AA and PL domains, respectively (Figure 2.1, Figure 2.2, Figure 2.3, Supplementary Data 2.2). GT domains are important for many biosynthetic processes (Lairson et al. 2008) and an organism would require a diverse repertoire of GT domains to mediate them. The most prevalent CAZyme domains were GT2 and GT4 (Figure 2.1), with GT2 being the only CAZyme domain to be present in every proteome studied, and GT4 being present in all but one proteome, that of *Ignicoccus hospitalis.* Even though *I. hospitalis* lacks predicted GT4 domains in its proteome, it has a symbiotic relationship with and is known to host the *Nanoarchaeum equitans* (Jahn et al. 2007; Podar et al. 2008), which has one predicted GT4 domain (Supplementary Data 2.3). GT2 and GT4 CAZymes are involved in the synthesis of many polymers (Lairson et al. 2008), and are some of the only GTs present in ancient archaea. They are possibly the evolutionary origin of most current GT families (Lairson et al. 2008) and are expected to be present in most GT-carrying organisms.

The most commonly predicted CBM domains in the dataset were CBM50, CBM48 and CBM22 (Figure 2.1). While the CBM50 domain was the most abundant CBM domain identified in the study, it was seen exclusively in bacterial proteomes (Supplementary Data 2.3). CBM50 recognizes and binds to linked residues of N-acetylglucosamine (GlcNAc; Steen et al. 2003) and N-acetylmuramic acid (MurNAc; Onaga and Taira 2008), the monomers of peptidoglycan and chitin, which make up the cell walls of bacteria (Schleifer and Kandler 1972) and fungi (Peberdy 1990), respectively. CBM50 domains are therefore important for the synthesis and maintenance of bacterial cell walls, potentially accounting for the prevalence of CBM50 domains in the dataset. CBM48 was the second most abundant CBM domain and was identified in archaeal and bacterial proteomes (Figure 2.1, Figure 2.2, Supplementary Data 2.2). CBM48 binds glycogen (Cantarel et al. 2009; Lombard et al. 2014), a well-known biopolymer that serves as an energy reserve for archaea (Horcajada et al. 2006) and bacteria (Dawes and Senior 1973). CBM48 domains could therefore play a role in synthesis and metabolism of glycogen in archaea and bacteria. CBM22 mainly binds xylan (Carvalho et al. 2015). This domain was only identified in the bacterial proteomes in this study, and especially in *Thermotoga and Caldicellulosiruptor spp* proteomes (Supplementary Data 2.3), both of which are known for producing thermostable lignocellulose degrading enzymes able to hydrolyse β-linked glycosidic bonds (Blumer-Schuette et al. 2014; Blumer-Schuette et al. 2008; Gibbs et al. 2000). Additionally, CBM22 can have a thermostabilizing effect (Khan et al. 2013; Lee et al. 1993), explaining the occurrence of CBM22 domains in extremely thermophilic xylanases. Interestingly, the most abundant CBM domain in archaeal proteomes (CBM44; Supplementary Data 2.3) can bind cellulose and xyloglucan (Najmudin et al. 2006), indicating potential to process lignocellulose.

Of the most common GH domains (GH57, GH109, GH13) identified in this study, GH57 domains target and hydrolyse α-linkages in many sugars such as starch and glycogen, galacto-oligosaccharides, galactomannans, galactolipids, pullulan and amylopectin (Lombard et al. 2014). As such, GH57 may play a role in the release of energy from stored glycogen in times of nutritional scarcity. The GH109 domain is prevalent in both bacteria and archaea, and has N-acetylgalactosaminidase activity (Liu et al. 2007; Lombard et al. 2014) able to hydrolyse the bonds between the sugars and peptides in glycoproteins. GH109 domains could play a role in cell envelope construction in archaea and bacteria, allowing for the modification of glycoproteins, which are a prominent component of many bacterial and archaeal S-layers (Albers and Meyer 2011; Rodrigues-Oliveira et al. 2017). Lastly, GH13 domains have many recorded activities, but predominantly hydrolyse α-linkages in sugars (Stam et al. 2006). Considering the variety in this class and its prominence across taxonomical range, GH13 domains are expected to be abundantly represented in the dataset.

CE10 domains were the most abundant class of CE identified in the study. However it was shown that the majority of CE10 substrates are not actually carbohydrates, and CE10s are therefore no longer considered CAZymes (Lombard et al. 2014). CE4 and CE1 were the next most highly represented CE domains in the dataset. Both CE4 and CE1 domains have displayed acetyl-xylan esterase (AXE) activity (Caufrier et al. 2003). Additionally, CE4 domains hydrolyse chitin and peptidoglycan (Blair et al. 2005), while CE1 domains show feruloyl esterase activity (Prates et al. 2001; Tarbouriech et al. 2005) and can hydrolyse bonds often found in lignin (Boerjan et al. 2003; Sarkanen and Ludwig 1971). While the presence of CE4 domains may be due to the need for cell wall biosynthesis and maintenance in extremely thermophilic archaea and bacteria, the ability of CE4 and CE1 domains to degrade xylan and lignin suggest that these organisms have capacity to metabolise lignocellulosic biomass. In addition to this, the most common PL domains identified in this study, PL11 and PL9 (Figure 2.1), have both been shown to process pectins (Lombard et al. 2014), an important component of plant primary cell walls (Mohnen 2008).

2.5.2 Lignocellulose degrading capacity

A number of CBM and GH domains identified in this study are potentially able to target and hydrolyse specific linkages in cellulose and/or xylan (Table 1). Cellulose and xylan are abundant in lignocellulosic material and are economically valuable commodities (Himmel et al. 2007). Using CAZymes from extremely thermophilic organisms could facilitate processing of lignocellulosic biomass. However, in order to break down a biopolymer (to either use the monomers in downstream processes or make extracting of other polymers easier and cheaper), an enzyme would need to perform two actions, specifically: i) targeting to a specific location or substrate, and ii) hydrolysis of the substrate.

Targeting of an enzyme to a specific location or substrate is not always required for enzyme function but generally increases efficiency of catalysis (Duan et al. 2017; Sainz-Polo et al. 2015). For CAZymes, targeting is generally mediated by CBMs (Lombard et al. 2014). By recognising and binding to certain residues, CBMs bring enzymes into close contact with substrates. A number of CBM domains that recognise cellulose, xylan or both were identified in this study (Table 1). These included CBM2, CBM3, CBM4, CBM6, CBM9, CBM13, CBM16, CBM22, CBM28, CBM35, CBM36, CBM37, CBM44, CBM54 and CBM60 domains.

CBM3 (Poole et al. 1992; Shimon et al. 2000; Tormo et al. 1996), CBM16 (Bae et al. 2008), CBM28 (Boraston et al. 2002; Boraston et al. 2003), and CBM44 (Najmudin et al. 2006) domains have been shown to bind to cellulose *in vitro*, with CBM28 shown to bind specifically to amorphous cellulose (Blake et al. 2006; Boraston et al. 2003). Conversely, CBM13 (Boraston et al. 2000; Schärpf et al. 2002), CBM22 (Charnock et al. 2000), CBM35 (Cantarel et al. 2009; Kellett et al. 1990; Lombard et al. 2014), CBM36 (Jamal-Talabani et al. 2004), CBM54 (Dvortsov et al. 2010; Dvortsov et al. 2009) and CBM60 (Montanier et al. 2010) domains recognise and bind xylan. The remaining CBM domains, CBM2 (Jing et al. 2009; Xu et al. 1995), CBM4 (Sunna et al. 2001; Zverlov et al. 2001), CBM6 (Abbott et al. 2009; Fernandes et al. 1999; van Bueren et al. 2005), CBM9 (Boraston et al. 2001; Goldstein et al. 1993; Winterhalter et al. 1995) and CBM37 (Xu et al. 2004) are able to bind both cellulose and xylan. While many domains may recognise the same biopolymer, each domain may target a different location on the biopolymer, such as amorphous and crystalline regions of cellulose, or residues adjacent to certain chemical modifications of the backbone of glucuronoxylan. Together, these CBM domains can target a large proportion of the structures of cellulose and glucuronoxylan and could allow for targeting of enzymes to specific locations, facilitating lignocellulose deconstruction through rationally designed enzymes containing CBM domains.

However, while CBM domains can disrupt the structure of biopolymers making them more amenable to degradation (Reese et al. 1950; Shoseyov et al. 2006), they are normally not sufficient for efficient hydrolysis. GH domains can facilitate hydrolysis and many in the dataset are predicted to target cellulose and xylan. For example, GH1 (Cairns and Esen 2010), GH9 (Gilkes et al. 1991; Henrissat et al. 1989), GH12 (Vlasenko et al. 2010), GH26 (Araki et al. 2000; Cartmell et al. 2008; Taylor et al. 2005), GH44 (Najmudin et al. 2010; Warner et al. 2010), GH48 (Barr et al. 1996) and GH74 (Bauer et al. 2005;

Chhabra and Kelly 2002; Desmet et al. 2007; Yaoi and Mitsuishi 2002; York et al. 1993) domains were shown experimentally to hydrolyse the β-1,4-glycosidic bonds present in cellulose. GH10, GH11 (Gebler et al. 1992; Henrissat 1991), GH39 (Morrison et al. 2016), GH43 (Cantarel et al. 2009; Lombard et al. 2014; Shallom et al. 2005), GH67 (Bronnenmeier et al. 1995; Ruile et al. 1997) and GH116 (Cobucci-Ponzano et al. 2010; Ferrara et al. 2014) were shown to hydrolyse xylan. Finally, GH3 (Harvey et al. 2000; Macdonald et al. 2014), GH5 (Henrissat 1991; Henrissat and Bairoch 1993; Henrissat and Bairoch 1996; Henrissat et al. 1989), GH8 (Gilkes et al. 1991; Henrissat et al. 1989; Lombard et al. 2014), GH30 (Cantarel et al. 2009; Henrissat 1991; Henrissat et al. 1989; Lombard et al. 2014) and GH51 (Eckert and Schneider 2003) domains have shown activity on both cellulose and xylan.

Additionally many domains identified in this study have acetyl-xylan esterase activity, such as the CE1, CE3, CE4, CE6, CE7 and CE12 domains (Lombard et al. 2014). Together, the thermostable GH, CBM and CE domains identified in this study could facilitate rational design of thermostable enzymes for the hydrolysis of lignocellulosic biomass (Botha et al. 2017). Breakdown of lignocellulosic biomass currently requires some combination of extreme temperature, pressure and/or pH (Alvira et al. 2010). By supplementing physical and chemical pre-treatment processes with thermostable biopolymer-degrading enzymes, the energy investment and associated economic costs may be reduced (Blumer-Schuette et al. 2014). Additionally, *in planta* expression of enzymes directly in lignocellulosic biomass is also a promising strategy for ameliorating recalcitrance to digestion (Kim et al. 2016; Mir et al. 2014; Mir et al. 2017). The CAZyme domains identified in this study could be used to rationally design thermostable enzymes targeted to specific regions or substrates, and would benefit the abovementioned strategies by allowing efficient and precise modification or degradation of biopolymers.

2.5.3 Future discovery

By subjecting the selection of extremely thermophilic organism proteomes to HMMER analysis, we identified putative domains representative of a large portion of known CAZyme domain families

(Figure 2.5). The diversity of domains present in bacteria and archaea represents approximately 30% to 80% of CAZyme domain families in each CAZyme class. The majority of previously described CE domains are present, but there are relatively few CE domain families compared to other CAZyme classes, and CE domains have high overlap in described functions (Lombard et al. 2014). In most other classes, less than half of published CAZyme domain families are represented (Figure 2.5). Since domains with capacity for lignocellulose degradation have been identified in the dataset (Table 2.1), and some lignocellulose targeting domains were only identified in the proteome of single organisms in this study (Table 2.3), sequencing and analysing the genomes of more extremely thermophilic organisms should allow for the identification of additional lignocellulose degrading CAZyme domains. A rarefaction curve of CAZyme domains from each organism proteome sampled shows that it has not yet reached a plateau (Figure 2.6), indicating that a greater variety of thermostable CAZyme domains could be identified. This is also supported by the identification of domain-, phylum- (Table 2.3), and species-specific (Table 2.3, Supplementary Data 2.6) CAZyme domains in this study. Additionally, while a defined number of CAZyme domains exist (Lombard et al. 2014), new families can be identified over time, and known families are sometimes divided into subfamilies, based on new sequence and structural data (Lombard et al. 2014; Terrapon et al. 2017). By sampling from more extremely thermophilic organisms, these data could also be captured, building an increasing repertoire of thermostable CAZyme domains.

2.6 Conclusion

The CAZyme domains from the extremely thermophilic organism proteomes identified in this study provide a pool of thermostable protein domains. These domains are abundant and diverse, covering a range of different activities and functions, some of which could be developed for industrial application. The ability to degrade lignocellulose is industrially and economically important, and the capacity to do so exists within the identified CAZyme domains. By using this data, a strategy may be developed to design industrially compatible enzymes and processes with the purpose of holistic breakdown of specific lignocellulosic biopolymers such as cellulose or xylan. Additionally, while a significant proportion of CAZyme family domains were represented in the dataset, more families, subfamilies and family variants are expected to be identified if more proteomes of thermophilic organisms are analysed. This work provides the basis for lignocellulose deconstruction and enzyme-engineering strategies that may lead to increased efficiency and lowered cost of bioconstruction, paving the way for a petrochemical free bioeconomy.

2.7 Acknowledgements

The author would to thank Ms Desre Pinard for her help with the R programming language as well as bioinformatics scripts used in the study.

2.8 References

Abbott DW et al. (2009) Analysis of the structural and functional diversity of plant cell wall specific family 6 carbohydrate binding modules. Biochemistry 48:10395-10404

Albers S-V, Meyer BH (2011) The archaeal cell envelope. Nature Reviews Microbiology 9:414

Alvira P, Tomás-Pejó E, Ballesteros M, Negro M (2010) Pre-treatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: a review. Bioresource Technology 101:4851-4861

André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M (2014) CAZyme discovery and design for sweet dreams. Current Opinion in Chemical Biology 19:17-24

Araki T, Hashikawa S, Morishita T (2000) Cloning, sequencing, and expression in *Escherichia coli* of the new gene encoding β-1,3-xylanase from a marine bacterium, *Vibrio sp*. strain XY-214. Applied and Environmental Microbiology 66:1741-1743

Artzi L, Bayer EA, Moraïs S (2017) Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nature Reviews Microbiology 15:83-95

Bae B, Ohene-Adjei S, Kocherginskaya S, Mackie RI, Spies MA, Cann IK, Nair SK (2008) Molecular basis for the selectivity and specificity of ligand recognition by the family 16 carbohydrate-binding modules from *Thermoanaerobacterium polysaccharolyticum* ManA. Journal of Biological Chemistry 283:12415-12425

Barr BK, Hsieh Y-L, Ganem B, Wilson DB (1996) Identification of two functionally different classes of exocellulases. Biochemistry 35:586-592

Bauer S, Vasu P, Mort AJ, Somerville CR (2005) Cloning, expression, and characterization of an oligoxyloglucan reducing end-specific xyloglucanobiohydrolase from *Aspergillus nidulans*. Carbohydrate Research 340:2590-2597

Bayer EA, Belaich J-P, Shoham Y, Lamed R (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. Annual Review of Microbiology 58:521-554

Berg IA et al. (2010) Autotrophic carbon fixation in archaea. Nature Reviews Microbiology 8:447-460

Blair DE, Schüttelkopf AW, MacRae JI, van Aalten DM (2005) Structure and metal-dependent mechanism of peptidoglycan deacetylase, a streptococcal virulence factor. Proceedings of the National Academy of Sciences 102:15429-15434

Blake AW, McCartney L, Flint JE, Bolam DN, Boraston AB, Gilbert HJ, Knox JP (2006) Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes. Journal of Biological Chemistry 281:29321-29329

Blumer-Schuette SE et al. (2014) Thermophilic lignocellulose deconstruction. FEMS microbiology reviews 38:393-448

Blumer-Schuette SE, Kataeva I, Westpheling J, Adams MW, Kelly RM (2008) Extremely thermophilic microorganisms for biomass conversion: status and prospects. Current Opinion in Biotechnology 19:210-217

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annual Review of Plant Biology 54:519-546

Boraston A, Ghaffari M, Warren R, Kilburn D (2002) Identification and glucan-binding properties of a new carbohydrate-binding module family. Biochemical Journal 361:35-40

Boraston A, Tomme P, Amandoron E, Kilburn D (2000) A novel mechanism of xylan binding by a lectin-like module from *Streptomyces lividans* xylanase 10A. Biochemical Journal 350:933-941

Boraston AB et al. (2001) Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A. Biochemistry 40:6240-6247

Boraston AB, Kwan E, Chiu P, Warren RAJ, Kilburn DG (2003) Recognition and hydrolysis of noncrystalline cellulose. Journal of Biological Chemistry 278:6120-6127

Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

Bronnenmeier K, Meissner H, Stocker S, Staudenbauer WL (1995) α-d-glucuronidases from the xylanolytic thermophiles *Clostridium stercorarium* and *Thermoanaerobacterium saccharolyticum*. Microbiology 141:2033-2040

Cabeen MT, Jacobs-Wagner C (2005) Bacterial cell shape. Nature Reviews Microbiology 3:601

Cairns JRK, Esen A (2010) β-Glucosidases. Cellular and Molecular Life Sciences 67:3389-3405

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Research 37:D233-D238

Cartmell A, Topakas E, Ducros VM, Suits MD, Davies GJ, Gilbert HJ (2008) The *Cellvibrio japonicus* mannanase CjMan26C displays a unique exo-mode of action that is conferred by subtle changes to the distal region of the active site. Journal of Biological Chemistry 283:34403-34413

Carvalho CC, Phan NN, Chen Y, Reilly PJ (2015) Carbohydrate-binding module tribes. Biopolymers 103:203-214

Caufrier F, Martinou A, Dupont C, Bouriotis V (2003) Carbohydrate esterase family 4 enzymes: substrate specificity. Carbohydrate Research 338:687-692

Charnock SJ, Bolam DN, Turkenburg JP, Gilbert HJ, Ferreira LM, Davies GJ, Fontes CM (2000) The X6 “thermostabilizing” domains of xylanases are carbohydrate-binding modules: structure and biochemistry of *the Clostridium thermocellum* X6b domain. Biochemistry 39:5013-5021

Chhabra SR, Kelly RM (2002) Biochemical characterization of *Thermotoga maritima* endoglucanase Cel74 with and without a carbohydrate binding module (CBM). FEBS Letters 531:375-380

Cobucci-Ponzano B et al. (2010) A new archaeal β-glycosidase from *Sulfolobus solfataricus* seeding a novel retaining β-glycan-specific glycoside hydrolase family along with the human non-lysosomal glucosylceramidase GBA2. Journal of Biological Chemistry 285:20691-20703

Cosgrove DJ, Jarvis MC (2012) Comparative structure and biomechanics of plant primary and secondary cell walls. Frontiers in Plant Science 3:204

Dawes EA, Senior PJ (1973) The role and regulation of energy reserve polymers in micro-organisms Advances in Microbial Physiology 10:135-266

Desmet T, Cantaert T, Gualfetti P, Nerinckx W, Gross L, Mitchinson C, Piens K (2007) An investigation of the substrate specificity of the xyloglucanase Cel74A from *Hypocrea jecorina.* FEBS Journal 274:356-363

Duan C-J, Huang M-Y, Pang H, Zhao J, Wu C-X, Feng J-X (2017) Characterization of a novel theme C glycoside hydrolase family 9 cellulase and its CBM-chimeric enzymes. Applied Microbiology and Biotechnology 101:5723-5737

Dvortsov I, Lunina N, Zverlov V, Velikodvorskaya G (2010) Substrate-binding properties of the family 54 module of *Clostridium thermocellum* Lic16A laminarinase. Molecular Biology 44:591-595

Dvortsov IA, Lunina NA, Chekanovskaya LA, Schwarz WH, Zverlov VV, Velikodvorskaya GA (2009) Carbohydrate-binding properties of a separately folding protein module from β-1,3-glucanase Lic16A of *Clostridium thermocellum.* Microbiology 155:2442-2449

Eckert K, Schneider E (2003) A thermoacidophilic endoglucanase (CelB) from *Alicyclobacillus acidocaldarius* displays high sequence similarity to arabinofuranosidases belonging to family 51 of glycoside hydrolases The FEBS Journal. 270:3593-3602

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755-763

Fernandes A, Fontes C, Gilbert H, Hazlewood G, Fernandes T, Ferreira L (1999) Homologous xylanases from *Clostridium thermocellum*: evidence for bi-functional activity, synergism between xylanase catalytic modules and the presence of xylan-binding domains in enzyme complexes. Biochemal Journal 342:105-110

Ferrara MC, Cobucci-Ponzano B, Carpentieri A, Henrissat B, Rossi M, Amoresano A, Moracci M (2014) The identification and molecular characterization of the first archaeal bifunctional exo-β-glucosidase/N-acetyl-β-glucosaminidase demonstrate that family GH116 is made of three functionally distinct subfamilies. Biochimica et Biophysica Acta - General Subjects 1840:367-377

Gebler J et al. (1992) Stereoselective hydrolysis catalyzed by related beta-1,4-glucanases and beta-1,4-xylanases. Journal of Biological Chemistry 267:12559-12561

Gerday C, Glansdorff N (2007) Physiology and biochemistry of extremophiles. ASM Press

Gibbs MD, Reeves RA, Farrington GK, Anderson P, Williams DP, Bergquist PL (2000) Multidomain and multifunctional glycosyl hydrolases from the extreme thermophile *Caldicellulosiruptor* isolate Tok7B. Current Microbiology 40:333-340

Gilkes N, Henrissat B, Kilburn D, Miller R, Warren R (1991) Domains in microbial beta-1,4-glycanases: sequence conservation, function, and enzyme families. Microbiological Reviews 55:303-315

Goldstein MA, Takagi M, Hashida S, Shoseyov O, Doi R, Segel I (1993) Characterization of the cellulose-binding domain of the *Clostridium cellulovorans* cellulose-binding protein A. Journal of Bacteriology 175:5762-5768

Harvey AJ, Hrmova M, De Gori R, Varghese JN, Fincher GB (2000) Comparative modeling of the three-dimensional structures of family 3 glycoside hydrolases. Proteins: Structure, Function, and Bioinformatics 41:257-269

Hendriks A, Zeeman G (2009) Pre-treatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology 100:10-18

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

Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochemical Journal 293:781-788

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

Henrissat B, Claeyssens M, Tomme P, Lemesle L, Mornon J-P (1989) Cellulase families revealed by hydrophobic cluster analysis. Gene 81:83-95

Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315:804-807

Horcajada C, Guinovart JJ, Fita I, Ferrer JC (2006) Crystal structure of an archaeal glycogen synthase: Insights into oligomerization and substrate binding of eukaryotic glycogen synthases. Journal of Biological Chemistry 281:2923-2931

Jahn U et al. (2007) *Ignicoccus hospitalis sp. nov.*, the host of '*Nanoarchaeum equitans*' International Journal of Systematic and Evolutionary Microbiology 57:803-808

Jamal-Talabani S, Boraston AB, Turkenburg JP, Tarbouriech N, Ducros VM-A, Davies GJ (2004) *Ab initio* structure determination and functional characterization of CBM36: a new family of calcium-dependent carbohydrate binding modules. Structure 12:1177-1187

Jing H, Cockburn D, Zhang Q, Clarke AJ (2009) Production and purification of the isolated family 2a carbohydrate-binding module from *Cellulomonas fimi.* Protein Expression and Purification 64:63-68

Kellett LE, Poole DM, Ferreira L, Durrant AJ, Hazlewood GP, Gilbert HJ (1990) Xylanase B and an arabinofuranosidase from *Pseudomonas fluorescens subsp. cellulosa* contain identical cellulose-binding domains and are encoded by adjacent genes. Biochemical Journal 272:369-376

Khan MIM, Sajjad M, Sadaf S, Zafar R, Niazi UH, Akhtar MW (2013) The nature of the carbohydrate binding module determines the catalytic efficiency of xylanase Z of *Clostridium thermocellum*. Journal of Biotechnology 168:403-408

Kim JY, Nong G, Rice JD, Gallo M, Preston JF, Altpeter F (2016) In planta production and characterization of a hyperthermostable GH10 xylanase in transgenic sugarcane. Plant Molecular Biology 4-5:465-478

Lairson L, Henrissat B, Davies G, Withers S (2008) Glycosyltransferases: structures, functions, and mechanisms. Biochemistry 77:521

Lee Y-E, Lowe S, Henrissat B, Zeikus JG (1993) Characterization of the active site and thermostability regions of endoxylanase from *Thermoanaerobacterium saccharolyticum* B6A-RI. Journal of Bacteriology 175:5890-5898

Leuschner C, Antranikian G (1995) Heat-stable enzymes from extremely thermophilic and hyperthermophilic microorganisms. World Journal of Microbiology and Biotechnology 11:95-114

Liu QP et al. (2007) Bacterial glycosidases for the production of universal red blood cells. Nature Biotechnology 25:454-464

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495

Macdonald SS, Blaukopf M, Withers SG (2014) N-Acetyl glucosaminidases from CAZy family GH3 are really glycoside phosphorylases, thereby explaining their use of histidine as an acid/base catalyst in place of glutamic acid. Journal of Biological Chemistry jbc-M114

Mir BA, Mewalal R, Mizrachi E, Myburg AA, Cowan DA (2014) Recombinant hyperthermophilic enzyme expression in plants: a novel approach for lignocellulose digestion. Trends in Biotechnology 32:281-289

Mir BA, Myburg AA, Mizrachi E, Cowan DA (2017) In planta expression of hyperthermophilic enzymes as a strategy for accelerated lignocellulosic digestion. Scientific Reports 7:11462

Mohnen D (2008) Pectin structure and biosynthesis. Current Opinion in Plant Biology 11:266-277

Montanier C et al. (2010) Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. Journal of Biological Chemistry 285:31742-31754

Morrison JM, Elshahed MS, Youssef N (2016) A multifunctional GH39 glycoside hydrolase from the anaerobic gut fungus *Orpinomyces sp.* strain C1A. PeerJ 4:e2289

Naik SN, Goud VV, Rout PK, Dalai AK (2010) Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews 14:578-597

Najmudin S et al. (2006) Xyloglucan is recognized by carbohydrate-binding modules that interact with β-glucan chains. Journal of Biological Chemistry 281:8815-8828

Najmudin S, Pinheiro BA, Prates JA, Gilbert HJ, Romão MJ, Fontes CM (2010) Putting an N-terminal end to the *Clostridium thermocellum* xylanase Xyn10B story: Crystal structure of the CBM22-1–GH10 modules complexed with xylohexaose. Journal of Structural Biology 172:353-362

Onaga S, Taira T (2008) A new type of plant chitinase containing LysM domains from a fern (*Pteris ryukyuensis*): roles of LysM domains in chitin binding and antifungal activity. Glycobiology 18:414-423

Orell A, Fröls S, Albers S-V (2013) Archaeal biofilms: the great unexplored. Annual Review of Microbiology 67

Orell A, Schopf S, Randau L, Vera M (2017) Biofilm lifestyle of thermophile and acidophile archaea. Biocommunication of Archaea. Springer, pp 133-146

Peberdy J (1990) Fungal cell walls—a review. Biochemistry of cell walls and membranes in fungi. Springer, pp 5-30

Podar M et al. (2008) A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans.* Genome biology 9:R158

Poole DM, Morag E, Lamed R, Bayer EA, Hazlewood GP, Gilbert HJ (1992) Identification of the cellulose-binding domain of the cellulosome subunit S1 from *Clostridium thermocellum* YS. FEMS Microbiology Letters 99:181-186

Prates JA, Tarbouriech N, Charnock SJ, Fontes CM, Ferreira LsM, Davies GJ (2001) The structure of the feruloyl esterase module of xylanase 10B from *Clostridium thermocellum* provides insights into substrate recognition. Structure 9:1183-1190

Reese ET, Siu RG, Levinson HS (1950) The biological degradation of soluble cellulose derivatives and its relationship to the mechanism of cellulose hydrolysis. Journal of Bacteriology 59:485

Rodrigues-Oliveira T, Belmok A, Vasconcellos D, Schuster B, Kyaw CM (2017) Archaeal S-layers: overview and current state of the art. Frontiers in Microbiology 8:2597

Ruile P, Winterhalter C, Liebl W (1997) Isolation and analysis of a gene encoding α-glucuronidase, an enzyme with a novel primary structure involved in the breakdown of xylan. Molecular Microbiology 23:267-279

Sainz-Polo MA, González B, Menéndez M, Pastor FJ, Sanz-Aparicio J (2015) Exploring multimodularity in plant cell wall deconstruction: structural and functional analysis of Xyn10C containing the CBM22-1-CBM22-2 tandem. Journal of Biological Chemistry jbc-M115

Sarkanen KV, Ludwig CH (1971) Lignins: occurrence, formation, structure and reactions Lignins: occurrence, formation, structure and reactions. Wiley-Interscience

Schärpf M, Connelly GP, Lee GM, Boraston AB, Warren RAJ, McIntosh LP (2002) Site-specific characterization of the association of xylooligosaccharides with the CBM13 lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A by NMR spectroscopy. Biochemistry 41:4255-4263

Schleifer KH, Kandler O (1972) Peptidoglycan types of bacterial cell walls and their taxonomic implications. Bacteriological Reviews 36:407

Shallom D et al. (2005) Biochemical characterization and identification of the catalytic residues of a family 43 β-D-xylosidase from *Geobacillus stearothermophilus* T-6. Biochemistry 44:387-397

Shimon LJ, Belaich A, Belaich J-P, Bayer E, Lamed R, Shoham Y, Frolow F (2000) Structure of a family IIIa scaffoldin CBD from the cellulosome of *Clostridium cellulolyticum* at 2.2 Å resolution. Acta Crystallographica Section D: Biological Crystallography 56:1560-1568

Shoseyov O, Shani Z, Levy I (2006) Carbohydrate binding modules: biochemical properties and novel applications. Microbiology and Molecular Biology Reviews 70:283-295

Stam MR, Danchin EG, Rancurel C, Coutinho PM, Henrissat B (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Engineering, Design and Selection 19:555-562

Steen A et al. (2003) Cell wall attachment of a widely distributed peptidoglycan binding domain is hindered by cell wall constituents. Journal of Biological Chemistry 278:23874-23881

Sunna A, Gibbs MD, Bergquist PL (2001) Identification of novel β-mannan-and β-glucan-binding modules: evidence for a superfamily of carbohydrate-binding modules. Biochemical Journal 356:791-798

Tarbouriech N, Prates JAM, Fontes CMGA, Davies GJ (2005) Molecular determinants of substrate specificity in the feruloyl esterase module of xylanase 10B from *Clostridium thermocellum.* Acta Crystallographica Section D 61:194-197

Taylor EJ et al. (2005) How family 26 glycoside hydrolases orchestrate catalysis on different polysaccharides: Structure and activity of *Clostridium thermocellum* lichenase, CtLic26A. Journal of Biological Chemistry 280:32761-32767

Team RC (2013) R: A language and environment for statistical computing. 201

Terrapon N, Lombard V, Drula E, Coutinho PM, Henrissat B (2017) The CAZy database/the carbohydrate-active enzyme (CAZy) database: principles and usage guidelines. A Practical Guide to Using Glycomics Databases. Springer, pp 117-131

Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, Steitz TA (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. The EMBO Journal 15:5739

van Bueren AL, Morland C, Gilbert HJ, Boraston AB (2005) Family 6 carbohydrate binding modules recognize the non-reducing end of β-1,3-linked glucans by presenting a unique ligand binding surface. Journal of Biological Chemistry 280:530-537

Vlasenko E, Schülein M, Cherry J, Xu F (2010) Substrate specificity of family 5, 6, 7, 9, 12, and 45 endoglucanases. Bioresource Technology 101:2405-2411

Warner CD et al. (2010) Tertiary structure and characterization of a glycoside hydrolase family 44 endoglucanase from *Clostridium acetobutylicum.* Applied and Environmental Microbiology 76:338-346

Winterhalter C, Heinrich P, Candussio A, Wich G, Liebl W (1995) Identification of a novel cellulose-binding domain the multidomain 120 kDa xylanase XynA of the hyperthermophilic bacterium *Thermotoga maritima.* Molecular Microbiology 15:431-444

Xu G-Y et al. (1995) Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. Biochemistry 34:6993-7009

Xu Q, Morrison M, Nelson KE, Bayer EA, Atamna N, Lamed R (2004) A novel family of carbohydrate-binding modules identified with *Ruminococcus albus* proteins. FEBS Letters 566:11-16

Yaoi K, Mitsuishi Y (2002) Purification, characterization, cloning, and expression of a novel xyloglucan-specific glycosidase, oligoxyloglucan reducing end-specific cellobiohydrolase. Journal of Biological Chemistry 277:48276-48281

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research 40:W445-W451

York WS, Harvey LK, Guillen R, Alberhseim P, Darvill AG (1993) Structural analysis of tamarind seed xyloglucan oligosaccharides using β-galactosidase digestion and spectroscopic methods. Carbohydrate Research 248:285-301

Zverlov VV, Volkov IY, Velikodvorskaya GA, Schwarz WH (2001) The binding pattern of two carbohydrate-binding modules of laminarinase Lam16A from *Thermotoga neapolitana*: differences in β-glucan binding within family CBM4 Microbiology 147:621-629

CHAPTER 3:

In planta expression of a chimeric xylanase in *Arabidopsis* shows targeting to the secondary cell wall but increases recalcitrance to enzymic degradation

Jonathan Botha[1,2,4], Eshchar Mizrachi[2,3], Alexander A. Myburg[2,3], Don A. Cowan[1,2,4]

[1]Centre for Microbial Ecology and Genomics, Department Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[2]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[3]Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa
[4]Genomics Research Institute (GRI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

In this chapter, I designed the chimeric enzyme Xyl22L. I amplified the synthesised coding sequence, and subcloned it into the appropriate vectors for plant transformation. I generated the transgenic plant lines and performed all experimental work reported in this chapter. I also performed all data analysis and prepared this manuscript. Prof. D.A. Cowan, Prof. A.A. Myburg and Dr E. Mizrachi provided advice, direction and supervision during the planning of this chapter, as well as guidance during the interpretation of the results. They also performed critical revision of the manuscript.

3.1 Abstract

Lignocellulosic biomass is an important second generation feedstock for the production of biomaterials, biochemicals and bioenergy. However, lignocellulosic feedstocks are notoriously recalcitrant to enzymic digestion, requiring expensive industrial pre-treatments in order to process them. In this study, a chimeric thermostable enzyme was designed, synthesized and heterologously expressed in *Arabidopsis thaliana*, in order to reduce biomass recalcitrance and increase efficiency of targeting of the enzyme to the secondary cell wall (SCW). The enzyme, Xyl22L, consisted of a thermostable glycoside hydrolase (GH11) catalytic domain derived from an extremely thermophilic metagenomic library, as well as C-terminal xylan-targeting carbohydrate binding module (CBM22) repeats obtained from *Eucalyptus grandis*. The enzyme was synthesised and characterised, and transgenic *A. thaliana* plants expressing Xyl22L under a constitutive promoter were produced. The growth and development, as well as recalcitrance to enzymic digestion of the transgenic plants was assessed. Accumulation and localisation of Xyl22L in the plants was also examined. The GH11 catalytic domain of Xyl22L proved to be inactive, but transgenic lines expressing Xyl22L showed increased biomass. Xyl22L accumulated at low levels in plant biomass, and localised to the SCWs of interfascicular fibre cells. Before heat treatment, dried biomass containing Xyl22L showed increased recalcitrance to enzymic digestion, suggesting that Xyl22L adhered to the SCW and prevented access of other digestive enzymes. This work showed that enzymes may be targeted to specific locations in the SCW by appending plant-derived CBMs to them, and provides groundwork for the synthesis of custom enzymes that target specific biopolymers or other features of SCWs in lignocellulosic biomass. Together, these findings could help to reduce recalcitrance of lignocellulosic biomass in a controlled and efficient manner.

3.2 Introduction

It is becoming increasingly important to find sustainable and environmentally friendly alternatives to current petrochemically derived products. Bioproducts such as biofuels, bioplastics, and bioadhesives, among others (Gallezot 2012), may offer a solution to this problem as they are derived from biological sources and can be biodegradable. However, most bioproducts require large amounts of simple polysaccharides and biopolymers (Perlack et al. 2005), that can be difficult and expensive to extract. Each feedstock used for bioproduct synthesis has associated advantages and challenges (McKendry 2002), which need to be overcome in order for them to be economically viable on a large scale.

The first generation of feedstocks were mainly food crops such as maize and sugarcane, as these plants had large amounts of readily accessible polysaccharides (mostly starch based), which could easily be converted to fermentable sugars for biofuel. However, these feedstocks are not sustainable sources for biomaterial synthesis, as they compete with food production and have a lower net energy ratio [energy returned vs energy invested; (Naik et al. 2010)]. In order to overcome this problem, other sources of plant biomass such as agricultural waste, perennial grasses and tree species were identified as the second generation of feedstocks for bioproduct synthesis. These included hardwood tree genera such as *Eucalyptus* and *Populus*, and other plants such as *Agave*, willow, alfalfa, *Miscanthus*, switchgrass, bitter cassava, wild sugarcane, hemp and water hyacinth, among others (Phitsuwan et al. 2013). Second generation crops do not compete with food production, can grow in a wide range of conditions and environments, can be harvested year-round, and produce large amounts of lignocellulosic biomass for conversion (Hendriks and Zeeman 2009; Himmel et al. 2007).

However, second generation feedstocks are not without their challenges. Plant secondary cell walls (SCWs), comprising the bulk of lignocellulosic biomass, are highly recalcitrant to enzymic breakdown, due mainly to physical barriers formed by the structural complexity and cross-linking of lignin, cellulose and hemicelluloses in the cell wall (Busse-Wicher et al. 2016; Cosgrove 2005; Cosgrove 2014).

This can be overcome in part by submitting plant biomass to a variety of industrial pre-treatments such as milling or grinding, steam explosion or ammonia fibre explosion (Alvira et al. 2010). Industrial pre-treatments are harsh, requiring a significant investment of energy, and also produce degradation products which inhibit downstream processes (Jönsson and Martín 2016).

Carbohydrate Active Enzymes (CAZymes) are proteins which are active on oligosaccharides, polysaccharides and glycoconjugates, and may be able to supplement current industrial pre-treatments. CAZymes are organised into a hierarchy of protein domain families that perform broad functions (Cantarel et al. 2009; Lombard et al. 2014), such as GlycosylTransferases (GTs), Glycoside Hydrolases (GHs), Carbohydrate Esterases (CEs), Polysaccharide Lysases (PLs), Carbohydrate-Binding Modules (CBMs) and Auxiliary Activity families (AAs). These enzymes could help to digest complex biopolymers in lignocellulosic biomass, thereby reducing the cost, energy investment and degradation products associated with industrial pre-treatments.

Industrial pre-treatments involve extreme conditions such as high temperatures and pressures or extreme pH, under which most enzymes are not functional. Extremely thermophilic organisms grow optimally at temperatures exceeding 70°C (Gerday and Glansdorff 2007; Leuschner and Antranikian 1995), and may offer a solution to this problem. Due to the ecological niches which they inhabit (Rothschild and Mancinelli 2001), extremely thermophilic organisms are an excellent source of thermostable enzymes for use in industrial applications (Blumer-Schuette et al. 2014).

Heterologous *in planta* expression of thermostable enzymes may allow for normal growth of the plant at mesophilic temperatures, as well as accumulation of the enzyme in plant tissues. This could reduce the need for processing and enzyme loading in the harvested biomass, thereby producing self-processing plants for biomaterials extraction [i.e. autohydrolysis; (Mir et al. 2014; Mir et al 2017; Montalvo-Rodriguez et al. 2000; Ziegler et al. 2000)]. Additionally, this process may be improved by tailoring enzymes (André et al. 2014; Elleuche 2015) to specific purposes using individual CAZyme domains (Botha et al. 2017). By pairing a thermostable xylan-degrading domain with a plant derived binding domain, it may be possible to produce an enzyme which specifically localises to the plant SCW, facilitating activity of the xylanase domain and increasing efficiency of lignocellulosic biomass degradation. In this study, we design, synthesise and characterise a chimeric xylanase (designated Xyl22L), consisting of a hyperthermophilic xylan degrading GH11 domain from an extremely thermophilic metagenomic library, and *Eucalyptus grandis* xylan binding CBM22 repeats to target the synthetic enzyme to the SCW. We heterologously express Xyl22L in *Arabidopsis*, investigate its accumulation and localisation in the plant, and assess its effect on the growth and development of transgenic *A. thaliana* plants as well as the recalcitrance of the plant biomass to enzymic saccharification.

3.3 Materials and Methods

3.3.1 Synthesis and cloning of Xyl22L

A literature survey was performed for xylanase candidates that were active at high temperature and extreme pH, in order to identify a candidate that would operate under standard lignocellulose industrial pre-treatment conditions (Supplementary Data 3.1). The xylanase *JX125044* (*Mxyl*, GENBANK accession: AFP81696) was identified from a compost-soil metagenome (Verma et al. 2013), and was selected based on its thermostability and tolerance for high pH. The coding sequence of *JX125044* was modified with a tobacco pathogenesis-related protein 1a (Pr1a) signal peptide and a plant-specific kozak consensus sequence (GCCACCATGG) at the 5' end (Mir et al. 2017). Additionally, the endogenous CBM60 domain was replaced with three tandem copies of *Eucalyptus grandis* CBM22, obtained from the *E. grandis* xylanase encoded by *Eucgr.F00108*, and was designated Xyl22L. All three CBM22 repeats in *Eucgr.F00108* were inserted in an effort to preserve the wild-type structures and functions of the CBM22 domains. DNA sequences were sent to GenScript Corporation (Piscataway, NJ, USA) for *de novo* synthesis and were amplified using Phusion® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA) with an initial denaturation step of 30 s at 98°C, then 30 cycles of 10 s at 98°C, 30 s at 62°C and 30 s at 72°C, followed by a final extension step of 72°C for 10 min. PCR reaction products were visualised via agarose gel electrophoresis, excised and gel purified using the NucleoSpin Gel and PCR clean-up kit (Machary Nagel, Dűren, Germany). Adenine overhangs were added to the blunt ended products by incubating 20 µl of PCR product with 1 µl dNTPs (2.5 mM), 5 µl Roche buffer (10 x) and 2 µl Roche Taq polymerase (Merck, Modderfontein, South Africa; 5 U/µl). Reactions were made up to a final volume of 50 µl with dH_2O and incubated at 72°C for 10 min in order to facilitate TA cloning into the pCR™8/GW/TOPO® entry vector (ThermoFisher Scientific, Waltham, MA, USA). Chemically competent *E. coli* DH5α cells were transformed via heat shock with the recombinant cassette. The transformation mixtures were grown on plates containing Spectinomycin (100 µg/ml) to select for colonies containing the insert.

Colonies were screened for correct insert orientation by colony PCR. Briefly, colonies were picked using pipette tips and resuspended in 20 µl of dH_2O. 5 µl of suspension was then used as template in PCR reactions. Two sets of PCR reactions were performed on each colony, one containing a vector-specific forward primer and gene-specific reverse primer (Supplementary Table 1), and a second containing a vector-specific forward primer and gene-specific forward primer (Supplementary table 1). PCR reactions were performed using Excel Taq polymerase (Smobio, Hsinchu City, Taiwan) as per the manufacturer's instructions. The reaction mixture was first heated for 6 min at 94°C, followed by 30 cycles of 30 s at 94°C, 30 s at 55°C and 1 min at 72°C, with a final extension step of 72°C for 10 min. Colonies which produced a strong band determined by agarose gel electrophoresis in the first reaction, but not the second, were selected for further analysis. Plasmid DNA was extracted from selected colonies using the GeneJET Plasmid Miniprep kit (ThermoFischer Scientific) and was sequence-verified using Sanger sequencing (Macrogen Inc., Seoul, Korea), with M13 and gene-specific primers (Supplementary table 1). Sequence verified Xyl22L (Supplementary File 3.1) was cloned into the pMDC32 destination vector (Curtis and Grossniklaus 2003) using Gateway™ LR Clonase™ II Enzyme Mix (ThermoFisher Scientific), and sequence was verified using Sanger sequencing (Macrogen Inc.) with M13 and gene-specific primers (Supplementary table 1).

3.3.2 Generation of transgenic *Arabidopsis thaliana* plants

Transgenic *Arabidopsis thaliana* plants were generated by *Agrobacterium tumifaciens* mediated transformation via a floral dipping method (Clough and Bent 1998). Briefly, *A. thaliana* Col-0 plants were grown under long day conditions for four weeks to encourage bolting. At four weeks, the inflorescence stems were clipped back in order to encourage further bolting and flower production. Four days later, the *A. thaliana* plants were dipped in pre-prepared solutions of *A. tumifaciens* LBA4404, containing *Xyl22L* in pMDC32. Dipping involved submerging the rosette and inflorescence stems in the *A. tumifaciens* solution for 5 s with gentle agitation, after which the plants were placed in a container in a horizontal position and covered in plastic for 16-24 hours. The dipping was then repeated one week later. Plants were grown to maturity under long day conditions (16 h light) and seed was collected, dried and sieved to remove excess debris. The T_0 seed was washed with 70% EtOH for 5 min, then in a freshly prepared solution of 10% bleach and 0.1% Triton X-100 for 15 min, and subsequently rinsed 3-4 times with distilled water. Washed and sterilised seed was sown onto germination medium (1mM, KNO_3, 0.8% Bacto agar) containing hygromycin (20 mg/ml) and cefotaxime (100 mg/ml) in order to select for transgenic plants. Seeds were kept in the dark at 4°C for two days before being transferred to long day conditions to allow germination and growth. After two weeks of growth, plantlets (T_1) showing healthy growth and no yellowing were transferred to Jiffy™ pots (Jiffy Products International, Norway), incubated under long day conditions, and allowed to reach maturity and to produce seed (T_2). Leaf samples were taken from T_1 plants in order to extract DNA for PCR analysis to confirm the presence of the constructs in the plants (Supplementary Figure 3.1). The T_2 seed was washed and sown onto selective medium and grown as previously described. T_3 seed was collected from these plants, and a chi-square analysis was used to determine homozygosity and select homozygous plant lines for each transformation event. T_3 seed was washed and planted as previously described. After two weeks of growth on selective medium containing hygromycin (20 mg/ml), the number of resistant and sensitive plants were counted, and used to calculate chi-square values for a typical monohybrid cross (Supplementary Data 3.2). Plant lines showing skewed ratios with significantly more plants resistant to selection were considered homozygotes for the inserted constructs and selected for further analysis. Plants of each homozygous line were transferred to Jiffy™ pots (Jiffy Products International) and allowed to grow to maturity (6 weeks) under short day conditions for further analysis. Three plants were set aside and grown under long day conditions for seed collection (T_3).

3.3.3 RT-qPCR analysis

Total RNA was extracted from four week old plant leaves using the SV Total RNA isolation system (Promega, Madison, WI). RNA was poly-dT prepared by mixing 1 µg of purified RNA and 0.5 µl poly dT primer (PolyTVN; 100 µM), and adjusted to 5 µl with RNAse-free water. The mixture was incubated for 5 min at 70^{o}C, and then incubated on ice for at least 5 minutes. cDNA was synthesised using Improm-II™ reverse transcriptase (Promega) as per the manufacturer's instructions. Synthesised cDNA was screened for contamination via PCR with intron-spanning *Act2* primers (Supplementary Table 3.1, Supplementary Figure 3.2). qPCR was performed on the QuantStudio 12K Flex Real-Time PCR System (ThermoFischer Scientific) using *Xyl22L*-specific primers (Supplementary Table 1) and SYBR® Select Master Mix (ThermoFischer Scientific). Expression data (Supplementary File 3.2) was analysed with the Biogazelle qbase+ software package (ver. 3.1, Zwijnaarde, Belgium, www.qbaseplus.com), using the double delta Ct method of quantification (Hellemans et al. 2007; Schmittgen and Livak 2008). *Arabidopsis Act2* and *Ubq5* (Supplementary table 1) were used as reference genes in the experiment.

3.3.4 Plant phenotyping and microscopy

Pictures were taken of plants at six and eight weeks old for comparison (Supplementary File 3.3). At eight weeks old, the above-ground tissues of transgenic plants were harvested and dried in a laminar flow hood. The weights of the plant tissues were measured before drying and then daily over two weeks. When the weight measurements ceased to change for three consecutive days, the plant material was considered properly dry. The dry weights of the plants were then measured for comparison. In order to investigate plant SCW structure, hand sections of six week old inflorescence stems were made in 70% EtOH, after which they were stained for two minutes with phloroglucinol solution [95% EtOH, 2% phloroglucinol powder (Merck, Modderfontein, South Africa)]. A cover slip was placed over the sections and they were immediately visualised by light microscopy and a 40 X objective lens.

3.3.5 Immunostaining and confocal microscopy

Six week old plant stems were fixed in methanol. Briefly, samples were placed in 100% methanol for 20 min. The methanol was then replaced and the sample was incubated at 60°C for 3 min, after which water was added to reduce the concentration of the methanol. The incubation and addition of water was repeated an additional six times, until a final concentration of 20% methanol was reached. Samples were then stored in dH_2O at 4°C until they were embedded in wax.

For each sample, a 1 cm piece of inflorescence stem was subjected to a butanol dehydration series in order to remove all water from the sample (Supplementary Table 2). The samples were then embedded in paraffin wax using a standard protocol (Zeller 1999) that was modified to use butanol instead of xylene. Paraffin wax blocks were stored at 4°C until they were sectioned. The paraffin wax blocks were cut down to the appropriate size and were sectioned at 15 µm on a rotary microtome. The paraffin wax sections containing the sample were washed in dH_2O and smoothed in a 40°C water bath, before being fixed to microscope slides using Haupts solution (1% gelatin, 15% glycerol, 2% phenol). The slides were dried overnight on a slide warmer set to 40°C, with a dust cover.

Once dry, the slides were washed in 100% xylol for 10 min in order to dissolve the paraffin wax. The slides were then hydrated through an ethanol series by washing them in ethanol concentrations of 100%, 100%, 95%, 70%, 50%, 30%, 0% (dH20) and 0% (dH20) for 2 min each, after which they were incubated in blocking buffer (5% skim milk powder in PBS, pH8) for 15 min at room temperature. This was followed by an incubation in blocking buffer containing primary antibody at a 1:100 dilution overnight at 4°C. The slides were washed three times in PBS (0.8% NaCl, 0.02% KCl, 0.144% Na_2HPO_4, 0.024% $KH_2PO_{2)}$) and incubated in the dark in blocking buffer containing the secondary AlexaFluor 514 fluorescent antibody (ThermoFischer Scientific) at a 1:250 dilution for 2 hours at room temperature. The slides were washed an additional three times in PBS. The sections were mounted in Vectashield antifade medium (Vector Laboratories, Burlingame, USA), and sealed with vulcanizing solution (i.e. rubber cement). The sections were then stored at 4°C in the dark and visualised within 24 hours on the Zeiss LSM 880 confocal microscope (Carl Zeiss Microscopy, Jena, GmbH) using a 514 nm laser, with 500 nm – 570 nm bandpass (BP) and 635 nm – 735 nm BP emission filters for antibody and autofluorescence signals, respectively. Brightness settings were adjusted until autofluorescence was not easily visible and used as standard settings for all images.

3.3.6 Enzyme assays

Enzyme activity under various pHs and temperatures was assessed through the use of the DNS method (Miller 1959) to detect reducing sugar content (Supplementary Data 3.5). For the pH assays, 50 µg (1 mg/ml) of recombinant protein was incubated with 0.95 ml beechwood xylan (1%, Merck) in buffers ranging from pH 3.0 to 12.0. Citrate (pH 3.0 to 6.0), K-phosphate (pH 7.0 to 8.0) and glycine-NaOH (pH 9 to 12) buffers were used, all at concentrations of 50 mM. The mixture was incubated for 30 min at 80°C, after which the reaction was terminated by adding an equal volume of DNS reagent and boiling the mixture for 10 min. For the temperature assays, 50 µg (1 mg/ml) of recombinant protein was incubated with 0.95 ml 1% beechwood xylan (Merck) in 50 mM glycine-NaOH buffer (pH 9) at temperatures ranging between 40°C and 100°C. The mixture was incubated for 30 min at the appropriate temperature, after which the reaction was terminated by adding an equal volume of DNS reagent and boiling the mixture for 10 min. The optical absorbance at 540 nm was measured and a standard curve was used to convert the measurements into quantities of reducing sugar. The standard curve was constructed by measuring and plotting the optical absorbance at 540 nm for a series of known concentrations of xylose. One unit of enzyme activity was defined as the amount of enzyme required for the release of 1 µmol of reducing sugar (xylose) per minute, per mg of tissue, under the assay conditions. Assays were performed in triplicate, and blanked with deactivated enzyme controls (enzyme boiled for 15 min) for each pH and temperature studied.

For the xylanase activity of plant protein extracts, 1 mg (20 µg/ul) of plant protein extract was incubated with 150 µl beechwood xylan (Merck) in glycine-NaOH buffer (pH 9) for 3 hours at 60°C, after which the reaction was terminated by adding an equal volume of DNS reagent, and boiling the mixture for 10 minutes. The optical absorbance at 540 nm was measured and a standard curve was used to convert the measurements into quantities of reducing sugar. The standard curve was constructed by measuring and plotting the optical absorbance at 540 nm for a series of known concentrations of xylose. One unit of enzyme activity was defined as the amount of enzyme required for the release of 1 µmol of reducing sugar (xylose) per minute under the assay conditions. Assays were performed in triplicate for each plant line, and normalised to de-activated enzyme controls (enzyme boiled for 15 min) for each plant line.

3.3.7 Dry weight measurements and biomass sugar release assays

Wild-type and transgenic eight week old plants were collected. Multiple replicates (between 19 and 34 for each plant line) were bulked into one group per plant line. The weight of each plant was recorded. Each group of plants was then divided into two subgroups. Subgroup one was immediately subjected to heat treatment of 80°C in an oven for two hours, while subgroup 2 was immediately incubated at room temperature for two hours. After incubations, all plants were allowed to dry at room temperature in a laminar flow cabinet. The weights of all plants were recorded after drying and used to calculate average dry weight for each plant line (Supplementary Data 3.3). After drying, heat treated and non-heat treated plants were kept separate and stored at room temperature for six months, after which they were ground into fine powder in liquid nitrogen. The ground tissue was then subject to sugar release assays. Briefly, 10 mg of plant tissue was further homogenised in 320 µl of sodium acetate buffer (50 mM NaAc, pH 5), using ceramic beads and the Thermo Savant FastPrep 120 instrument (GMI, Ramsey, USA), after which 80 µl of enzyme solution (2 U/ml cellulase from *Trichoderma reesei* ATCC 26921, Merck) was added. The mixture was briefly vortexed and incubated at 37°C for 2 hours, with shaking at 300 rpm. After incubation, the samples were immediately cooled on ice, and then centrifuged (13 000 g, 4°C) for 45 min. The supernatant for each sample was collected and subjected to DNS assays as described above, using glucose as a standard. Enzyme assays were performed in triplicate for each plant line, and normalised to de-activated enzyme controls (cellulase boiled for 15 min) for each plant line. Reactions using Avicel (Fluka, Bucharest, Romania) as a substrate were used as a positive control.

3.3.8 Plant protein extractions

Frozen plant tissue was homogenised to a fine powder in liquid nitrogen. Approximately 4 volumes of protein extraction buffer (50 mM sodium phosphate pH 6.5, 0.5 mM NaCl and 2 mM PMSF) was added to the ground tissue, and incubated at 4°C for 20 min, with occasional vortexing. The resulting mixture was centrifuged at 13 000 g for 30 min at 4°C. The supernatant was transferred to new 1.5 ml Eppendorf tube and was stored at -20°C.

3.3.9 Western blotting

A total of 15 µl of crude lysate from each plant, prepared as described above, was separated by SDS-PAGE on 8% Tris-glycine gels. Protein was transferred from the gels to PVDF membranes using the iBlot2 system and protein stacks (ThermoFisher Scientific). The membrane was incubated in 50 ml of blocking buffer (5% milk powder in TBS-T) for 1 hour, with gentle agitation, then washed three times for 5 min each in 50 ml of TBS-T buffer (50 mM Tris-Cl, 150 mM NaCl, 0.05% Tween-20, pH 7.5). From this point, all incubations were carried out with gentle agitation on a rotating shaker. The membrane was incubated in 50 ml TBS-T buffer containing a custom synthesised antibody against both CBM22_L and JX_WT (polyclonal, raised in rabbit to the antigen CYQSSGSSDITVGGT) at a dilution of 1:1500 overnight at 4°C. After this, the blot was washed three times for 5 min each in 50 ml of TBS-T buffer and was incubated in 50 ml TBS-T containing anti-rabbit secondary antibodies conjugated to horseradish peroxidase at a dilution of 1:10 000 for 1 hour at room temperature. The blot was washed as described previously and was visualised using chemiluminescence and exposure to x-ray film, with SuperSignal™ West Pico chemiluminescent substrate (ThermoFischer Scientific) and CL-Xposure™ x-ray film (ThermoFischer Scientific).

3.3.10 Statistical tests

Two-tailed Student’s t-tests assuming unequal variance were performed in order to evaluate the statistical significance between means in the datasets. Bonferroni adjustments (Bland and Altman

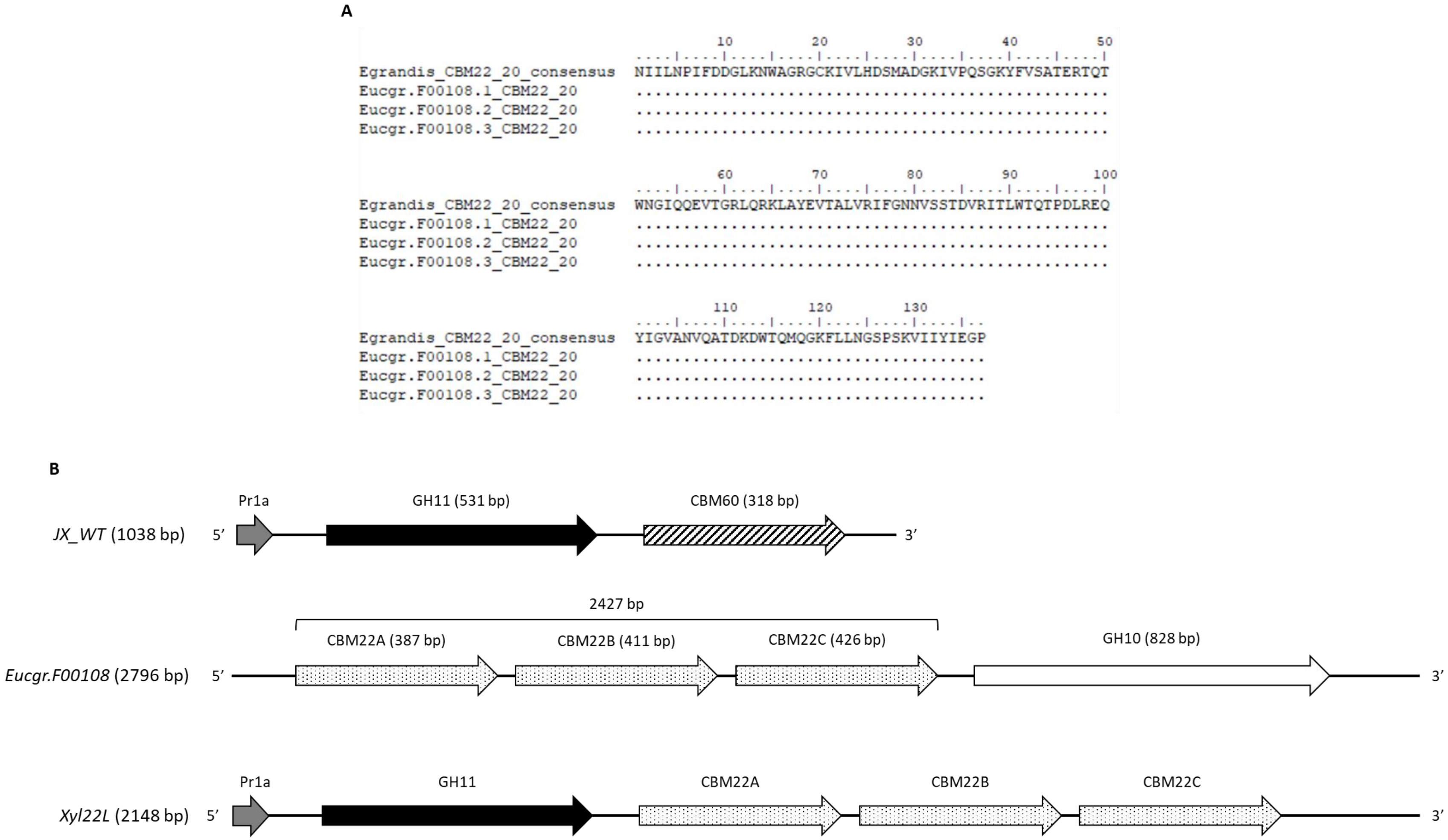
1995) were applied to all statistical tests. However, due to the high stringency of Bonferroni adjustments (Perneger 1998), data was interpreted based on both adjusted and non-adjusted values.

3.4 Results

3.4.1 Design of enzyme and generation transgenic plants

The coding sequence of *JX125044* (*Mxyl*, GENBANK accession: AFP81696; Verma et al. 2013), isolated from a soil compost metagenome and tested by Mir et al. (2017), was modified to incorporate the xylan-binding CBM22 domains from *E. grandis Eucgr.F00108*. Briefly, the full coding sequence of the three CBM22 repeats present in *Eucgr.F00108* (Figure 3.1A) was substituted into *JX125044,* replacing the native CBM60 domain (Figure 3.1B). The resulting coding sequence consisted of (in 5'to 3' order) a tobacco pathogenesis-related protein 1a (Pr1a) signal peptide, a plant-specific Kozak consensus sequence (GCCACCATGG), a GH11 domain and three CBM22 domains (CBM22A, -B, and -C), and was designated *Xyl22L* (Figure 3.1B).

T1 transgenic plants were generated through a floral dipping method (Clough and Bent 1998) and PCR analysis was performed in order to confirm the presence of the Xyl22L expression cassette in the T1 plants (Supplementary Figure 3.1, Supplementary File 3.1). Gene-specific primers were used (Supplementary Table 1) to amplify the CDS of *Xyl22L,* which should result in a band of 2148 bp. Bands of the appropriate size were identified in 32L_1_9, 32L_4, 32L_5, 32L_6, 32L_7, 32L_11, 32L_3_2, and 32L_4_2 (Supplementary Figure 3.1). The six plant lines which did not produce bands were excluded from the study. The T1 transgenic plants that produced bands of the expected size in the PCR analysis were taken to the T2 generation in order to identify homozygotes, via Chi-square analysis (Supplementary Data 3.2). The selected homozygous T3 lines were 32L_1_9_2, 32L_3_2_7, 32L_4_1_1, 32L_4_2_5, 32L_5_1_4 and 32L_6_3, which were designated Xyl22LA, -B, -C, -D, -E and -F, respectively.



```
                                   10        20        30        40        50
                          ....|....|....|....|....|....|....|....|....|....|
Egrandis_CBM22_20_consensus  NIILNPIFDDGLKNWAGRGCKIVLHDSMADGKIVPQSGKYFVSATERTQT
Eucgr.F00108.1_CBM22_20      ..................................................
Eucgr.F00108.2_CBM22_20      ..................................................
Eucgr.F00108.3_CBM22_20      ..................................................

                                   60        70        80        90       100
                          ....|....|....|....|....|....|....|....|....|....|
Egrandis_CBM22_20_consensus  WNGIQQEVTGRLQRKLAYEVTALVRIFGNNVSSTDVRITLWTQTPDLREQ
Eucgr.F00108.1_CBM22_20      ..................................................
Eucgr.F00108.2_CBM22_20      ..................................................
Eucgr.F00108.3_CBM22_20      ..................................................

                                  110       120       130
                          ....|....|....|....|....|....|..
Egrandis_CBM22_20_consensus  YIGVANVQATDKDWTQMQGKFLLNGSPSKVIIYIEGP
Eucgr.F00108.1_CBM22_20      .....................................
Eucgr.F00108.2_CBM22_20      .....................................
Eucgr.F00108.3_CBM22_20      .....................................
```

Figure 3.1 Representations of constructs and sequences used to produce Xyl22L. A: Alignment of the CBM22 region from *E. grandis Eucgr.F00108*. The consensus amino acid sequence and position is displayed at the top of the alignment. Dots represent identity to the consensus. Where residues differ, they are indicated by the appropriate letter in the alignment. Shaded residues indicate similarity to the consensus. B: Schematic representations of constructs used in this study. The name of the gene is listed to the left of each schematic. The individual features/domains are represented by differently filled arrows, with the direction of the arrow indicating direction of transcription. The name of the feature/domain is indicated above each arrow.

3.4.2 Protein expression/accumulation in plants

To determine if Xyl22L accumulated in the tissue of transgenic plant lines, qPCR analyses and western blots were performed. cDNA was synthesised from RNA extracted from the leaves of four week old transgenic plants, and was used to quantify the expression of *Xyl22L* (Figure 3.2, Supplementary File 3.2). Almost all plants lines expressed *Xyl22L* at relatively low levels, except for Xyl22LF plants which showed relatively high expression (Figure 3.2). Western blots were also performed on TSP extracted from four week old transgenic plant leaves (Figure 3.3), using a rabbit-derived antibody to target both Xyl22L and JX_WT. The antibody successfully bound to both Xyl22L and JX_WT (Supplementary Figure 3.3). However, the antibody also showed cross-reactivity with other proteins in the plant (Figure 3.3), although none were in the expected size range of Xyl22L (approximately 77 kDa). Using the antibody, a faint band of the correct size range was identified in Xyl22LB, -C, -D and -E. While Xyl22LF showed the highest expression of *Xyl22L*, it did not show a band of the appropriate size in the western blot (Figure 3.3).

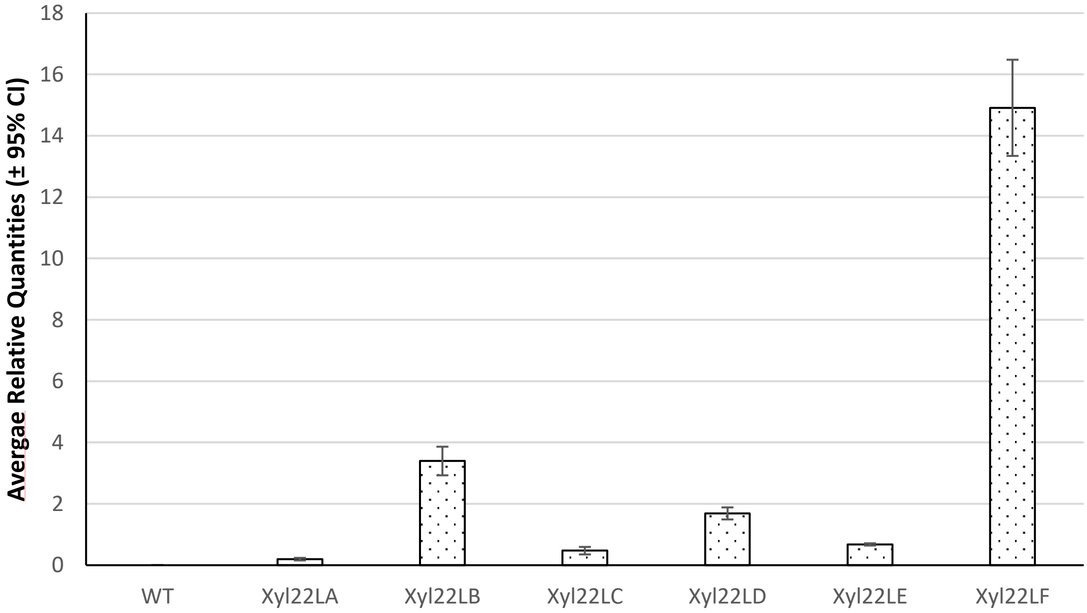


Figure 3.2 Relative expression of *Xyl22L* in transgenic plant lines. qPCR was performed to determine expression level of *Xyl22L* in various transgenic plant lines. The name of the transgenic plant line is indicated at the bottom of the graph. Expression was quantified as average relative expression normalised to two *A. thaliana* reference genes, *AtAct2* and *AtUbq5.* A total of three biological repeats with three technical repeats were performed for each plant line. T-bars indicate standard error of the mean of biological replicates.

3.4.3 Phenotyping of plants

To determine the effect of heterologous expression of Xyl22L on growth and development, the phenotypes of the transgenic plants were examined. Transgenic plants were grown and compared at six and eight weeks old (Figure 3.4A and B, Supplementary File 3.3). There were no discernible developmental differences between transgenic and wild-type plants at any stage, indicating that expression of Xyl22L *in planta* had no significant effect on growth and development of the plants. Phloroglucinol staining of stem cross sections (Figure 3.5, Supplementary File 3.4) also showed that interfascicular fibres and vascular bundles of transgenic plants were indistinguishable from those of wild-type plants. When compared to transgenic plants expressing JX_WT, there also appeared to be little difference. However, when dry-weights of all above-ground tissues of the plants were compared (Figure 3.6) there was a significant increase in biomass in five out of six Xyl22L transgenic lines, when compared to wild-type plants and JX_WT transgenics. The increase in biomass also strongly correlated (*r* = 0.76, Supplementary Data 3.3) with the level of expression of the transgene (Figure 3.2). Considering that the Xy22L expressing plants did not appear to be larger than wild-type plants, the increased biomass may be due to an increase in leaves, siliques and number of stems, rather than an increase in size of the stem itself.

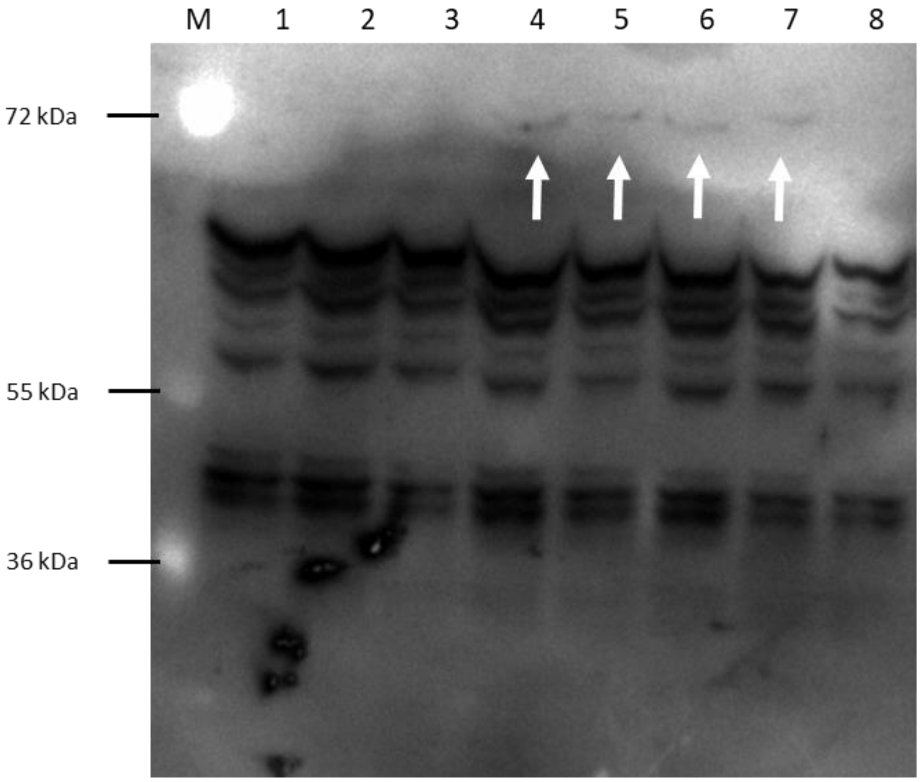


Figure 3.3 Western blot of total soluble protein (TSP) extracted from four week old transgenic plants. M indicates the Thermo Scientific PageRuler Plus Prestained Protein Ladder. Blots of TSP are shown for 1: wild-type, 2: 32E (empty vector control), 3: Xyl22LA, 4: Xyl22LB, 5: Xyl22LC, 6: Xyl22LD, 7:Xyl22LE, 8: Xyl22LF. The white arrows indicate possible Xyl22L protein.

3.4.4 Immunolocalisation

Immunolocalisation studies were performed on transgenic plants to determine whether Xyl22L bound to the SCW. Six week old plant stem cross sections were labelled with fluorescent antibodies raised against Xyl22L and visualised via confocal fluorescent microscopy (Figure 3.7, Supplementary Figure 3.4, Supplementary File 3.5). Fluorescence was seen in chloroplasts in all cases, indicating cross reactivity of the antibody with these structures. Minor labelling was seen throughout the SCWs of the 32E empty vector plant lines. Of the six transgenic Xyl22L expressing lines examined, Xyl22LA, -C, -D, and -F, showed labelling along the inside perimeter of the SCW, while Xyl22LB and –E did not show this pattern (Figure 3.7). WT plants, as well as the line expressing JX_WT did not show any specific labelling of the SCW. This indicates that the CBM22 repeats of Xyl22L may be functional, allowing targeting of the expressed enzyme to the cell wall. The lack of labelling in JX_WT also indicates that the *E. grandis* CBM22 repeats may allow for better targeting to the cell wall.

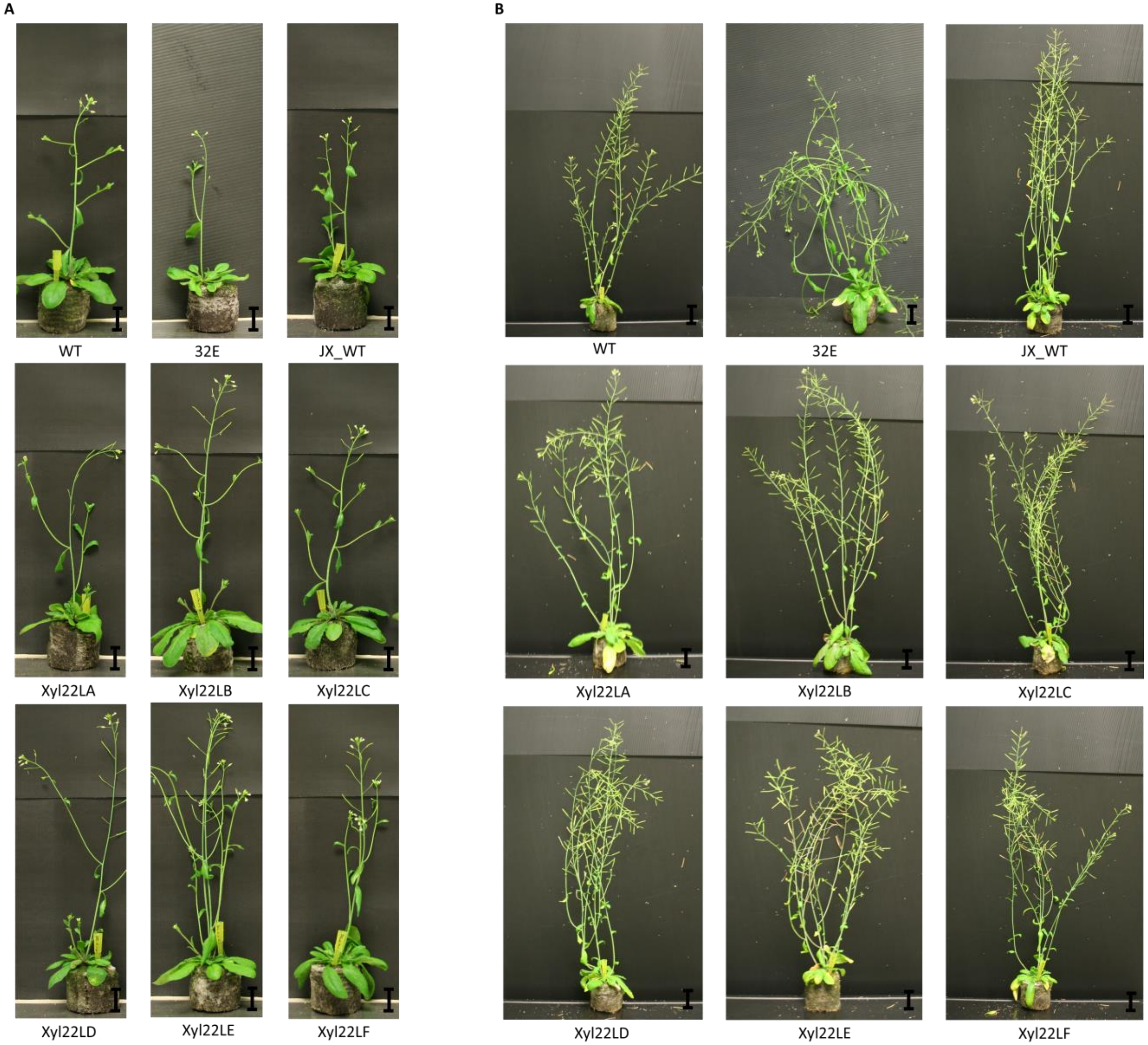


Figure 3.4 Growth of transgenic plants. A: Six week old transgenic plants. One representative of each transgenic line is shown, compared to control plants. B: Eight week old transgenic plants. One representative of each transgenic line is shown, compared to control plants. Scale bars represent 20 mm. In all cases the plant line is indicated below the image. All raw images and replicates are contained in Supplementary File 3.3.

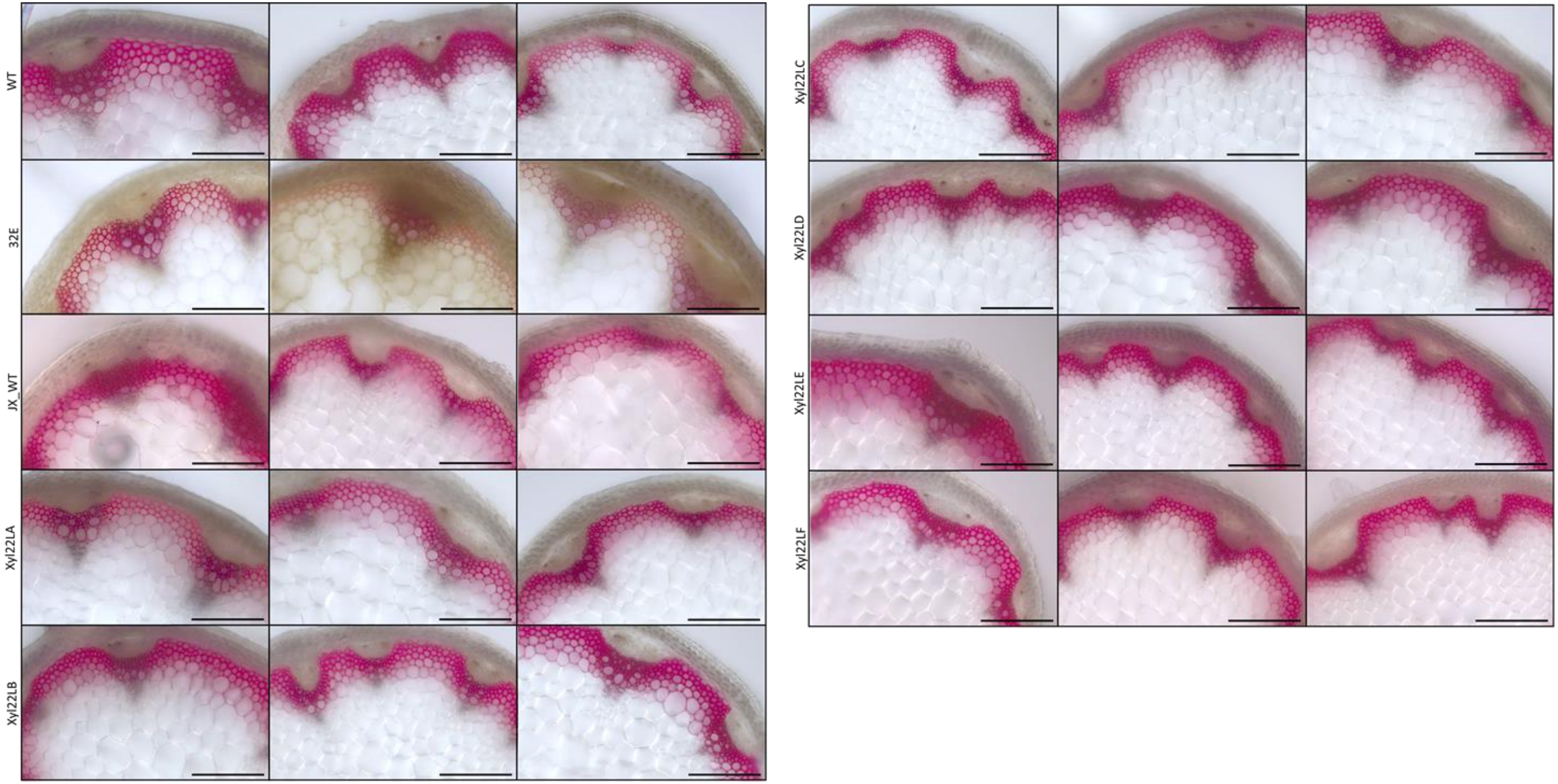


Figure 3.5 Phloroglucinol stained cross sections of lower inflorescence stems. One representative of each transgenic line is shown, compared to wild-type and empty vector controls. The plant line is indicated on the left of the set of images. Scale bars represent 50 µm. All raw images and replicates are contained in Supplementary File 3.4.

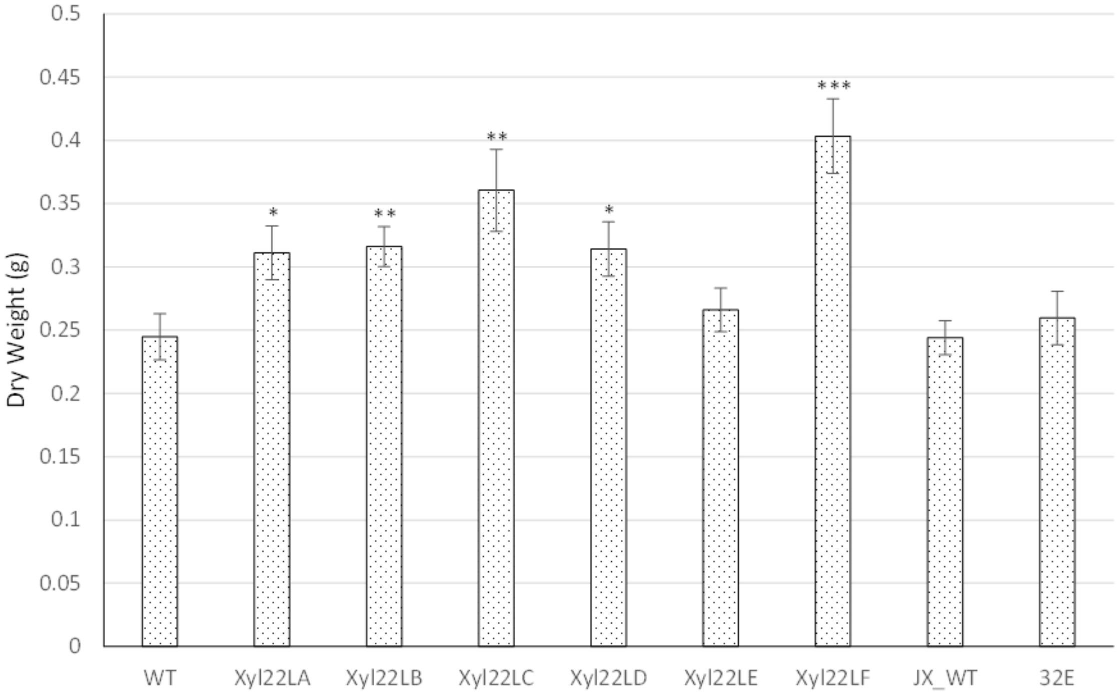


Figure 3.6 Dry weight (g) measurements of transgenic plants. The plant line is indicated at the bottom of the graph. The weight of the dried plant biomass is indicated on the y-axis. Asterisks indicate statistical increase compared to WT plants based on a Student's t-test. *: p < 0.05, ** : p < 0.01, ***: p < 0.001. For each plant line *n* is between 19 and 34.

3.4.5 Sugar release assays

Considering that the heterologous expression of Xyl22L in plants resulted in a growth phenotype and that Xyl22L appeared to bind to the SCW, experiments were performed to determine the effect that heat treatment had on recalcitrance of transgenic tissues to enzymic hydrolysis. Plant materials were collected at eight weeks old. In each case, half of the material was subjected to heat treatment at 80°C for 2 hours while the other half was kept at room temperature for 2 hours. The material was then dried in a laminar flow cabinet, ground in liquid nitrogen and subjected to sugar release assays (Figure 3.8). Before heat treatment, only one plant line, Xyl22LE, showed a significant decrease in sugar release, relative to wild-type plants. After heat treatment, no plant line showed a significant difference compared to WT. However, while wild-type plants showed no difference in sugar release before and after heat treatment, most transgenic plant lines showed decreased sugar release without heat treatment when compared with post-heat treated plant material of the same transgenic line.

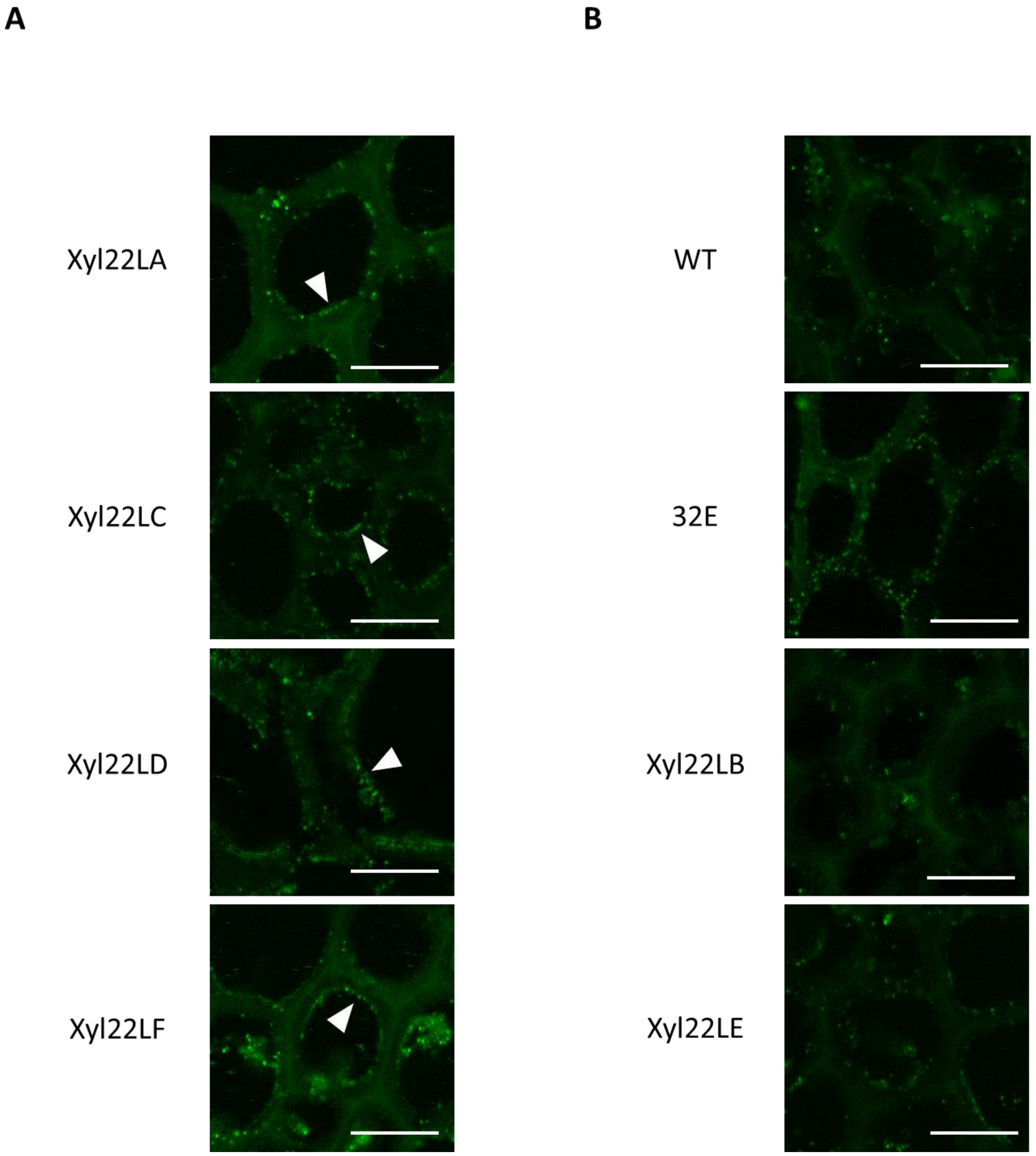


Figure 3.7 Fluorescent confocal microscopy of representative transgenic *A. thaliana* interfascicular fibre cross sections. The plant line is indicated on the left of each image. The signal from the secondary antibody is shown in the channel. A: Plant lines showing labelling near secondary cell walls. B: Plant lines showing no labelling near secondary cell walls. Scale bars indicate 10 µm. Additional replicates as well as raw images can be seen in Supplementary Figure 3.4 and Supplementary File 3.5.

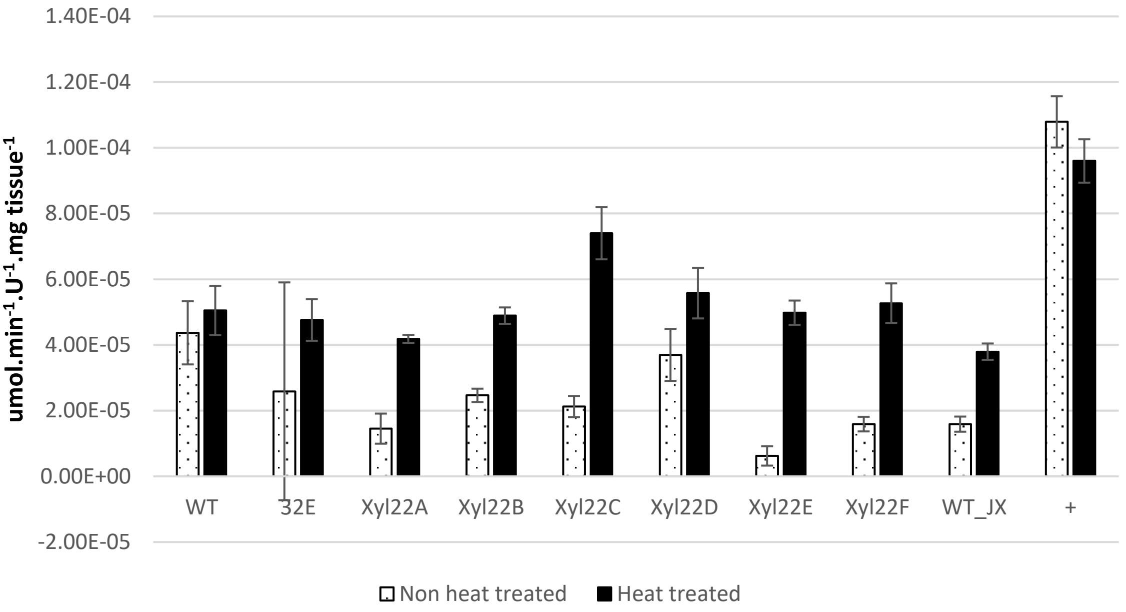


Figure 3.8 Reducing sugar release from cellulase treated plant tissues. The amount of reducing sugar liberated when heat treated and non-heat treated plant tissues were subjected to hydrolysis by *T. reesei* cellulase. The plant line is indicated at the bottom of the figure. The amount of sugar released is indicated on the Y-axis. WT: Wild-type control. +: Reaction carried out using Avicel as a substrate. 32E: Empty vector control. All measurements were normalised to reactions with inactivated cellulase (boiled for 15 min) from the same plant line. T-bars indicate standard error (*n* = 3).

3.4.6 Synthesis of proteins and protein characterisation

Xyl22L and JX125044 containing the Pr1a signal peptide (designated JX_WT, Figure 3.1A) were commercially synthesised and characterised in order to determine the properties of the chimeric enzyme. The optimum temperature and pH for enzyme activity were determined with DNS assays. JX_WT showed maximum relative activity at 50°C and pH 9 (Figure 3.9, Supplementary Data 3.5). The T_{opt} value was not consistent with previous studies, which identified T_{opt} for the wild-type JX125044 to be 80°C when expressed in *E. coli* (Verma et al. 2013), and JX_WT expressed in *A. thaliana* (Mir et al. 2017). The optimum pH of catalysis was identical to that previously described (Mir et al. 2017; Verma et al. 2013). Synthesised Xyl22L showed no activity at any pH or temperature, indicating that the synthesised protein had no xylanase activity. Additionally total soluble protein (TSP) extracted from *Xyl22L* transgenic plant lines did not show increased xylanase activity compared to wild-type plants (Figure 3.10, Supplementary Data 3.6), further suggesting that the catalytic domain of Xyl22L is non-functional. In previous studies, TSP extracted from JX_WT expressing plants showed increased xylanase activity relative to wild-type plants (Mir et al. 2017), which was not seen in this study for Xyl22L.

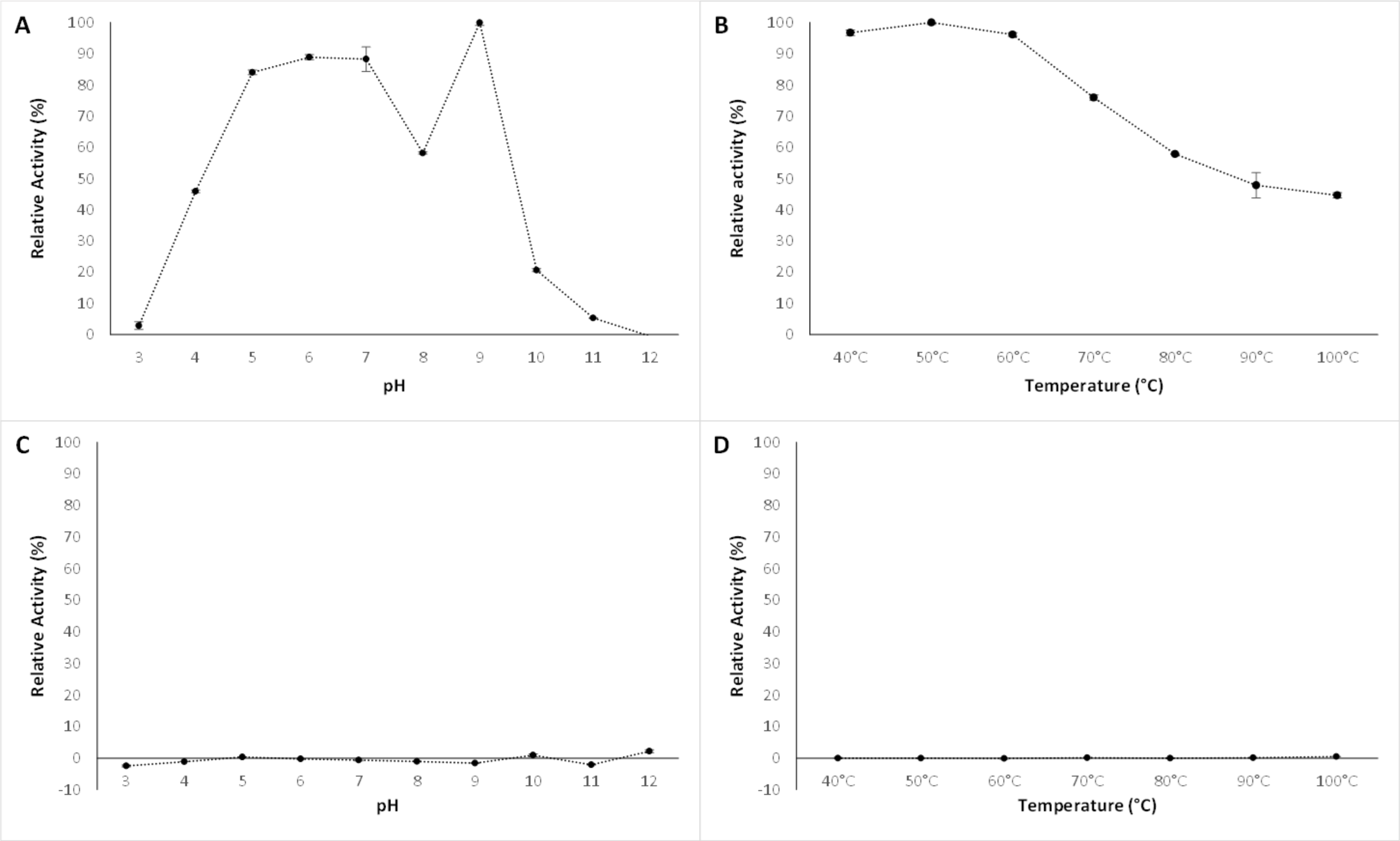


Figure 3.9 Optimum temperature and pH enzyme assays for WT_JX and Xyl22L, expressed as relative xylanase activity (%). In all cases, the relative activity is displayed on the Y-axis and the condition being tested (pH or temperature) is displayed on the X-axis. A: Optimum pH of WT_JX. B: Optimum temperature of JX_WT. C: Optimum pH of Xyl22L. D: Optimum temperature of Xyl22L. Readings for pH were normalised to JX_WT at pH 9, and readings for temperature were normalised to JX_WT at 50°C. All measurements were blanked to de-activated enzyme controls. Standard error is indicated by T bars (*n*=3).

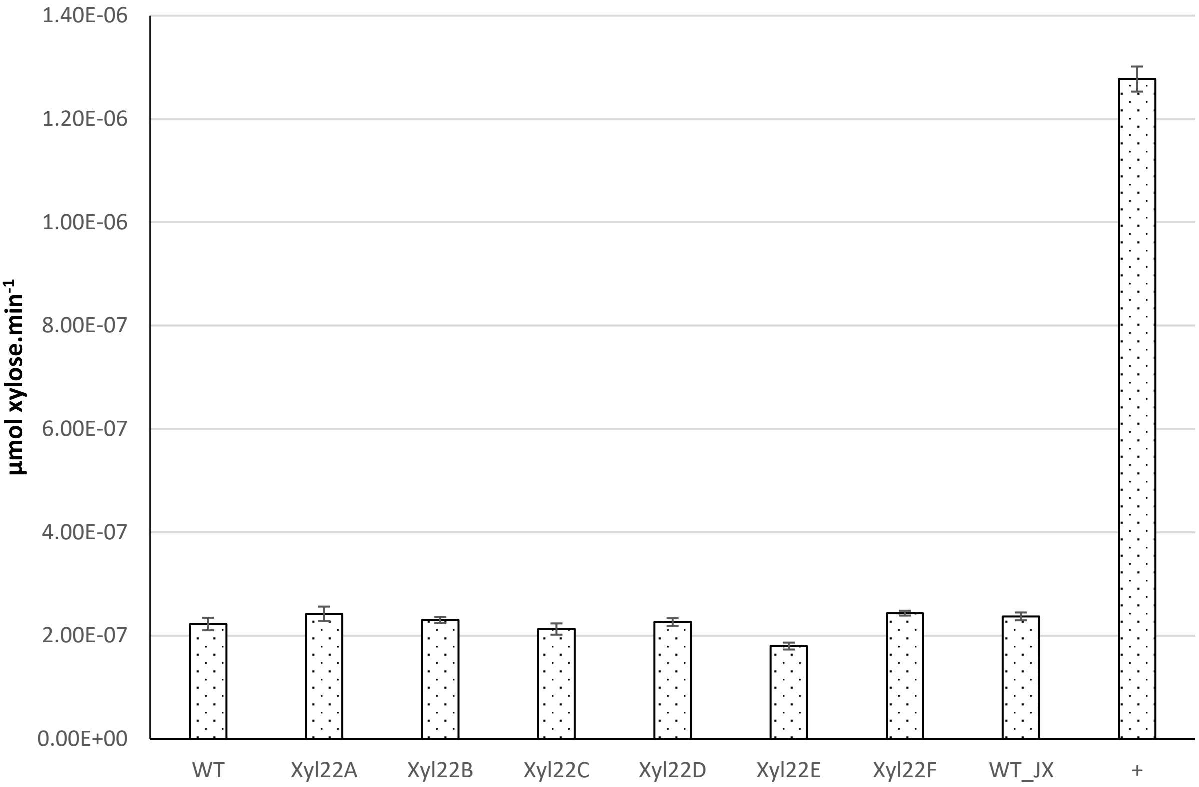


Figure 3.10 Xylose liberated from beechwood xylan (BWX) by plant TSP extracts. The amount of xylose liberated when TSP extracts were incubated with BWX (1%, 50 mM glycine-NaOH buffer, pH9) for 3 h at 60°C. The plant line from which TSP was extracted is listed at the bottom of the graph. WT: Wild-type control. +: Reaction carried out with pure synthesised JX-WT. All measurements were normalised to inactivated TSP extracts (boiled for 15 min) from the same plant line. T-bars indicate standard error (*n* = 3).

3.5 Discussion

In planta expression of extremely thermophilic CAZymes is a promising strategy for the reduction of recalcitrance of lignocellulosic feedstocks used for the synthesis of bioproducts (Mir et al. 2014). However, the efficiency of this strategy may be improved by using CBMs to target CAZymes to specific locations or biopolymers in the lignocellulosic biomass, and plant derived CBMs are likely to be most effective at targeting CAZymes to plant biopolymers. In this study, a chimeric protein was designed that consisted of a thermostable GH11 domain (Verma et al. 2013) and xylan targeting CBM22 domains derived from *Eucalyptus grandis* (*Xyl22L,* Figure 3.1A and B). The protein was commercially synthesised, characterised and heterologously expressed in *Arabidopsis thaliana*. This is the first report of such an enzyme being expressed in plants, and of an extremely thermophilic hydrolase being targeted to plant biopolymers using plant derived CBMs. The effect on growth of the plants, recalcitrance to enzymic digestion of the biomass, and the ability of the enzyme to adhere to the secondary cell wall (SCW) were assessed.

3.5.1 *In planta* expression of Xyl22L has no negative effect on plant growth

Due to the relatively low activity of extremely thermophilic enzymes at mesophilic temperatures, heterologous expression of the Xyl22L in *A. thaliana* should have little to no effect on plant growth and development, allowing for accumulation of enzyme *in planta* without the associated deleterious effects (Castiglia et al. 2016; Mir et al. 2014; Mir et al. 2017).

Xylan is a major component of the plant SCW, and any modifications to xylan could have a significant impact on SCW structure and function (Brown et al. 2007; Busse-Wicher et al. 2016; Lee et al. 2007; Peña et al. 2007; Ratke et al. 2018; Rennie and Scheller 2014). Heterologous expression of xylanases and acetyl-xylan esterases in plants have also yielded biomass that is less recalcitrant to enzymic digestion (Chen et al. 2017; Mir et al. 2014; Mir et al. 2017; Pawar et al. 2016). We found that JX_WT was able to hydrolyse beechwood xylan at various temperatures and pHs (Figure 3.1B, Figure 3.9A and

B). JX_WT showed activity over most pHs and temperatures, with maximum activity recorded at pH 9 and 50°C, respectively (Figure 3.9A and B). While the optimum pH was in line with previously published studies (Mir et al. 2017; Verma et al. 2013), the optimum temperature was significantly lower than the previously noted 80°C. This is most likely due to a combination of the modifications and expression systems used during protein synthesis. The original JX125044 protein (Verma et al. 2013) and JX_WT (Mir et al. 2017) both show optimal activity at 98°C, despite expression in *E. coli* and *A. thaliana,* respectively. However, JX_WT contains the plant-derived Pr1A localisation signal, which is normally cleaved off of mature proteins (Hammond-Kosack et al. 1994; Ziegler et al. 2000). When JX_WT is expressed in *A. thaliana*, Pr1a will be correctly processed and the signal will be cleaved off, allowing for normal folding of the protein. When JX_WT is expressed in *E. coli*, the cell may not be able to properly process the plant-based signal peptide, resulting in a change in tertiary structure of the enzyme and concomitant lowering of the T_{opt}.

Xyl22L showed little to no xylanase activity across all tested temperatures and pHs (Figure 3.9C and D, Supplementary Data 3.5). Additionally, total soluble protein (TSP) from *Xyl22L* expressing transgenic plant lines failed to liberate more xylose from beechwood xylan than TSP from wild-type plants (Figure 3.10, Supplementary Data 3.6). This suggests that xylanase activity was abolished by substitution of the *E. grandis* CBM22 repeats into the C-terminal of JX_WT. CAZymes are modular and CAZyme domains operate independently. However, the insertion, deletion or addition of domains can affect the function of a CAZyme (Botha et al. 2017). In the case of Xyl22L, the increased size of the inserted CBM22 repeats may be hindering proper folding of the enzyme catalytic domain. Additionally, the native linker sequence in Xyl22L may either be too short or rigid to facilitate the proper folding of the GH11 catalytic domain as well as the CBM22 repeats (George and Heringa 2002). Enzyme function may also be dependent on interactions between two domains within the protein (Venditto et al. 2015). They may share features such as secondary structures or disulphide bridges, resulting in reduced or lost function if either domain is removed or altered. The original CBM60 domain in JX_WT may be required for proper function of the catalytic domain, and substitution with a different CBM may have had a deleterious effect on the catalytic domain. It should be noted that while xylanase activity of Xyl22L appears to be absent, there was some evidence that the CBM22 repeats were still functional. Considering that the CBM22 repeats in this study were derived from the plant species *E. grandis*, they have a higher chance of folding correctly in a plant-based intracellular environment like *A. thaliana*, as opposed to *in vitro* systems.

TSP from JX_WT transgenic plants was previously shown to liberate more xylose than wild-type extracts (Mir et al. 2017), but we were not able to replicate the result in this study (Figure 3.10, Supplementary Data 3.6). This could be due to two factors: the first is that protein extracts in Mir et al. (2017) were heat treated in order to precipitate and remove mesophilic proteins, resulting in purer thermostable protein extracts for the assay and preventing extraneous plant proteins from interfering with the reaction. In the experiments performed in this study, heat treatment of the protein extract from transgenic plant lines appeared to remove Xyl22L from the extracts; therefore, this purification step could not be used. The second factor is the lower reaction temperature for the assay (60°C, as opposed to 80°C in previous studies), as determined by characterisation of synthesised JX_WT (Figure 3.9B). This would also reduce activity of the thermostable xylanase.

Six and eight week old transgenic plants did not show obvious differences in growth and development, when compared to control plants (Figure 3.3). Additionally, cross sections of lower inflorescence stems did not show any abnormalities in SCW structure (Figure 3.4). However, measurements of dry plant weight showed an increase in total biomass in transgenic lines, relative to wild-type plants (Figure 3.5), which seemed to be positively correlated with transgene expression levels (r = 0.76). Additionally, there was no increase in biomass of 32E empty vector lines or JX_WT transgenics, relative to wild-type plants (Figure 3.6). This suggests that the increase in biomass is in response to expression of Xy22L, and not an artefact of random insertion into the genome caused by *Agrobacterium* mediated transformation (Clough 2005). The overexpression of CBMs in plants has been shown to have an effect on the cell wall, resulting in altered physical characteristics such as increased biomass (Guillén et al.

2010; Nardi et al. 2015; Safra-Dassa et al. 2006; Shoseyov et al. 2006). The CBM22 repeats in Xyl22L may have a similar effect in the transgenic plant lines, resulting in elongated cells. However, the stem cross sections show cell size and number (Figure 3.4), but give no indication of the longitudinal dimensions of the cells. The transgenic plants were also not noticeably taller, so the increase in biomass could be due to greater numbers, or increased thickness of inflorescence stems on a single plant (Roberts and Shirsat 2006), though this was not formally assessed. Additionally, increased biomass could be due to changes in leaf size and number, which along with other alterations in phenotype have been associated with heterologous expression of enzymes in plants (Safra-Dassa et al. 2006; Tsai et al. 2012).

3.5.2 Xyl22L accumulates in the plant and binds to the secondary cell wall

An extremely thermophilic enzyme heterologously expressed *in planta* should allow for accumulation of the protein over the lifetime of the plant, with no deleterious effects on plant growth and development at mesophilic temperatures. CBMs can target enzymes to specific biopolymers (Hervé et al. 2010), and the inclusion of an *E. grandis* derived CBM22 could increase efficiency of adhesion of the chimeric enzyme to the SCW, allowing for targeting and “pre-packaging” of extremely thermophilic hydrolases in lignocellulosic biomass (Mir et al. 2014). Additionally, mature fibre cells in trees are dead (Plomion et al. 2001), increasing the importance of proper enzyme targeting during growth and development of the fibre cells. CBM22 domains have well-documented xylan-binding function (Araki et al. 2006; Cantarel et al. 2009; Lombard et al. 2014; Najmudin et al. 2010; Sainz-Polo et al. 2015), and the *E. grandis* derived CBM repeats should have a particular affinity for glucuronoxylan, the dominant hemicellulose in hardwoods (Scheller and Ulvskov 2010).

Western blots of TSP extracted from four week old leaves showed the presence of the protein in the Xyl22LB, -C and -D plant lines (Figure 3.3). In all cases, the protein band was very faint, indicating that only a small amount of protein was present. These plant lines showed relatively low expression of *Xyl22L*, while the highest expressing line, Xyl22LF, did not have Xyl22L present in the TSP fraction

(Figure 3.6). Considering that the addition of the *E. grandis* CBM22 repeats may have been responsible for abolishing the activity of the GH11 catalytic domain, Xyl22L may have been misfolded, resulting in an unstable protein that would be quickly degraded in the plant. Xyl22L contained the Pr1a signal peptide, which causes a protein to be secreted in the apoplast (Hammond-Kosack et al. 1994; Ziegler et al. 2000). Proteases are abundant in plant apoplasts (Delannoy et al. 2008; Pillay et al. 2014), and a denatured protein would be highly susceptible to proteolysis. Also, plant-based protein expression systems have been known to degrade and eliminate proteins based on improper post-translational modification, folding and/or accumulation *in planta* (Doran 2006; Hellwig et al. 2004; Kusnadi et al. 1997). Lastly, Xyl22L may be targeted to and locked into the SCW, hindering extraction and reducing yield in the soluble protein fraction.

Fluorescent confocal microscopy of Xyl22LA, -C, -D and -F showed fluorescence lining the inside of the SCW that was not present in WT lines (Figure 3.7, Supplementary Figure 3.4). The empty vector control line, 32E, also showed some fluorescence, but this was distributed throughout the SCW. However, 32E had much thinner cell walls, possibly an artefact of random insertion into the genome inherent in *Agrobacterium* mediated transformation (Clough 2005), or the presence of the *ccdb* cytotoxicity gene in unmodified pMDC32 (Curtis and Grossniklaus 2003). The thinner walls were probably more easily infiltrated by the antibody, resulting in increased background fluorescence. The lack of signal seen in Xyl22LB and -D may be because of insertional effects attributed to *Agrobacterium* mediated transformation (Clough 2005), or due to the use of an antibody that was not optimised for histochemical applications. Additionally, fluorescence was seen in the chloroplasts of control and transgenic lines (Supplementary Figure 3.4), indicating that this was due to cross-reactivity of the antibody used for labelling. Antibodies specifically designed to detect Xyl22L in fluorescent applications would help to mitigate this issue in future work.

SCWs consist of three layers, designated S1-S3, which vary in proportions of cellulose, hemicellulose and lignin, as well as cellulose microfibril angle (Mellerowicz et al. 2001; Mellerowicz and Sundberg

2008; Plomion et al. 2001). S3 is the innermost layer, and is abundant in xylan (Mellerowicz et al. 2001; Plomion et al. 2001). The fluorescence identified in this area suggests that the CBM22 repeats may have targeted the protein to xylan, allowing Xyl22L to adhere to the SCW. Even though xylan is abundant throughout the entire SCW (Kim and Daniel 2012; McCartney et al. 2005), fluorescence was only observed around the inner-most layer. This could be due to an inability of Xyl22L to penetrate the cell wall, caused by a combination of increased size from the addition of the CBM22 repeats, and possible misfolding of the GH11 catalytic domain. It is noted that the CBM22 repeats derived from *E. grandis* have not been experimentally characterised. While the closest reciprocal blast hit in *A. thaliana* [*AtXyn1*, (Suzuki et al. 2002)] was shown to translocate to the apoplast and SCW, the *E. grandis* ortholog may have a slightly different function, and recognise a more specific part of xylan, such as one of the many side-chains associated with the xylan backbone (Busse-Wicher et al. 2016) or the reducing end tetrasaccharide (Peña et al. 2007). Together, the increase in plant biomass (Figure 3.6) and fluorescent microscopy (Figure 3.7) suggest that the CBM22 repeats present in Xyl22L may be functional and that the protein may have interacted with the SCW, even if the GH11 domain of Xyl22L is not functional.

3.5.3 Xyl22L increases recalcitrance of the secondary cell wall

The heterologous expression of an extremely thermophilic CAZyme *in planta* has the potential to reduce recalcitrance of lignocellulosic tissue to digestion (Mir et al. 2014; Mir et al. 2017). Expression of a thermostable xylanase, for example, could allow for degradation of SCW xylan at high temperatures (Borkhardt et al. 2010), allowing for easier access of other enzymes and more efficient hydrolysis of biopolymers.

While Xyl22L did not retain detectable endo-β-1,4-xylanase activity (Figure 3.9C and D, Figure 3.10), the presence of functional CBM22 repeats could potentially reduce recalcitrance, due to loosening of the SCW or disruption of biopolymer structures (Guillén et al. 2010; Nardi et al. 2015; Safra-Dassa et al. 2006; Shoseyov et al. 2006). Xylan is an important structural component of SCWs (Meents et al.

2018); therefore, any disruption in the structure has the potential to affect recalcitrance to hydrolysis. Sugar release assays showed that plant lines expressing JX_WT, which were previously shown to be less recalcitrant to hydrolysis (Mir et al. 2017; Verma et al. 2013), did not show a significant difference compared to the transgenic lines in this study (Figure 3.8, Supplementary Data 3.4). Since the experimental controls (wild-type biomass and Avicel, as positive and negative controls, respectively) behaved as expected, this could be explained by differences in method. While Mir et al. (2017) examined release of xylose from transgenic plant lines expressing JX_WT during heat treatment, the authors did not test whether the biomass was less recalcitrant to digestion by commercial cellulase after heat treatment. Additionally, Mir et al. (2017) carried out heat treatments in buffers of optimum pH for JX_WT, while in this study, heat treatments were applied to freshly harvested biomass at physiological pH. Since JX_WT is sensitive to pH (Mir et al. 2017; Verma et al. 2013), this may have reduced the effectiveness of JX_WT *in planta*.

Unexpectedly, before heat treatment, sugar release appeared to be lower from transgenic Xyl22L expressing plant lines, though there was no statistically significant difference ($p < 0.05$) in most plant lines. However, one plant line (Xyl22LE) did show a significant decrease in release of reducing sugars (i.e. increased recalcitrance to enzymic digestion), relative to wild-type plants before heat treatment (Figure 3.8, Supplementary Data 3.4). After heat treatment, none of the transgenic lines showed significant difference in sugar release compared to wild-type. However, all transgenic plant lines showed a significant increase ($p < 0.05$) in sugar release from heat treated biomass, compared to non-heat treated biomass of the same transgenic line. The densely packed and highly cross-linked biopolymers of the SCW (Cosgrove and Jarvis 2012) contribute to recalcitrance of lignocellulosic biomass by creating spatial constraints on enzyme access (Himmel et al. 2007). Before heat treatment, Xyl22L may be increasing spatial constraints by coating the SCW and preventing access of hydrolytic enzymes. Prolonged heat treatment would cause Xyl22L to denature, thereby dissociating it from the SCW, loosening spatial constraints, and allowing easier access of hydrolytic enzymes. Through this mechanism, non-heat treated transgenic biomass would be more recalcitrant to enzymic digestion than wild-type plants, but heat treated transgenic biomass would show no difference. This, along with previous experiments (Figure 3.6, Figure 3.7), reinforces the likelihood that the CBM22 repeats in Xyl22L are indeed functional, and able to bring Xyl22L in to close proximity with the SCW even if the GH11 domain is non-functional. If the catalytic GH11 domain of Xyl22L was functional, we would expect the transgenic biomass to be less recalcitrant to enzymic digestion after heat treatment. Normally, active hydrolases can offset this effect by creating additional space for enzyme infiltration through biopolymer degradation (Ding et al. 2012; Himmel et al. 2007; Yang and Wyman 2004). However, this study suggests that the degradation of the heterologously expressed enzymes during heat treatments also creates space and contributes to this effect.

3.6 Conclusion

The *in planta* expression of thermostable hydrolases in lignocellulosic biomass has the potential to reduce recalcitrance of the tissue, and mitigate the costs associated with biomass processing. In this study, we expressed a thermostable chimeric xylanase (Xyl22L) in *A. thaliana*, consisting of an extremely thermophilic catalytic domain and CBMs from an *E. grandis* xylanase, in order to increase the efficiency of lignocellulosic biomass hydrolysis. The rationale for the study was that the addition of the CBM22 repeats from a fast growing woody plant to an extremely thermophilic xylanase would lead to increased targeting and accumulation of the chimeric enzyme in the SCW. This would allow for the preloading of lignocellulosic biomass with inactive xylanase targeted to the SCW, which could later be activated by heat pre-treatment of the biomass. This was expected to further reduce the recalcitrance of lignocellulosic biomass to enzymic saccharification from what could be achieved with a xylanase alone. Instead, we found that the catalytic domain of the chimeric xylanase did not retain function whether expressed *in vitro* or *in planta*. Furthermore, *in planta* expression of the synthetic xylanase increased plant biomass and that the added CBM22 repeats from *E. grandis* likely allowed Xyl22L to target xylan molecules in the SCW. Additionally, it was hypothesised that *in planta* expressed enzyme may physically block access to biopolymers, thereby reducing the efficiency of hydrolysis before heat treatment. Together, these findings suggest that the addition of a mesophilic CBM domain can lead to targeting of synthetic enzymes to a desired region of SCWs, which could allow for more specific and efficient hydrolysis or modification of biopolymers. The unexpected increase in biomass caused by *in planta* expression of Xyl22L should also be further investigated, in order to determine the exact mechanism by which this takes place. Finally, the experiment should be repeated with a chimeric enzyme that retains xylanolytic function to determine if the increase in biomass can be maintained while recalcitrance to enzymic digestion is reduced. Applying these findings to current industrial pre-treatments could reduce cost and increase efficiency of processing of lignocellulosic biomass, increasing the viability of deriving a range of bio-based products from lignocellulosic feedstocks such as wood, providing an alternative to current petrochemical-derived products.

3.7 Acknowledgements

The author would to thank the following people for their contributions to the work: Dr Steven Hussey for advice on protein expression, Ms Caryn Oates for help with wax embedding and microscopy, and Dr Bilal Mir for providing seed for the transgenic JX_WT plant line.

3.8 References

Alvira P, Tomás-Pejó E, Ballesteros M, Negro M (2010) Pre-treatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: a review. Bioresource Technology 101:4851-4861

André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M (2014) CAZyme discovery and design for sweet dreams. Current Opinion in Chemical Biology 19:17-24

Araki R, Karita S, Tanaka A, Kimura T, Sakka K (2006) Effect of family 22 carbohydrate-binding module on the thermostability of Xyn10B catalytic module from *Clostridium stercorarium.* Bioscience Biotechnology and Biochemistry 70:3039

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. Bmj 310:170

Blumer-Schuette SE et al. (2014) Thermophilic lignocellulose deconstruction. FEMS Microbiology Reviews 38:393-448

Borkhardt B, Harholt J, Ulvskov P, Ahring BK, Jørgensen B, Brinch-Pedersen H (2010) Autohydrolysis of plant xylans by apoplastic expression of thermophilic bacterial endo-xylanases. Plant Biotechnology Journal 8:363-374

Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR (2007) Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. The Plant Journal 52:1154-1168

Busse-Wicher M, Grantham NJ, Lyczakowski JJ, Nikolovski N, Dupree P (2016) Xylan decoration patterns and the plant secondary cell wall molecular architecture. Biochemical Society Transactions 44:74-78

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Research 37:D233-D238

Castiglia D et al. (2016) High-level expression of thermostable cellulolytic enzymes in tobacco transplastomic plants and their use in hydrolysis of an industrially pretreated *Arundo donax L.* biomass. Biotechnology for Biofuels 9:1

Chen M-H, Kim SM, Raab RM, Li X, Singh V (2017) Heterologous expression of thermoregulated xylanases in switchgrass reduces the amount of exogenous enzyme required for saccharification. Biomass and Bioenergy 107:305-310

Clough SJ (2005) Floral dip: agrobacterium-mediated germ line transformation. Methods in Molecular Biology 286:91-102

Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. The Plant Journal 16:735-743

Cosgrove DJ (2005) Growth of the plant cell wall. Nature Reviews Molecular Cell Biology 6:850-861

Cosgrove DJ (2014) Re-constructing our models of cellulose and primary cell wall assembly. Current Opinion in Plant Biology 22:122-131

Cosgrove DJ, Jarvis MC (2012) Comparative structure and biomechanics of plant primary and secondary cell walls. Frontiers in Plant Science 3:204

Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes *in planta*. Plant Physiology 133:462-469

Delannoy M, Alves G, Vertommen D, Ma J, Boutry M, Navarre C (2008) Identification of peptidases in *Nicotiana tabacum* leaf intercellular fluid. Proteomics 8:2285-2298

Ding S-Y, Liu Y-S, Zeng Y, Himmel ME, Baker JO, Bayer EA (2012) How does plant cell wall nanoscale architecture correlate with enzymatic digestibility? Science 338:1055-1060

Doran PM (2006) Foreign protein degradation and instability in plants and plant tissue cultures. Trends in Biotechnology 24:426-432

Elleuche S (2015) Bringing functions together with fusion enzymes—from nature's inventions to biotechnological applications. Applied Microbiology and Biotechnology 99:1545-1556

Gallezot P (2012) Conversion of biomass to selected chemical products. Chemical Society Reviews 41:1538-1558

George RA, Heringa J (2002) An analysis of protein domain linkers: their classification and role in protein folding. Protein Engineering, Design and Selection 15:871-879

Gerday C, Glansdorff N (2007) Physiology and biochemistry of extremophiles. ASM Press

Guillén D, Sánchez S, Rodríguez-Sanoja R (2010) Carbohydrate-binding domains: multiplicity of biological roles. Applied Microbiology and Biotechnology 85:1241-1249

Hammond-Kosack KE, Harrison K, Jones J (1994) Developmentally regulated cell death on expression of the fungal avirulence gene Avr9 in tomato seedlings carrying the disease-resistance gene Cf-9. Proceedings of the National Academy of Sciences 91:10445-10449

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biology 8:R19

Hellwig S, Drossard J, Twyman RM, Fischer R (2004) Plant cell cultures for the production of recombinant proteins. Nature Biotechnology 22:1415

Hendriks A, Zeeman G (2009) Pre-treatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology 100:10-18

Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP (2010) Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. Proceedings of the National Academy of Sciences 107:15293-15298

Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315:804-807

Jönsson LJ, Martín C (2016) Pre-treatment of lignocellulose: formation of inhibitory by-products and strategies for minimizing their effects. Bioresource Technology 199:103-112

Kim JS, Daniel G (2012) Immunolocalization of hemicelluloses in *Arabidopsis thaliana* stem. Part I: temporal and spatial distribution of xylans. Planta 236:1275-1288

Kusnadi AR, Nikolov ZL, Howard JA (1997) Production of recombinant proteins in transgenic plants: practical considerations. Biotechnology and Bioengineering 56:473-484

Lee C, O'Neill MA, Tsumuraya Y, Darvill AG, Ye Z-H (2007) The *irregular xylem9* mutant is deficient in xylan xylosyltransferase activity. Plant and Cell Physiology 48:1624-1634

Leuschner C, Antranikian G (1995) Heat-stable enzymes from extremely thermophilic and hyperthermophilic microorganisms. World Journal of Microbiology and Biotechnology 11:95-114

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495

McCartney L, Marcus SE, Knox JP (2005) Monoclonal antibodies to plant cell wall xylans and arabinoxylans. Journal of Histochemistry & Cytochemistry 53:543-546

McKendry P (2002) Energy production from biomass (part 1): overview of biomass. Bioresource Technology 83:37-46

Meents MJ, Watanabe Y, Samuels AL (2018) The cell biology of secondary cell wall biosynthesis. Annals of Botany 121:1107-1125

Mellerowicz EJ, Baucher M, Sundberg B, Boerjan W (2001) Unravelling cell wall formation in the woody dicot stem. Plant Cell Walls. Springer, pp 239-274

Mellerowicz EJ, Sundberg B (2008) Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. Current Opinion in Plant Biology 11:293-300

Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. Analytical Chemistry 31:426-428

Mir BA, Mewalal R, Mizrachi E, Myburg AA, Cowan DA (2014) Recombinant hyperthermophilic enzyme expression in plants: a novel approach for lignocellulose digestion. Trends in Biotechnology 32:281-289

Mir BA, Myburg AA, Mizrachi E, Cowan DA (2017) In planta expression of hyperthermophilic enzymes as a strategy for accelerated lignocellulosic digestion. Scientific Reports 7:11462

Montalvo-Rodriguez R, Haseltine C, Huess-LaRossa K, Clemente T, Soto J, Staswick P, Blum P (2000) Autohydrolysis of plant polysaccharides using transgenic hyperthermophilic enzymes Biotechnology and bioengineering 70:151-159

Naik SN, Goud VV, Rout PK, Dalai AK (2010) Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews 14:578-597

Najmudin S, Pinheiro BA, Prates JA, Gilbert HJ, Romão MJ, Fontes CM (2010) Putting an N-terminal end to the *Clostridium thermocellum* xylanase Xyn10B story: Crystal structure of the CBM22-1–GH10 modules complexed with xylohexaose. Journal of Structural Biology 172:353-362

Nardi CF, Villarreal NM, Rossi FR, Martínez S, Martínez GA, Civello PM (2015) Overexpression of the carbohydrate binding module of strawberry expansin2 in *Arabidopsis thaliana* modifies plant growth and cell wall metabolism. Plant Molecular Biology 88:101-117

Pawar PMA et al. (2016) Expression of fungal acetyl xylan esterase in *Arabidopsis thaliana* improves saccharification of stem lignocellulose. Plant Biotechnology Journal 14:387-397

Peña MJ et al. (2007) Arabidopsis irregular xylem8 and irregular xylem9: implications for the complexity of glucuronoxylan biosynthesis. The Plant Cell Online 19:549-563

Perlack RD, Wright LL, Turhollow AF, Graham RL, Stokes BJ, Erbach DC (2005) Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply. Oak Ridge National Laboratory DTIC Document

Perneger TV (1998) What's wrong with Bonferroni adjustments. Bmj 316:1236-1238

Phitsuwan P, Sakka K, Ratanakhanokchai K (2013) Improvement of lignocellulosic biomass *in planta*: A review of feedstocks, biomass recalcitrance, and strategic manipulation of ideal plants designed for ethanol production and processability. Biomass and Bioenergy 58:390-405

Pillay P, Schlüter U, Van Wyk S, Kunert KJ, Vorster BJ (2014) Proteolysis of recombinant proteins in bioengineered plant cells. Bioengineered 5:15-20

Plomion C, Leprovost G, Stokes A (2001) Wood formation in trees. Plant physiology 127:1513-1523

Ratke C et al. (2018) Downregulating aspen xylan biosynthetic GT 43 genes in developing wood stimulates growth via reprograming of the transcriptome. New Phytologist 219:230-245

Rennie EA, Scheller HV (2014) Xylan biosynthesis. Current Opinion in Biotechnology 26:100-107

Roberts K, Shirsat A (2006) Increased extensin levels in Arabidopsis affect inflorescence stem thickening and height. Journal of Experimental Botany 57:537-545

Rothschild LJ, Mancinelli RL (2001) Life in extreme environments. Nature 409:1092-1101

Safra-Dassa L, Shani Z, Danin A, Roiz L, Shoseyov O, Wolf S (2006) Growth modulation of transgenic potato plants by heterologous expression of bacterial carbohydrate-binding module. Molecular Breeding 17:355-364

Sainz-Polo MA, González B, Menéndez M, Pastor FJ, Sanz-Aparicio J (2015) Exploring multimodularity in plant cell wall deconstruction: structural and functional analysis of Xyn10C containing the CBM22-1-CBM22-2 tandem. Journal of Biological Chemistry jbc-M115

Scheller HV, Ulvskov P (2010) Hemicelluloses. Annual Review of Plant Biology 61

Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative C T method. Nature Protocols 3:1101

Shoseyov O, Shani Z, Levy I (2006) Carbohydrate binding modules: biochemical properties and novel applications. Microbiology and Molecular Biology Reviews 70:283-295

Suzuki M, Kato A, Nagata N, Komeda Y (2002) A xylanase, AtXyn1, is predominantly expressed in vascular bundles, and four putative xylanase genes were identified in the *Arabidopsis thaliana* genome. Plant and Cell Physiology 43:759-767

Tsai AYL, Canam T, Gorzsás A, Mellerowicz EJ, Campbell MM, Master ER (2012) Constitutive expression of a fungal glucuronoyl esterase in *Arabidopsis* reveals altered cell wall composition and structure. Plant Biotechnology Journal 10:1077-1087

Venditto I et al. (2015) Family 46 Carbohydrate-binding modules contribute to the enzymatic hydrolysis of xyloglucan and β-1,3–1,4-glucans through distinct mechanisms. Journal of Biological Chemistry 290:10572-10586

Verma D, Kawarabayasi Y, Miyazaki K, Satyanarayana T (2013) Cloning, expression and characteristics of a novel alkalistable and thermostable xylanase encoding gene (Mxyl) retrieved from compost-soil metagenome PLoS One 8:e52459

Yang B, Wyman CE (2004) Effect of xylan and lignin removal by batch and flowthrough pre-treatment on the enzymatic digestibility of corn stover cellulose. Biotechnology and Bioengineering 86:88-98

Zeller R (1999) Fixation, embedding, and sectioning of tissues, embryos, and single cells. Current Protocols in Pharmacology 7:11-14

Ziegler MT, Thomas SR, Danna KJ (2000) Accumulation of a thermostable endo-1,4-β-D-glucanase in the apoplast of *Arabidopsis thaliana* leaves. Molecular Breeding 6:37-46

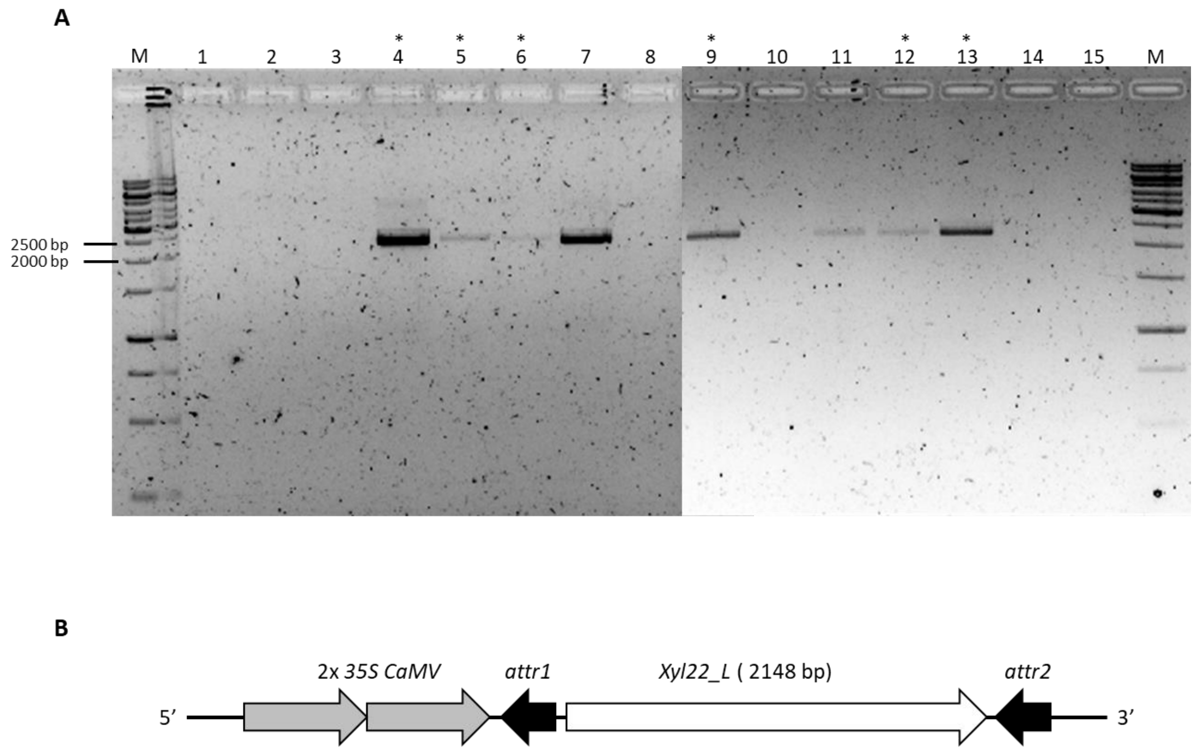
3.8 Supplementary Tables and Figures

Supplementary Table 3.1 Primers used in this study

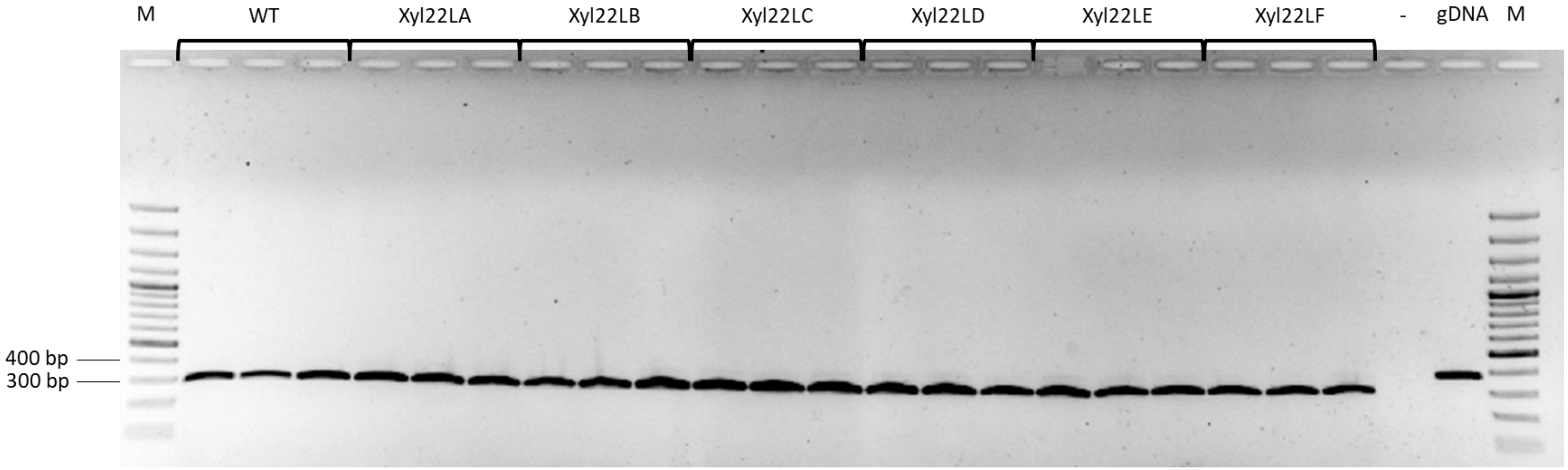
Primer name	Primer Sequence	T_m	Application
M13F-pUC	GTTTTCCCAGTCACGAC	58°C	Screening for insert orientation, sequencing of Gateway™ vectors (Supplementary File 3.1)
M13R-pUC	CAGGAAACAGCTATGAC	58°C	
Xyl22L_F	ATGGGATTTGTTCTTTTCTC	62°C	Amplification of the full length *Xyl22_L* CDS (Supplementary Figure 3.1, Supplementary File 3.1)
Xyl22L_R	TCAAGGAGTTGGACCCTGAA	62°C	
Xyl22L_F_int_seq	GAGAAGGACTGGAGGTACAA	58°C	Internal sequencing primers for *Xyl22*_L CDS (Supplementary File 3.1)
Xyl22L_R_int_seq	CAACACCGATGTACTGTTCC	58°C	
Act2_F	TGGAATCCACGAGACAACCT	62°C	Used for screening of cDNA for gDNA contamination (Supplementary Figure 3.2) and to amplify and quantify a reference gene in qPCR analysis (Figure 3.2)
Act2_R	TGGACCTGCCTCATCATACT	62°C	
Ubq5_F	GGTGGTGCTAAGAAGAGGAA	60°C	Amplification/quantification of *AtUbq5* reference gene in qPCR analysis (Figure 3.2)
Ubq5_R	TCGATCTACCGCTACAACAG	60°C	
Xyl22_qPCR_F	TTCGTGTCTGCTACTGAGAG	60°C	Amplification/quantification of *Xyl22L* in qPCR analysis (Figure 3.2)
Xyl22_qPCR_R	CAACACCGATGTACTGTTCC	60°C	

Supplementary Table 3.2 The butanol dehydration series used in sample preparation for confocal fluorescent microscopy.

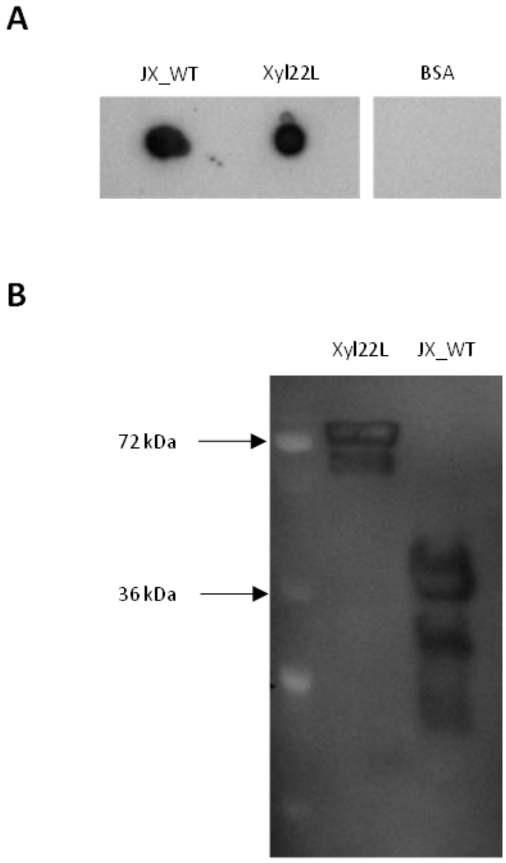
Step	Butanol (cm^3)	100% EtOH (cm^3)	dH_2O (cm^3)	Time (hrs)
A	50	60	90	1
B	80	60	60	1
C	110	50	40	1
D	140	40	20	2
E	170	30	0	2
F	200	0	0	Overnight



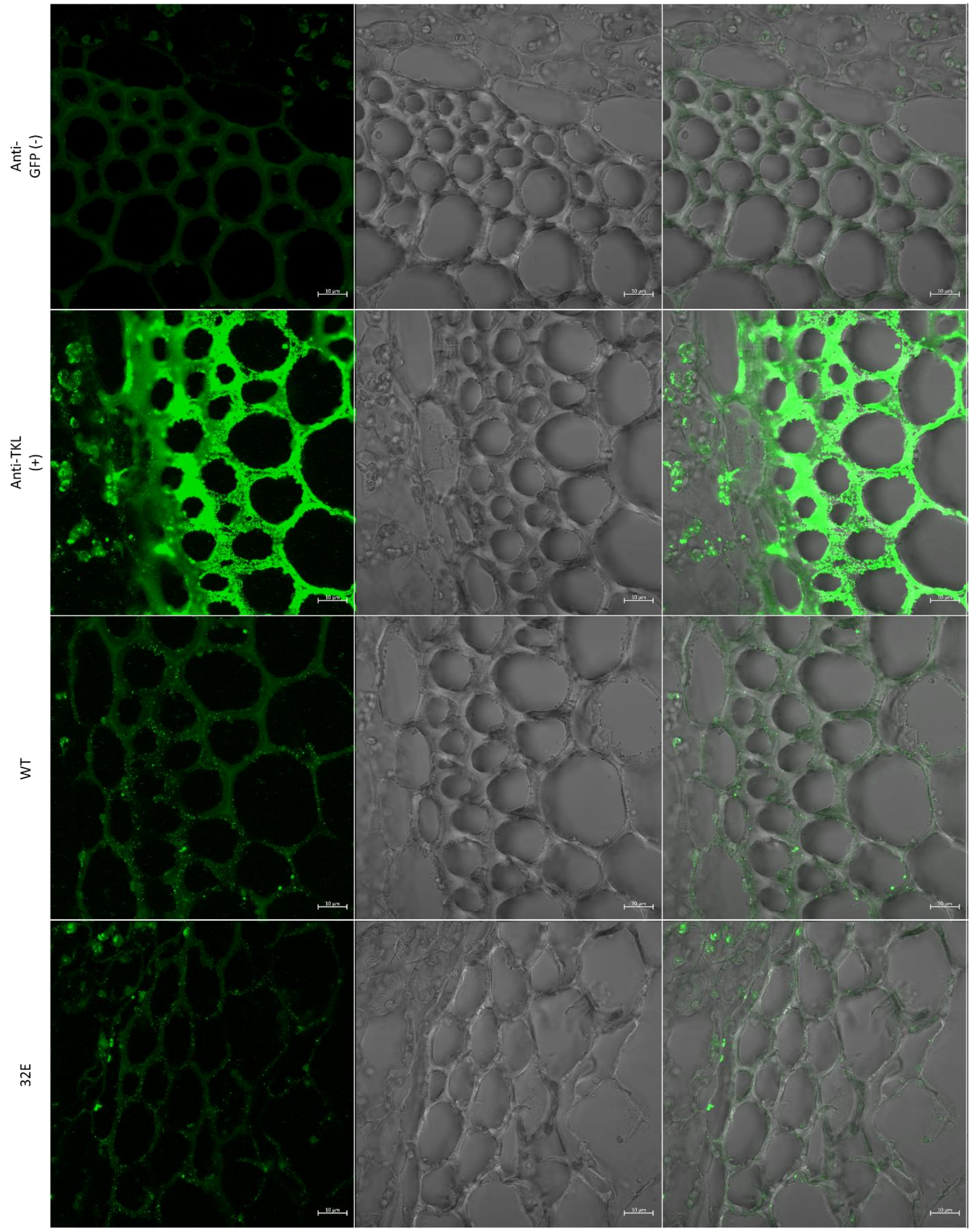
Supplementary Figure 3.1 Validation of expression cassette in T1 transgenic *Arabidopsis* lines. A: PCR amplification of the *Xyl22_L* CDS from T1 transgenic *Arabidopsis* lines. M indicates the Generuler 1 kb Fermentas Molecular Marker, with the relevant band sizes highlighted to the left of the figure. Successful amplification of *Xyl22L* is expected to result in band 2148 bp in size. The numbers above the figure indicate separate transgenic events as follows: 1: 32L_1, 2: 32L_2, 3: 32L_3, 4: 32L_4, 5: 32L_5, 6: 32L_6, 7: 32L_7, 8: 32L_8, 9: 32L_1_9, 10: 32L_10, 11: 32L_11, 12: 32L_3_2, 13: 32L_4_2, 14: 32L_5_2 and 15: Wild-type (negative control). Asterisks indicate transformation events for which homozygous plant lines were obtained. B: Schematic representation of the expression cassette *in pMDC32* used for transformation. Each features/CDS is represented by arrows, with the direction of the arrow indicating direction of transcription. The name of the feature/CDS is located above each arrow.

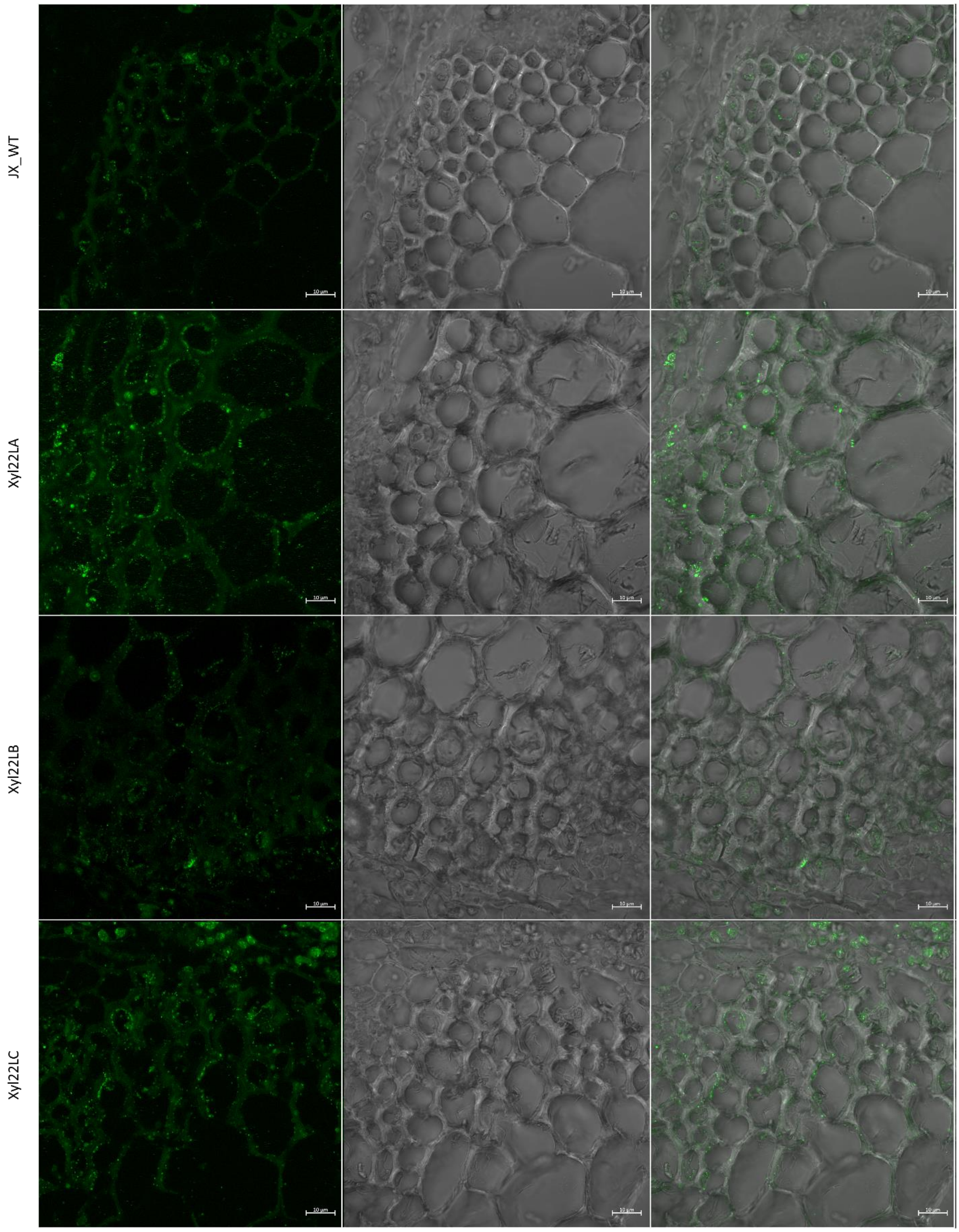


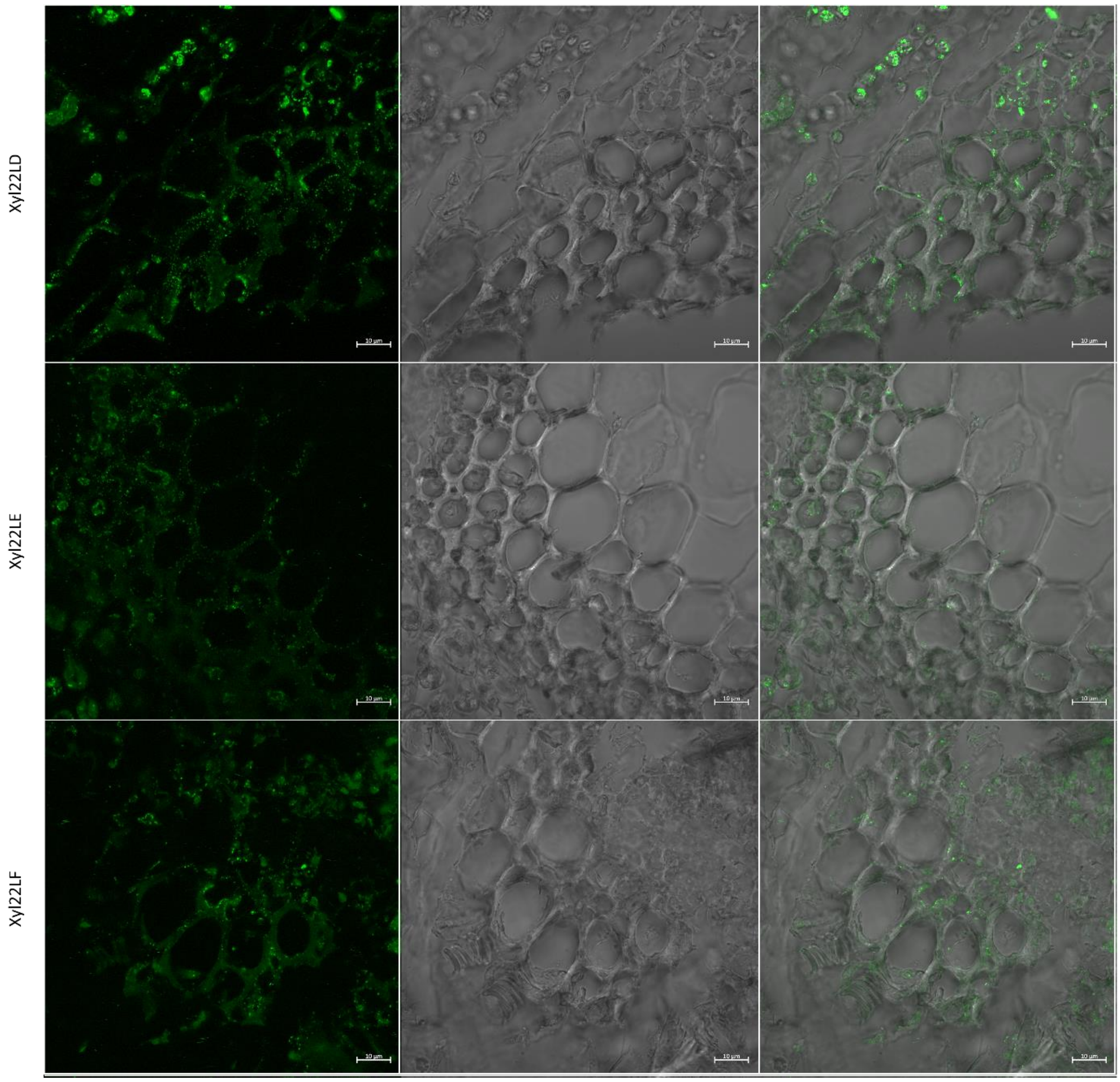
Supplementary Figure 3.2 Testing of cDNA for gDNA contamination. The plant line is indicated above the wells, all three replicates for each plant line are shown**.** M: Fermentas 100 bp molecular marker. -: No template control. gDNA: gDNA template control. PCR was performed using intron-spanning *Act2* primers (Supplementary table 1). A band of approximately 300 bp indicates no gDNA contamination. A band of approximately 400 bp indicates the presence of gDNA.



Supplementary Figure 3.3 Testing of primary antibody against synthesised JX_WT (32 kDa) and Xyl22L (77 kDa). A: Dot blot showing reactivity of primary antibody with JX_WT and Xyl22L that has been expressed in and purified from *E. coli*. BSA is used as a negative control. B: Western blot using primary antibody showing the size range of Xyl22L and JX_WT that has been expressed in and purified from *E. coli*. Smaller bands indicate degraded protein.


Anti-GFP (-)
Anti-TKL (+)
WT
32E
10 µm


JX_WT
Xyl22LA
Xyl22LB
Xyl22LC
10 μm



Supplementary Figure 3.4 Fluorescent confocal microscopy of transgenic *A. thaliana* stem cross sections. The plant line is indicated on the left of each image. The channel is indicated on the top of each image. The labels are as follows: Signal – The fluorescent signal detected in the sample. Light – An image obtained using white light. Merged – A composite picture of the merged Signal and Light channels. Primary antibodies raised against green fluorescent protein (GFP) were used as the negative control. Primary antibodies raised against *A. thaliana* transketolase (TKL) were used as a positive control. Scale bars indicate 10 µm.

CHAPTER 4:

Concluding Remarks

4.1 Summary of findings

Lignocellulosic biomass is recalcitrant to enzymic digestion (Himmel et al. 2007), which is a significant barrier to the adoption of an economically and environmentally sustainable strategy for the synthesis of biomaterials. Lignocellulosic biomass normally requires extreme industrial pre-treatments in order to access and utilise the constituent biopolymers. Industrial pre-treatments are economically and energetically expensive, and can result in degradation products that inhibit downstream processes (Alvira et al. 2010; Hassan et al. 2018). One promising strategy to overcome this issue is to heterologously express Carbohydrate Active enZymes (CAZymes) *in planta*, thereby promoting autohydrolysis, and reducing the need for external enzyme loading and additional pre-treatments (Mir et al. 2014). This strategy may be further improved by using enzymes from extremely thermophilic organisms, as they are not typically active at mesophilic temperatures, allowing for accumulation in the biomass without adversely affecting growth and development of the plant (Mir et al. 2014; Mir et al. 2017). However, while deconstruction of lignocellulose by thermophilic enzymes has been examined (Blumer-Schuette et al. 2014), the full extent to which extremely thermophilic organisms can degrade lignocellulosic biomass is unknown. Additionally, the ability to specifically target enzymes from extremely thermophilic organisms to biopolymers by combining them with mesophilic plant derived protein domains has not been assessed.

In this thesis, a survey of protein domains from the proteomes of extremely thermophilic organisms was provided, and the capacity for lignocellulose degradation within these domains was investigated. Additionally, a chimeric enzyme consisting of a GH11 domain from an extremely thermophilic metagenomic library, and CBM22 domains derived from *Eucalyptus grandis* was designed and synthesised. The enzyme was heterologously expressed in *Arabidopsis thaliana* and the effect on growth and development of the plant, as well as on the recalcitrance of the biomass to enzymic degradation was investigated.

In Chapter 1 (Botha et al. 2017), we provided an introduction to synthetic biology and protein domains, as well as a list of putative CAZyme domains identified from extremely thermophilic proteomes. The domains were described as a "toolbox" that could be used to assemble recombinant enzymes for a variety of tasks. The discovery of new domains for the toolbox was addressed, as was how current domains may be further modified in order to make them more suitable for a given application. Finally, the use of the toolbox for the degradation of lignocellulosic biomass was addressed.

Chapter 2 provided an update and closer examination of the dataset produced in Chapter 1. The diversity and abundance of CAZyme domains in extremely thermophilic proteomes was examined, as well as the capacity for lignocellulose degradation within the set of domains. Whether or not new extremely thermophilic CAZyme domains would be identified as more genomes were sequenced was also addressed. Significant differences were identified in CAZyme domain diversity and abundance between archaea and bacteria, mainly relating to structural differences such as cell wall composition, and nutritional strategy. Putative lignocellulose degrading CAZyme domains were prominent in the dataset, but found mainly in bacteria, though some unique lignocellulose degrading CAZyme domains, such as GH116 (Ferrara et al. 2014) were identified in archaea. Finally, it was found that as more genomes of extremely thermophilic organisms are sequenced, novel variants of currently known CAZyme domains, as well as domains from currently unrepresented CAZyme classes may be identified.

In Chapter 3, a chimeric enzyme (Xyl22L) was designed and synthesised that consisted of a thermostable xylan-degrading GH11 domain (Verma et al. 2013) and xylan-targeting CBM22 repeats derived from *E. grandis.* The ability of the chimeric enzyme to degrade xylan was investigated. The enzyme was also heterologously expressed in *A. thaliana*, and the effect on growth, development and digestibility of the plants was determined. Additionally, the ability of Xyl22L to adhere to the secondary cell wall (SCW) was assessed. Xyl22L was not able to hydrolyse xylan, indicating that addition of the CBM22 repeats had abolished the xylanase activity of the GH11 domain. As expected, heterologous expression of Xyl22L had no negative effect on the growth and development of the plant, though transgenic plant lines showed an increase in biomass relative to wild-type plants. Xyl22L accumulated in transgenic plant biomass, and was able to adhere to the SCW, indicating that the CBM22 repeats were functional. Finally, before heat treatment, transgenic plants showed an increase in recalcitrance to enzymic digestion compared to wild-type. After heat-treatment, there was no difference. This indicates that before heat treatment, Xyl22L was coating the SCW and preventing access by other hydrolytic enzymes. After treatment the enzyme was denatured, allowing access to the biopolymers. Together, this showed that even though hydrolase activity in Xyl22L was lost, the CBM22 repeats were still functional and able to target Xyl22L to the SCW.

4.2 Contributions of the thesis to current knowledge

In Chapter 2, it was found that there is a diverse range of CAZyme domains present in extremely thermophilic organisms. While some domains were relatively common, others are unique to phyla, genera or species. The differences in composition are mostly due to the different lifestyles and ecological niches that these organisms inhabit. It was also found that extremely thermophilic organisms contain significant capacity for the degradation of lignocellulosic biomass. The pool of CAZyme domains identified in this study could potentially be used to construct custom synthetic enzymes that are tailored to needs of a particular situation. For example, by combining a xyloglucan targeting CBM with a GH domain that degrades xylan, it may be possible to target and degrade xylan specifically associated with xyloglucan in lignocellulosic biomass. Relatively few genomes of extremely thermophilic organisms have been sequenced, and the potency of this strategy will increase as novel domains and domain variants are discovered. Finally, many thermostable enzymes have been investigated in the past (Blumer-Schuette et al. 2014; Mir et al. 2014), but relatively little attention has been given to determining the individual CAZyme domain content in extremely thermophilic organisms. The chapter provides a survey of putative domains from these organisms, and by extension provides insight into the strategies and mechanisms by which extremely thermophilic organisms survive and evolve.

In Chapter 3, it was found that heterologously expressing an enzyme *in planta* can affect the growth and development of the plant, as well as the recalcitrance of the biomass to enzymic digestion. This is especially significant considering that the catalytic domain of Xyl22L seemed to be non-functional. Therefore, any changes in growth, development and recalcitrance may be attributed to the action of the CBM22 repeats contained within Xyl22L, and not to digestion of xylan in the SCW. Xyl22L was also able to accumulate in the biomass and associate with the cell wall, and plant lines expressing Xyl22L were more recalcitrant to sugar release before heat treatments but not after, indicating that Xyl22L was bound to the SCW and preventing access of hydrolytic enzymes. This work is proof of concept that CBMs may be used to modify enzymes and help them target specific polymers or locations in lignocellulosic biomass. Additionally, the differences in recalcitrance of biomass before and after heat treatments shows that the enzyme itself may prevent access by other hydrolases. These findings are important to many white (industrial) biotechnological applications. Autohydrolysis with enzymes designed to target specific biopolymers would allow for a reduction in the cost and energy required to process lignocellulosic feedstocks as well as a decrease in formation of degradation products. While producing self-hydrolysing biomass containing thermostable enzymes (Mir et al. 2014) remains an attractive option for the mitigation of recalcitrance to enzymic digestion, the work in this chapter shows that these strategies may be improved through synthesising custom enzymes that are more efficiently targeted to a substrate. This work is also one of the first reports of the expression of a synthesised chimeric enzyme *in planta* for the purpose of autohydrolysis, and, therefore, provides valuable insight and knowledge for the improvement and application of this strategy.

4.2 Limitations of the thesis

There are a number of factors that should be kept in mind when engaging with the work in Chapter 2. The first, and most important, is that most of these domains have not been experimentally validated. CAZyme domains are classified based on similarity to pre-existing crystal structures of known proteins and the assumption is made that all CAZyme domains within a particular class will behave more or less consistently, but this is often not the case (Cantarel et al. 2009; Lombard et al. 2014). Similarly, not all domains defined in the chapter may be thermostable. Therefore, the only way to be absolutely certain of domain function is through experimental characterisation, in both *in vitro* and *in vivo* contexts. Regardless, the list of CAZyme domains in the chapter serves as a good starting point, and proteins designed using this resource may be improved and fine-tuned through rational design, as well as directed evolution strategies (Botha et al. 2017; Turner 2009). The second factor is the method by which the domains were identified. HMM-based scans were performed using the HMMER package (Eddy 1998; Finn et al. 2011). Like any package, HMMER has biases and drawbacks and so some false positives probably feature in the dataset. Additionally, some domains may have been missed (false negatives). Experimental validation would help to identify false positives, and some false negatives could be identified by performing a scan with relaxed stringency, though at the cost of increasing false positives. The third factor to consider is how to design synthetic enzymes using the domains identified in the chapter. Many other elements play a role in proper protein function, such as linker sequences between domains, internal bonds and bridges, as well as shared secondary and tertiary structures, among others. Any of these factors may prevent a synthetic protein from functioning properly, but was beyond the scope of the study. The last factor to consider relates to cellulosomes. Cellulosomes are important machinery for the deconstruction of lignocellulose in microorganisms (Artzi et al. 2017). Part of the dataset suggested the presence of cellulosomes, but they were not discussed at length in the chapter.

While there were interesting findings from the work performed in Chapter 3, there were some issues that hindered the project. The first and most obvious of these is that Xyl22L had no catalytic function, possibly due to the substitution of the native CBM60 domain in JX_WT with the CBM22 repeats, making further characterisation (optimum pH, temperature, specific activities and thermostability) of the enzyme redundant. This may also have affected protein stability, as well as function of the CBM22 repeats, and the extent of this effect is difficult to quantify without an appropriate comparison. These kinds of problems may be circumvented in future work through more rigorous protein design (André et al. 2014). Close interrogation of the literature surrounding selected protein domains and in depth *in silico* analysis of protein sequences will allow for the identification of important structural features (e.g. linker sequences, disulphide bridges, shared structures etc.), and is more likely to result in a functional protein. Additionally, the domains used should be experimentally characterised, or at least derived from an experimentally characterised protein. The second issue is that the antibody used to label Xyl22L showed cross reactivity with other plant proteins. While this was not an issue for western blots, as there was no cross reactivity with proteins in the expected size range of Xyl22L, this may have reduced sensitivity of both western blots as well as the fluorescent confocal microscopy performed in the chapter. This issue may be remedied by redesigning antibodies until an appropriately specific antibody is obtained. The third issue is that the cause of the increased biomass in transgenic plant lines is not well understood. A strong phenotype was not expected from the transgenic plants described in the chapter, and so the growth experiments were not designed to capture this kind of data. Experiments designed to investigate the structure and composition of the SCW, as well the morphology of the plants may help to explain the phenotype. Additionally, more transgenic plant lines with more diverse levels of *Xyl22L* expression would help to increase the statistical rigor of these experiments. The fourth issue is that it is not clear whether Xyl22L can bind to xylan. *In vitro* xylan binding assays did not provide an answer, due to difficulties associated with native PAGE (Wittig and Schägger 2005; Wittig and Schägger 2008), and in this case confocal fluorescent microscopy gave no indication of whether the protein is associating directly with the cell wall or the cell membrane. Finally, the plant line expressing JX_WT did not behave as expected. Plants expressing JX_WT were previously characterised (Mir et al. 2017) and one of these plant lines were used as a comparison in the chapter. JX_WT plants previously showed increased sugar release from the transgenic biomass, and total soluble protein (TSP) extracts from JX_WT plants were able to hydrolyse beechwood xylan. This was not seen in the data presented in Chapter 3, even though experimental controls behaved as expected. The differences may be ascribed to differing experimental approaches, but without fully recharacterising the plant line in the chapter, it is difficult to determine whether there is an issue with the plant lines or an error in the experiments.

Despite the above-mentioned issues, this work not only facilitates enzyme engineering strategies for more efficient lignocellulose deconstruction, but also provides an example of such an enzyme, and the effect that heterologous expression can have on lignocellulosic biomass. While the enzyme was ultimately unable to degrade xylan, this work showed that it is possible to target an enzyme to the vicinity of the cell wall through the addition of plant derived CBMs, allowing for more specific enzyme action. Applying these findings to current industrial processes may lead to a decrease in cost and increase in throughput of the processing of lignocellulose, allowing for a more economically viable and environmentally sustainable bioproduct industry.

4.3 Future work

The work performed in this thesis provides a good starting point for the design of chimeric enzymes, but further research would be beneficial. For a start, expressing and characterising some of the domains identified in Chapter 2 *in vitro* would help to legitimise the toolbox of extremely thermophilic protein domains for synthetic biology. Some of these domains could then in turn be used to synthesise chimeric enzymes for the degradation or modification of lignocellulosic biomass. For chapter 3, repeating the study with a newly designed protein would help to address some of the previously unanswered questions. Only adding a single CBM22 domain, or using a previously characterised plant CBM know to bind to xylan, such as those found in AtXyn1 (Suzuki et al. 2002), may allow for the catalytic domain to remain functional. Additionally, using a catalytic domain that naturally occurs alone in enzymes will increase the likelihood that it will function regardless of which domain it is paired with. Additionally, it would be interesting to target various well-characterised catalytic domains to specific locations and side chains of SCW biopolymers through fusion with appropriate CBMs. These recombinant enzymes could be characterised and expressed *in planta*, and their effect on the plant biomass could be determined. Fully characterising the CBM22 repeats used in Chapter 3 may also be beneficial. Determining how they interact with biopolymers in the SCW and comparing them to orthologs from *A. thaliana* would not only potentially provide a tool for targeting tree hemicellulose, but also provide insight into the differences in SCW between herbaceous plants and trees. Concerning the transgenics produced in Chapter 3, it would be interesting to investigate the cause of the increase biomass more closely. Through microscopy, phenotyping, growth trials and wood chemical analysis, it may be possible to determine the effect that the CBM22 repeats have on the structure and composition of the SCW, or if the increase in biomass is due to a morphological change, such as increased leaves or inflorescence stems. Finally, once an enzyme that can reduce the recalcitrance of lignocellulosic biomass to enzymic digestion has been successfully produced, it would be interesting to express it in a tree species, such as poplar or *Eucalyptus*, and assess its effect on growth and development of the biomass, as well as the recalcitrance of the biomass to digestion.

4.4 References

Alvira P, Tomás-Pejó E, Ballesteros M, Negro M (2010) Pre-treatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: a review. Bioresource Technology 101:4851-4861

André I, Potocki-Véronèse G, Barbe S, Moulis C, Remaud-Siméon M (2014) CAZyme discovery and design for sweet dreams. Current Opinion in Chemical Biology 19:17-24

Artzi L, Bayer EA, Moraïs S (2017) Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nature Reviews Microbiology 15:83-95

Blumer-Schuette SE et al. (2014) Thermophilic lignocellulose deconstruction. FEMS Microbiology Reviews 38:393-448

Botha J, Mizrachi E, Myburg AA, Cowan DA (2017) Carbohydrate active enzyme domains from extreme thermophiles: components of a modular toolbox for lignocellulose degradation. Extremophiles:1-12

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Research 37:D233-D238

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755-763

Ferrara MC, Cobucci-Ponzano B, Carpentieri A, Henrissat B, Rossi M, Amoresano A, Moracci M (2014) The identification and molecular characterization of the first archaeal bifunctional exo-β-glucosidase/N-acetyl-β-glucosaminidase demonstrate that family GH116 is made of three functionally distinct subfamilies. Biochimica et Biophysica Acta - General Subjects 1840:367-377

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39:W29-W37

Hassan SS, Williams GA, Jaiswal AK (2018) Emerging technologies for the pre-treatment of lignocellulosic biomass. Bioresource Technology

Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315:804-807

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42:D490-D495

Mir BA, Mewalal R, Mizrachi E, Myburg AA, Cowan DA (2014) Recombinant hyperthermophilic enzyme expression in plants: a novel approach for lignocellulose digestion. Trends in Biotechnology 32:281-289

Mir BA, Myburg AA, Mizrachi E, Cowan DA (2017) In planta expression of hyperthermophilic enzymes as a strategy for accelerated lignocellulosic digestion. Scientific Reports 7:11462

Suzuki M, Kato A, Nagata N, Komeda Y (2002) A xylanase, AtXyn1, is predominantly expressed in vascular bundles, and four putative xylanase genes were identified in the *Arabidopsis thaliana* genome. Plant and Cell Physiology 43:759-767

Turner NJ (2009) Directed evolution drives the next generation of biocatalysts. Nature Chemical Biology 5:567-573

Verma D, Kawarabayasi Y, Miyazaki K, Satyanarayana T (2013) Cloning, expression and characteristics of a novel alkalistable and thermostable xylanase encoding gene (Mxyl) retrieved from compost-soil metagenome. PLoS One 8:e52459

Wittig I, Schägger H (2005) Advantages and limitations of clear-native PAGE. Proteomics 5:4338-4346

Wittig I, Schägger H (2008) Features and applications of blue-native and clear-native electrophoresis. Proteomics 8:3974-3990