Genome-wide analysis of carbohydrate active enzyme diversity and expression in *Eucalyptus grandis*

by

Desré Pinard

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences
Department of Genetics
University of Pretoria
Pretoria
December 2013

Under the supervision of Prof. Alexander A. Myburg

and co-supervision of Prof. Fourie Joubert and Dr. Eshchar Mizrachi

Declaration

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

_____

Desré Simone Pinard

December 2013

Thesis Summary

Genome-wide analysis of carbohydrate active enzyme diversity and expression in *Eucalyptus grandis*

Desré Pinard

Supervised by **Prof. A.A. Myburg**

Co-supervised by **Prof. Fourie Joubert** and **Dr. Eshchar Mizrachi**

Submitted in partial fulfillment of the requirements for the degree ***Magister Scientiae***

Department of Genetics

University of Pretoria

The woody biomass derived from tree species forms a vital part of the world's economy, and a thorough understanding of the processes of carbon sequestration and carbohydrate metabolism in trees is paramount in ensuring efficient and sustainable use of this biomass. To date, there is still much to be learnt about wood formation and polysaccharide deposition in plant cell walls. The enzymes responsible for the synthesis, degradation, and modification of polysaccharides and glycosidic bonds are known as Carbohydrate Active enZymes

(CAZymes) are organized into functional classes and families based on amino acid sequence. CAZymes in plant genomes can be analyzed using the functional protein domains that form the proteins in order to better understand the functional potential of the carbohydrate metabolism strategy employed by plants. The glycosyltransferase (GT) class of CAZyme domains is responsible for the synthesis of glycosidic bonds, and the glycosylhydrolases (GH), polysaccharide lyase (PL), and carbohydrate esterase (CE) CAZyme domain classes degrade and modify these bonds. The final class of CAZyme domains is the non-enzymatic carbohydrate binding modules (CBMs), which act to increase the activity of the enzymatic CAZyme domain classes via specific binding to polysaccharides, disruption of the cell wall polysaccharide matrix, and proximity effects when appended to enzymatic CAZyme domains in complex CAZyme domain containing proteins.

In this project, we used comparative genomics and transcriptomics of CAZyme domains to analyze the functional building blocks of plant carbohydrate metabolism to gain insight into the process of wood formation, with a specific focus on the biomass feedstock crop, *Eucalyptus grandis*. The aim of this project was to compare the CAZyme domain frequency, diversity and complexity across plant genomes representative of the major land plant lineages and green algae species to identify any delineating factors that contribute to wood formation in tree species. In addition, we analyzed the expression levels of CAZyme domains in the transcriptomes of source and sink tissues in *E. grandis* and *Populus trichocarpa* to better understand the expression investment in carbohydrate metabolism in different tissues of divergent tree species.

The results show conservation of a fundamental functional strategy for carbohydrate metabolism across land plant evolution. The ratio of CAZyme domain frequency is maintained in land plants, with GTs contributing ≈40% of the genomic CAZyme domain content, highlighting the importance of polysaccharide synthesis in plants. The diversity of CAZyme domain families within each class cannot be used to differentiate the genomes of

major land plant lineages (lycophytes and bryophytes, monocots, and dicots) from one another, however, species-specific differences in CAZyme domain family diversity are observed. The complexity of CAZyme domain containing proteins shows that CAZyme domains are not very promiscuous, repeated CAZyme domains within a protein are more common than unique combinations of CAZyme domains within a protein, which are also conserved for the most part. The analysis of CAZyme domain expression in six tissues in *E. grandis* showed that in the wood forming tissue, immature xylem, GT domain families responsible for cellulose and hemicellulose biosynthesis formed the majority of the transcript abundance, a pattern not seen in the other tissues analyzed. This pattern was conserved in *P. trichocarpa*, highlighting the conserved mechanism for wood formation between divergent tree species.

The results of this study reveal the conservation of the fundamental functional machinery responsible for carbohydrate metabolism in land plants, and highlight the importance of differential regulation of this machinery to wood formation. The long-term goal of improving the production of lignocellulosic biomass from trees will be achieved by fully understanding the regulatory mechanisms controlling the concerted expression of these CAZyme domain-containing genes.

Preface

A major focus of recent times has been the need to maximize the utilization of natural resources in an environmentally friendly and efficient manner. In this regard, research into the production of lignocellulosic biomass from woody crop species has become a priority. Lignocellulosic biomass is an important natural resource, used in the paper and pulp industry among others, the manufacturing of high-value chemical cellulose derivatives, along with the potential to efficiently produce bioethanol. The secondary cell walls of the xylem tissue of trees are the major source of the biopolymers that are important in industry: the cellulosic biopolymers, namely cellulose and hemicelluloses. The secondary cell walls are comprised of cellulose microfibrils embedded in a matrix of hemicelluloses, pectin, lignin, and cell wall proteins. The thick secondary cell wall has a higher percentage of cellulose and hemicelluloses compared to the primary cell wall. Despite the large amount of research performed in the field, there are still many unknowns with regards to secondary cell wall/ wood formation.

The enzymes responsible for the synthesis, modification and degradation of polysaccharides, including those in the cell wall, are collectively known as Carbohydrate Active enZymes (CAZymes). CAZyme protein domains are classified into five functional classes based on their activity. Glycosyltransferases (GT) synthesize glycosidic bonds, which are broken via hydrolysis by glycosyl/glycoside hydrolases (GH). Glycosidic bonds are further modified and degraded non-hydrolytically by carbohydrate esterases (CEs), which break the bonds of acetyl esters, and polysaccharide lyases, which cleave glycosidic bonds using an elimination mechanism. The final class, carbohydrate binding modules (CBMs), are non-enzymatic domains that commonly co-occur with enzymatic domains, and function by binding specific carbohydrate biopolymers and increasing the activity of enzymatic domains. Previous studies have shown a relationship between the frequency and diversity of genomic and expressed CAZymes, and the carbohydrate metabolic lifestyle of bacterial and plant

species. To date, no large-scale comparative analysis of CAZymes across plant species has been performed, thus our understanding of the contribution of CAZymes to the woody habit can be improved using comparative *in silico* analyses.

The aim of this MSc was to characterize the genomic CAZyme frequency and diversity in twenty-two sequenced plant species, and to compare the expressed catalogue of CAZymes across six tissues in *Eucalyptus grandis* along with two tissues in *Populus trichocarpa*. By combining a comprehensive comparative genomic analysis of plant CAZymes with comparisons of expressed CAZymes in the woody and non-woody tissues of two valuable crop species, we can gain insight into the metabolism of carbohydrates in wood forming tissues.

Chapter 1: This chapter is a comprehensive review of protein domains, carbohydrate active enzymes and how they are related to secondary cell wall biosynthesis. The dynamics of protein domain evolution are reviewed, along with the methods of identifying protein domains across genomes. In order to fully appreciate the mechanisms by which CAZymes contribute to wood formation, the functions of CAZymes are discussed as they are responsible for cell wall polysaccharide biosynthesis, in addition they are involved in plant signaling, defense and storage polysaccharide biosynthesis. The biopolymers present in the cell wall are considered, as they are the main targets of CAZymes of interest to this study. Each class of CAZymes is discussed with known examples of the CAZyme domain families in each class, specifically those that have been shown to be involved in cell wall biosynthesis. The studies that have analyzed genome-wide CAZyme content in plants are reviewed, along with the importance of woody biomass, and specifically *E. grandis* as a crop species.

Chapter 2: The results of comparative analysis of CAZyme domain frequency and diversity in the genomes of twenty-two plant species CAZyme are presented. The plant species analyzed represent the major lineages of plant evolution, including eudicots, monocots,

lycophytes and bryophytes, and green algae. In addition to a comparison of the expressed CAZyme investment in six different tissues (immature xylem, phloem, young leaf, mature leaf, flowers and shoot tips) of *E. grandis* were analyzed in order to identify CAZyme domain families in involved in xylogenesis. The comparison of expressed CAZymes in the immature xylem and young leaf of *E. grandis* and *P. trichocarpa* is discussed. Finally, the impact and conclusions of the study are presented.

Chapter 3: Found at the end of the dissertation, the Concluding Remarks discuss the findings of the study and how fit into the context of the research that has already been performed. The future prospects of the findings in terms of improvements, limitations, and impact are examined.

Outcomes: Results from this study have been presented at the South African Society of Genetics conference in 2011, in an oral presentation entitled: "The CAZyme repertoire of woody perennials." Results were also presented at the Plant and Animal Genomes conference in 2012, in an oral presentation entitled: "Comparative genomics of CAZyme domain frequency and diversity in *Eucalyptus grandis* and other land plants."

Acknowledgements

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

- Prof. Zander Myburg for his guidance, support, and patience, allowing me to learn and grow as a scientist in an excellent environment.
- Prof. Fourie Joubert for his support in the Bioinformatics Department, and most of all for being a friendly supporter with novel insight throughout this study.
- Dr. Eshchar Mizrachi for his infinite patience, willingness to listen and guide, and for keeping me in line. Most importantly, for providing me with the best example of how to work, love science, and strive to be better.
- Dr. Nicky Creux for being so generous with her time; listening to problems, celebrating the victories, providing advice, and always having my back. Nicky has consistently been my biggest supporter, she taught me how to teach, which is a gift I can never repay. If I could be half as awesome as Nicky, I would count myself extremely lucky.
- Dr. Anna Kersting for introducing me to protein domains and genomes, and for always being on hand (or Skype) for help and advice.
- Dr. Charles Hefer for all his patience in teaching me Python and data analysis, and for laughing at, and with, me at my more ridiculous/stupid requests/mistakes.
- Andrew dos Santos for helping me to learn how to code, and for being the best example of how being lazy can make for being a stellar programmer (i.e. keep it simple, stupid.)
- Mrs. Marja O'Neill for all the coffee and support, and for being a great example of professionalism, and for listening to me and swiftly bringing me back down to earth when I needed it, all without never letting me forget that she was still in my corner.

- Big thanks to Ritesh and Jono who put up with my lab antics with grace and patience, and for being the best lab mates and friends. You guys made coming into the lab so much fun, even when I didn't want to be there.
- All the staff and student of FMG, especially everyone in Lab Z: Nicky, Ritesh, Jono, Steve, and Matt. Lab Z is without a doubt the best place to do research, filled with people willing to listen and contribute, all the diverse discussions, the excellent music, and support have been invaluable to me.
- The National Research Foundation (NRF), The Human Resources and Technology for Industry Programme (THRIP), and Sappi for funding.
- The Department of Genetics and the Forestry and Agricultural Biotechnology Institute at the University of Pretoria for the world class infrastructure and support, and for the constant reminder that Science is a community driven pursuit.
- Tamaryn Hine, massive thanks for all the statistics help, support, reality checks, and inspiration.
- To all my friends, especially the boys, Ty, Dougie, and Koeks, for cheering me up every time I needed it last year, and for patiently listening to me rage about science even though you had no idea what I was talking about.
- Pooja and Karen, you guys are the best, and I would have lost even more of my sanity without your unconditional love, support, jokes, beers, coffee and smoke breaks.
- Lastly, to my family, Mom, Dad, Charlé, and Russ. I cannot express how grateful I am to all of you. Financial support through some very tough times is the absolute least of what you guys have done for me. Each of you is an incredible inspiration and source of strength for me everyday. Just knowing that I can come home to my best friends and a cold beer around the kitchen table every weekend is everything to me.

Table of Contents

List of Figures and Tables

Chapter 1

Literature review:

Carbohydrate Active enZymes and wood development

1.1. Introduction

Eucalyptus is the most widely grown short rotation woody crop genus, owing to excellent growth and wood fiber qualities for the production of paper and pulp, as well as chemical cellulose (Grattapaglia *et al.*, 2012). The last few years have seen a large amount of genomic and transcriptomic data available for *Eucalyptus*, including the newly sequenced reference genome of *Eucalyptus grandis* (Myburg *et al*., in review; Myburg *et al*., 2011; Hefer *et al*., 2011). Efforts are now underway to integrate the genome sequence data with other molecular data (gene expression, biochemical pathways), as well as growth and wood properties, to discover what contributes to the superior fiber properties of *Eucalyptus* spp. An approach that would be informative in terms of analyzing protein-coding genes that are responsible for secondary cell wall biopolymer synthesis is a protein domain centric analysis of the genome. An example would be analyzing the known Carbohydrate Active enZyme (CAZyme) domain containing proteins in *E. grandis*. CAZyme protein domains are a broad collection of the enzymatic and ancillary domains responsible for the synthesis, degradation and modification of polysaccharides (Cantarel *et al.*, 2009), and are especially relevant in the context of secondary cell wall (SCW) formation.

The availability of the genome sequence of *E. grandis* will allow for research into what determines the fiber cell properties of this commercially important short rotation woody crop species (Myburg *et* al., 2011; Paiva *et al.*, 2011; Grattapaglia *et al.*, 2012). By using *in silico* analysis of protein domains, the carbohydrate active enzymes (CAZyme genes) present in the genome can be identified. CAZymes are important in understanding how the secondary cell wall is synthesized, modified and degraded during the lifespan of the plant. Through a genome-wide analysis of gene models that contain CAZymes, a greater understanding of the enzymes responsible for cell wall formation within *Eucalyptus* can be gained, along with a greater general knowledge of carbohydrate metabolism in plants.

The rapid advancement of sequencing technology in recent years has provided researchers with massive amounts of genomic and transcriptomic data to analyze, including microarray and expressed quantitative trait loci (eQTL) data (Auerbach *et al.*, 2002). Expressed sequence tag (EST) and next generation sequencing (NGS) data have also proven to be extremely valuable in the identification and quantification of expressed genes in different plant tissues as they relate to metabolic processes and how they differ between tissues (Déjardin *et al*, 2004; Ward *et al*, 2012). NGS sequencing platforms for the assembly of transcriptomes using short reads include the Illumina platform, which has been used to *de novo* assemble the transcriptome of a *Eucalyptus* hybrid (Mizrachi *et al*., 2010). For this transcriptomic data to have relevance in a biological context, the functions of the genes and transcripts being analyzed must be identified (Koestler *et al.*, 2010). Together with transcriptome data from various cell and tissue types under different conditions, as well as proteomic and metabolomic data, the interactions between the proteins and metabolites in the cell can be modeled and analyzed by algorithms developed by bioinformaticists and biologists (Auerbach *et al.*, 2002; Keurentjes *et al.*, 2007; Verwoerd, 2011). The goal of modeling these interactions is to be able to understand how the cell, and ultimately the organism functions as a system (Mizrachi *et al.*, 2011).

The proteome of an organism is the total complement of proteins present in a biological unit, where the biological unit can be the whole organism, an organ, a tissue type, a single cell or an organelle (Abril *et al*., 2011). The proteome of a single organism shows variation at the tissue and cellular level, dependant on a myriad of external and internal factors (Abril *et al*., 2011). At the tissue level, the proteome can give invaluable insights into the metabolic functioning of that tissue, such as the identity and abundance of enzymes present that are involved in carbohydrate metabolism in sink tissues such as wood (Abril *et al*., 2011). A major setback to the use of proteomics at a large scale is the lack of high-throughput proteome technologies (Renuse *et al*., 2011). However, the differential gene

expression that dictates the proteome content of each tissue type can be inferred from their transcriptomes to a certain extent (Kryvych *et al.*, 2010). The proteins ultimately present in individual cells and tissue types differ from the mRNA transcripts from which they are derived, due to mRNA silencing, mRNA degradation and post-translational modification of proteins (Krol *et al.*, 2010; Remmerie *et al.*, 2011). When analyzing whole genomes and transcriptomes however, the primary transcript and transcript levels are generally taken as a proxy for protein level, as shown by the strong association of transcript regulation and xylogenesis (Hertzberg *et al.*, 2001). Annotation of genome sequence data is the first step to identify the functional elements contained within it (not only gene products, but promoters, transposable elements etc.), and is essential in understanding the biological implications of genomic data on the cell (Filipski & Kumar, 2005).

Despite the extensive research that has been performed in the field of wood formation, there are still many outstanding questions regarding carbohydrate metabolism in woody tissues (Mellerowicz & Sundberg, 2008). It is widely accepted that transcriptional regulation plays a major role in wood formation (Hertzberg *et al.*, 2001; Demura & Fukuda, 2007; Mellerowicz & Sundberg, 2008). The contribution of genomic CAZyme domain content to the differences in carbohydrate metabolism between divergent plant species is as yet unknown, as comparisons of CAZyme domain content have been focused on certain enzymatic classes in a handful of plant species (Henrissat *et al.*, 2001; Djerbi *et al.*, 2005; Geisler-Lee *et al.*, 2006; Tyler *et al.*, 2010). The relationship of CAZyme domain diversity and evolution to plant organizational complexity and carbohydrate metabolism is poorly understood. This review deals with contribution of the enzymatic CAZyme domains that synthesize, modify and degrade carbohydrates in plant species, with a focus on wood formation and *Eucalyptus*.

1.2. Protein domains

The three-dimensional structure of a protein is a consequence of the chemical and physical properties of the amino acids of which it is composed (Chothia *et al.*, 1977). Motifs are conserved amino acid sequences, and have characteristic certain secondary structures, such as α-helices and β-pleated sheets within protein domains, e.g. hair-pin loops formed within cystatin protease proteins have a highly conserved QXVXG motif (Dutt *et al.*, 2010; Schaeffer & Daggett, 2011). Domains themselves are made up of multiple motifs, characterized by the presence of a broader class of re-occurring structures, such as TIM barrels or αβ-plaits (Orengo *et al.*, 1997; Schaeffer & Daggett, 2011). Proteins, in turn, can be composed of single or multiple domains. Protein domains are independently folding units of protein structure that reoccur in genes throughout the genome of any organism with varying levels of amino acid homology (Lam & Blumwald, 2002; Caetano-Anollés *et al.*, 2009).

The biological activity of a protein is determined by the interactions between the domains present in that protein, and how they are orientated with each other impact the protein's functionality (Littler & Hubbard, 2005). For example, the Arm repeat domain (PF00514) is found in all eukaryotic organisms, and is implicated in a number of cellular processes such as signal transduction, nuclear import and ubiquitination (Samuel *et al.*, 2006). However, in plants, when it is associated with the U-box domain (PF04564) in the ARC-1 protein, it enables self-recognition to prevent self-pollination (Samuel *et al.*, 2006). Thus it is important to examine single domains in context with all other domains that they are found with to understand the contribution that they make to the overall functional diversity.

1.2.1. Domain interactions

Domains are grouped into families on the basis of sequence homology, and into super-families on the basis of structural homology (Russell *et al.*, 1998). The distinction between the two allows for interactions between domains to be better understood in a functional

context as sequence conservation in protein domains does not always reflect function as well as structure does (Vogel *et al.*, 2004). The majority of protein domain super-families interact with only one other super-family of domains. The exceptions are promiscuous domain super-families involved in widely used biological functions like DNA binding and ATP/GTP hydrolysis that are found in every organism (Littler & Hubbard, 2005). Littler *et al.* determined that out of 79 protein domain super-families, 33 formed multi-domain combinations, and 46 formed combinations with only one other domain super-family (Littler & Hubbard, 2005). Multi-domain combination super-families tend to vary the orientation in which they are found in the protein chain more often than single partner domain super-families. It is thought that there are a finite number of domain fold variations possible in all proteins which evolve independently, and that all proteins consist of various combinations of these domains to result in the "protein universe" (Littler & Hubbard, 2005). Thus by characterizing the functional and evolutionary relationships between domains, and domain super-families, a clearer understanding of protein evolution may emerge.

1.2.2. Domain evolution

Evolutionary events involving protein domains are classified into three classes: insertions/deletions, exchanges and duplications/repetitions (Bjorklund *et al.*, 2005). Björklund et al. found in their analysis of domain evolution that domain insertions were four times as frequent as deletions, suggesting that domain deletions are a less common mechanism of protein evolution (Bjorklund *et al.*, 2005). In proteins with more than two domains, insertions/deletions and duplications usually occur in the N- and C- terminals. Events involving more than one domain are generally duplications at the N- and C- terminals, with insertions and exchanges of more than one domain being rare. A mechanism by which enzymes may evolve new substrate specificities is by the addition of a novel binding domain (Bjorklund *et al.*, 2005). Binding domains are often added to existing catalytic domain architectures, thus modifying the substrate specificity of existing enzymes (Bjorklund *et al.*,

2005). For a review of the mechanisms and implications of domain evolution, see Vogel *et al.*, (2004) and a comprehensive analysis of domain gains, losses and rearrangements in plants has been done (Kersting *et al.*, 2012).

Domains can evolve new functions through single amino acid changes, insertions and deletions. Changes of this nature are found frequently in domains that have been duplicated, as the selective constraint on their sequence relaxes, and changes become fixed (Buljan & Bateman, 2009). Further more, duplications within gene families allow for neo-functionalization through regulatory divergence (Duarte *et al.*, 2006; Shakhnovich & Shakhnovich, 2008; Zou *et al.*, 2009). These changes in function after duplication is a factor that has impacted the ability of a core set of domains to produce the wide range of protein function on which existing biochemical pathways and organisms rely. For example, studies have shown that an ancestor of the immunoglobulin domain involved in invertebrate immune response has a function related to kinase signaling (Buljan & Bateman, 2009), thus lineage-specific changes in a protein domain are essential in evolutionary processes. An interesting example of domain evolution is the case of the SEX4 (starch excess) mutants in *Arabidopsis thaliana* and the laforin mutants of the *EPM2A* gene in humans; both genes contain dual specificity phosphatase (DSP) domains and carbohydrate-binding modules (CBMs) from family 20 CBMs (Moorhead *et al.*, 2009). The DSP domain dephosphorylates specific substrates in different organisms, in humans, the substrate for dephosphorylation is glycogen, and in plants it is starch. In humans, the CBM20 domain is at the N-terminal, and in *A. thaliana* it is found at the C-terminal (Gentry *et al.*, 2009). It has been found that by inserting the laforin DSP domain into *sex4* mutants in *A. thaliana*, the starch excess phenotype can be reversed, highlighting the functional implications of lineage-specific domain rearrangement and the wide evolutionary conservation of domain function (Gentry *et al.*, 2007).

In the context of plant evolution, domain emergence has been shown to play an important role (Kersting *et al.,* 2012). Domain emergence is the appearance of novel protein folds that form domains in the highly dynamic sequence space of evolving genomic regions (Caetano-Anollés & Nasir, 2012). The Viridiplantae lineage has 545 unique domains, mostly stress-response and secondary metabolite synthetic domains, compared to 30 in the insect lineage (Kersting *et al.,* 2012; Moore and Bornburg-Bauer, 2011). Domains related to photosynthesis and primary metabolic processes are not unique to the plant lineage, as photosynthesis also occurs in non-green algae and some bacteria (Kersting *et al.,* 2012). The importance of the emergence of novel domains in plants is suspected to be an adaptation to the long-lived, sessile lifestyle of plants, in addition to unique metabolic processes (Kersting *et al.,* 2012).

1.2.3. Domain discovery and annotation

There are a number of tools that are used to identify evolutionary events involving protein domains and homologous domain families. Using a position specific scoring matrix to identify distant evolutionary relationships between amino acid sequences can use variations on Basic Local Alignment Search Tools (BLAST) (Altschul et al., 1990), such as PSI-BLAST to detect protein domains. Hidden Markov models (HMMs) have however proven to be the most effective and widely used tool for detecting related protein domains (Eddy, 2001; Krogh *et al.*, 2001). HMMs are statistical models which are applied in bioinformatics to identify conserved secondary structures in proteins by aligning amino acid sequences and calculating the probability of transitions and fluctuations in the sequences according to a learnt set of parameters of mutations (insertions, deletions and amino acid substitutions) for that domain family (Eddy, 1998; Horan *et al.*, 2010; Reker *et al.*, 2010). This is done to identify domain families due to the fact that the protein fold and secondary structure of the domain family is often more highly conserved than the nucleotide sequence (Bradshaw *et al.*, 2011). In plants, it has been found that the functional domains present in the xylem transcriptome and genome

of *P. trichocarpa* is more highly conserved between selected vascular and non-vascular plants (77-85% at 1^{e-5}) than the nucleotide sequences (35-50% at 1^{e-50}) (Li *et al.*, 2010).

The principles behind automatic genome annotation domain searches in protein databases are precalculated Hidden Markov models (HMMs) for each protein domain (Hunter *et al.*, 2009; Yin *et al.*, 2012). There are many of these databases available to researchers. One such database is dbCAN (http://csbl.bmb.uga.edu/dbCAN/), which specifically uses HMMs to discover CAZymes in sequenced genomes, using domain family alignments taken from the CAZy database (www.cazy.org) (Yin *et al.*, 2012, Lombard *et al*., 2014). This database was developed as only 46% of CAZymes in the CAZy database had domain models in Pfam (pfam.sanger.ac.uk) (Punta *et al*., 2012), and to aid in the automatic annotation of CAZymes in whole genomes (Yin *et al.*, 2012).

1.3. Wood formation

Plant cells provide bulk of renewable carbon on the planet by sequestering carbon from the atmosphere during photosynthesis and storing it as cellulose and as other polysaccharides in the cell wall (Smith & Stitt, 2007). The plant cell wall is matrix composed of heterogeneous polysaccharide and phenolic biopolymers, and proteins (Keegstra *et al.*, 1973; Keegstra, 2010). The polysaccharides within the cell wall are of utmost importance as they give the plant structural and mechanical strength, and can be harnessed for biofuel energy (Hinchee *et al.*, 2009; Pauly & Keegstra, 2010; Somerville *et al.*, 2010).

The two types of cell walls found in woody angiosperms are primary and secondary cell walls. The primary cell wall is established early in the life cycle of all plant cells, and has a flexible matrix that consists of disorganized cellulose microfibrils cross-linked with hemicelluloses, pectins and proteins (Ray, 1967; Cosgrove, 2005; Harris & DeBolt, 2010). The primary cell wall (PCW) is deposited between the middle lamella and the plasma

membrane, and the secondary cell wall (SCW) is subsequently deposited between the primary cell wall and the plasma membrane (Plomion *et al.*, 2001; Cosgrove, 2005). The walls of secondary cells are composed of high amounts of cellulose and less pectin than the PCW, and are form the vessel and fiber cells of secondary xylem in wood (Mellerowicz *et al.*, 2001). Secondary xylem is formed in the load bearing vascular tissue of plants and functions as a transport system for water and dissolved nutrients, as reviewed in Brodribb (2009). Understanding the formation and structure of secondary cell wall biosynthesis will rely substantially on a full understanding of the proteins responsible for the synthesis, modification and degradation of SCW biopolymers.

A. thaliana is the model upon which most functional annotation in plant genomics is based due to the large amount of experimental evidence gathered in this species. Experimentally validated proteins coded by *A. thaliana* genes allow functional inferences to be made upon homologous genes in other plant species (Garcia-Hernandez *et al.*, 2002; Wienkoop *et al.*, 2010). As the model plant species, the *A. thaliana* genome was sequenced in 2000 (AGI, 2000), and since then, much headway has been made in annotating the functional regions of the genome (Jin *et al.*, 2010). Identification and quantification of plant cell wall proteins is extremely difficult, as isolating the membrane-bound proteins of the cell wall in appreciable amounts is a challenging task as they are embedded in the complex network of cell wall polysaccharides that can only be extracted separately, and often destructively (Jamet *et al.*, 2008; Abril *et al.*, 2011). Despite the difficulties associated with studying the proteins of the cell wall, it is important to have an understanding of the cell wall proteins, as they are responsible for the modification of the cell wall polysaccharides throughout the lifespan of the plant (Harris & DeBolt, 2010). Knowledge of the sub-cellular localization of proteins adds to the ability to predict their function with greater certainty. In plants, protein localization data is especially informative, as membrane bound proteins are likely to be involved in cell wall biosynthesis, chloroplast-localized proteins are likely to be involved in

photosynthesis, and hemicelluloses and pectins are formed by protein complexes in the Golgi apparatus (Li & Chiu, 2010; Crowell *et al.*, 2010; Oikawa *et al.*, 2012). Understanding of localization and protein properties may be useful in ascribing functions to proteins believed to be involved in polysaccharide biosynthesis.

1.3.1. Cellulose

Cellulose microfibrils are integrated into the cell wall by a membrane bound complex of cellulose synthase (CESA) proteins (Gardiner *et al.*, 2003; Carroll & Specht, 2011). The CESA proteins are found in the CESA rosette, which consists of six subunits, each comprising of six CESA proteins (Somerville, 2006; Joshi & Mansfield, 2007). Each CESA protein is thought to form a β-(1-4)-linked glucose chain (of which cellobiose is the repeating unit), which coalesce and crystallize to form a higher order cellulose microfibril. Cellulose microfibrils can be found in different lengths and the degree of polymerization can vary between different cell, tissue and plant types. However, other than these two variables, cellulose is the most homogenous biopolymer found in the plant cell wall (Somerville, 2006). Cellulose synthesis is an important target of research as cellulose is the most abundant biopolymer on earth. It is used to produce high-value chemical cellulose derivatives such as nitrocellulose and cellulose acetate, and as a future source of bio-ethanol. Elucidating the mechanism of its integration into the cell wall will enable researchers to utilize its potential most effectively (Mizrachi *et al.*, 2011).

1.3.2. Hemicellulose

Hemicelluloses are a diverse group of glycans, which have β-(1→4)- linked glucose, mannose, or xylose backbones and a wide variety of side-chains including xylose, galactose and arabinose (Scheller & Ulvskov, 2010). Hemicelluloses have a similar equatorial backbone as cellulose, but are branched with a variety of sugar moieties, which prevent them from forming microfibrils themselves (Scheible & Pauly, 2004). Their main function is to crosslink with cellulose and lignin, and strengthen the cell wall (Kryvych *et al.*, 2010;

Scheller & Ulvskov, 2010). Primary cell walls typically have lower percentage of cellulose and lignin than secondary cell walls, with a higher percentage of hemicelluloses and pectins (Reiter, 2002). The substrates used to synthesize the hemicelluloses are primarily nucleoside diphosphate sugars, generally UDP derivatives (Fry, 2004). In addition, the hemicelluloses found in primary cell walls tend to be more highly substituted than those found in secondary cell walls, which is evident in the fact that primary cell walls have more disorganized cellulose arrangement than secondary cell walls (Burton *et al.*, 2010).

The major hemicelluloses in plant cell walls are xylans and (gluco)-mannans that share the β-1-4-linked glucan chain backbone substituted with various residues to varying degrees (Scheller & Ulvskov, 2010). Glucuronoxylan is a major hemicellulose in the SCWs of dicots, it consists of a β-(1,4)-linked D-xylosyl backbone, to which α-D-glucuronic acid (GlcA), 4-*O*-methyl-α-D-glucuronic acid (MeGlcA), acyl and methyl group residues can be found at the O-2 position (Lee *et al.*, 2009). Xyloglucan has patterns of repetitive xylosyl residues linked to glucose at position 0-6 in the β-1-4-linked glucan chain backbone, and is prevalent in the PCW of angiosperms (Zabotina, 2012). Hemicelluloses and pectins are synthesized at the Golgi and transported to the cell wall by vesicles, which then fuse with the plasma membrane where the matrix polysaccharides integrate with the cellulose microfibrils to form the cell wall matrix (Cosgrove, 2005). The hemicelluloses allow the cellulose microfibrils to crosslink, giving the cell wall a rigid structure with gel-like properties that enables vertical growth and inter-cell porosity while the tissues are young, and a hollow lignified structure once the tissue has matured (Burton *et al.*, 2010). There are a vast number of hemicelluloses found in plant cell walls, due to the great variety of sugar moieties that can be appended to the backbone. Substituted hemicelluloses such as xylan play an important role in the structural integrity of the secondary cell wall, and as such the mechanism of their synthesis is an important area of research.

1.3.3. Other non-cellulosic biopolymers

Pectins are present in primary and secondary plant cell wall in varying degrees, also varying in the amount present between monocots and dicots (Mohnen, 2007). Pectins are complex molecules consisting of galacturonic acid containing polysaccharides (Mohnen, 2007). The exact structure of pectin is unknown at this time, as it is impossible to extract pectin from the cell wall intact (Atmodjo *et al.*, 2013). It is known that pectin consists of three different types of polysaccharides; Homogalacturonan (HG), Rhamnogalacturonan I (RGI) and Rhamnogalacturonan II (RGII) (Mohnen, 2007). Pectins contribute to cell wall growth by improving cell adhesion and strength via crosslinking with cellulosic polysaccharides and lignin (Caffall & Mohnen, 2009). Considering the importance of pectin in cell wall development and wood formation as evidenced by the high expression of pectin biosynthetic genes GAUT and GATL in *A. thaliana* stems, it is important to understand how pectin is synthesized (Lerouxel *et al.*, 2006; Minic *et al.*, 2009; Atmodjo *et al.*, 2011).

Lignin is polymerized from phenylpropanoids monomers, and functions to increase mechanical strength and protect the plant from microbial degradation; as such lignin limits the enzymatic digestion of the plant cell wall during industrial processing (Vanholme *et al.*, 2010). Lignification of secondary cell walls is most evident in the tracheary elements of secondary xylem, where the cell undergoes a marked differentiation between primary and secondary cell wall. Tracheary elements undergo rapid cell elongation and secondary cell wall deposition, during which the cell wall is enriched with phenolic compounds, after which it experiences programmed cell death and loses its cellular contents. What remains is a hollow, lignified tube that conducts water from the roots to the living tissues of the tree (Roberts & McCann, 2000). Lignin biosynthesis is regulated by the abundance of other sugars present in the tissue in addition to being induced by various abiotic and biotic factors. Lignin content affects the extractability of cellulosic biopolymers from the cell wall, and is therefore

and important aspect to consider in genetically engineering the cell wall for biotechnological applications (Vanholme *et al.*, 2008; Vega-Sánchez & Ronald, 2010).

1.3.4. Cell wall proteins

In addition to cellulose, hemicelluloses, pectins and lignin, there are several proteins that are embedded in the plant cell wall that function in signaling and cell wall growth and modification (Jamet *et al.*, 2006). The primary cell wall grows by a process known as biopolymer creep, in which the cell wall biopolymers increase the cell wall surface area by sliding across each other, as well as the concurrent synthesis of new biopolymers (Cosgrove, 2005). The mechanism by which biopolymer creep is achieved is largely due to a group of pH-dependant proteins known as expansins. Expansins cause loosening of the cell wall by disrupting the contact between cell wall polysaccharides and thereby allowing them to move past each other. The lines of evidence for this mechanism of action of expansins come from the fact that there are no catalytic domains present in any expansin proteins studied to date (Cosgrove, 2000; Choi *et al.*, 2008). Extensins are a group of hydroxyproline-rich glycoproteins found in the primary cell wall which function to crosslink to each other to form a network that aids cell wall development by reinforcing the polysaccharides in the cell wall (Showalter, 1993; Darley *et al.*, 2001). Lectins and Leucin-rich Repeat (LRR) proteins are known to be important for self- and nonself- recognition processes such as signaling and defence, and can be found in the plant cell wall (van Damme *et al.*, 2008; Jamet *et al.*, 2008). It is essential to take all the components of the cell wall into account when attempting to understand the synthesis, degradation and modification of cell wall polysaccharides

1.4. CAZymes

Carbohydrate active enzymes (CAZymes) are the enzymes responsible for the synthesis and remodeling of polysaccharides and other molecules in which glycosidic bonds are found. In the context of cell wall synthesis, CAZymes are the functional enzymes responsible for polysaccharide synthesis, modification and degradation. Glycosyl transferases (GTs),

glycosyl hydrolases (GHs), pectate lyases (PLs) and carbohydrate esterases (CEs) are all classified as CAZymes in the CAZy database (Henrissat & Davies, 2000). CAZy (www.cazy.org) is the database acknowledged as the seminal reference when working with carbohydrate active enzymes. CAZy is a manually curated database that groups carbohydrate active enzymes and their appended domains into well characterized, standardized families (Lombard *et al*., 2014). The system that is used to group the CAZymes and their associated domains into families is amino acid sequence based, and is correlated more closely with protein fold than enzyme/ligand specificity (Cantarel *et al.*, 2009). CAZy is a resource which aims to give the glycobiology researcher information regarding the structural features of the CAZyme family they are interested in by providing up to date links with resources such as Pfam (pfam.sanger.ac.uk), NCBI (www.ncbi.nlm.nih.gov/) and Interpro (www.ebi.ac.uk/interpro/) (Cantarel *et al.*, 2009). The CAZy database families are used to construct the HMMs for each domain family in dbCAN as discussed above.

1.4.1. Glycosyltransferases (GTs)

The enzymes that catalyze the synthesis of the glycosidic bonds are known as glycosyltransferases (GTs) (Scheible & Pauly, 2004). GTs are responsible for the formation of glycosidic bonds between the sugar donor substrate and the acceptor molecule, which may be a wide variety of molecules, from mono- to polysaccharides, lipids and proteins (Lairson *et al.*, 2008). GTs are present in every form of life and are involved in many functions within cells, including signaling and synthesis of biopolymers. The GT family of genes in plants is a large and diverse one, which catalyzes a wide range of reactions with many donor substrates (Henrissat & Davies, 2000; Henrissat *et al.*, 2001; Coutinho *et al.*, 2003; Palcic, 2011). Many members of GT families involved in biopolymer synthesis have been characterized in plants (Egelund *et al.*, 2004), among them are the families responsible for cell wall biopolymer synthesis.

The GT2 domain family is one of the largest of the known GT families; along with GT4, they comprise of over 50% of GTs in 94 described families (Hansen *et al.*, 2010) (www.cazy.org). The GT2 family contains the *CesA* superfamily of genes, members of which include the CESA protein encoding genes for cellulose biosynthesis, and the cellulose synthase like (CSL) protein encoding genes involved in hemicellulose biosynthesis, including xyloglucan, glucomannan and β-(1→3,1→4)-glucan (Arioli, 1998; Richmond & Somerville, 2000; Scheible & Pauly, 2004; Yin *et al.*, 2009; Scheller & Ulvskov, 2010; Carroll & Specht, 2011; Popper *et al.*, 2011; Dhugga, 2012). GT2s are inverting enzymes that change the conformation of the anomeric carbon of the transferred upon synthesis of the glycosidic bond (Hansen *et al.*, 2010). GT4 enzymes are retaining enzymes that maintain the anomeric carbon conformation upon synthesis, and they have been shown to have members involved in the synthesis of UDP-glucose from sucrose, and reversibly, sucrose formation for storage as starch via the *SuSy* gene (Haigler *et al.*, 2001).

Due to the heterogeneous nature of hemicelluloses, there are multiple GT families that have been shown to have family members involved in the biosynthesis of hemicelluloses, at least eight enzymes are proposed to be needed for the synthesis of glucuronoxylan alone (Lee *et al.*, 2009). GT43 and GT47 domain-containing enzymes have been shown to be involves in xylan backbone synthesis according to analysis of mutants of genes in these families (Brown *et al.*, 2009). Irregular xylem (*irx*) gene mutant plants have collapsed xylem vessel phenotypes and are useful in forward genetics approaches to discover genes involved in wood formation. IRX9 and IRX14 proteins, both members of the GT43 family, have been shown to be xylosyltransferases that act cooperatively in the elongation of the xylan backbone (Lee *et al.*, 2012). GT47 family members including fragile fiber 8 (FRA8), also known as IRX7, and glucuronosyltransferase 1 (GUT1), also known as IRX10, proteins are also involved in the synthesis of glucuronylxylan at the stage of reducing end primer synthesis and display severe to moderate IRX phenotypes in *A. thaliana* (Jung *et al.*, 2008; Brown *et al.*,

2009; Wu *et al.*, 2010). Glucuronic acid substitution of xylan (GUX) proteins GUX1, GUX2 and GUX3 from the GT8 domain family are responsible for adding glucuronic acid and 4-*O*-methylglucuronic acid side chains to the xylan backbone, mutations of the genes encoding these proteins cause secondary cell wall defects (Lee *et al.*, 2012). Furthermore, PARVUS, another GT8 protein, is involved in glucuronoxylan primer synthesis in *Arabidopsis* and *Populus* secondary cell wall formation (Jung *et al.*, 2008; Lee *et al.*, 2009).

GT domain families also function in regulatory and signaling pathways through the glycosylation of acceptor molecules. GT1 family domains have the most members in many genomes, including plants (Yonekura-Sakakibara & Hanada, 2011). GT1 enzymes are UDP glucosyltransferases (UGTs) that transfer UDP glucose to low molecular weight acceptor molecules. UGTs participate in the biosynthesis of secondary metabolites including terpenoids, phenylpropanoids and steroids in plants (Yonekura-Sakakibara & Hanada, 2011). The UGT72E gene cluster in *A. thaliana* has shown to be responsible for the glycosylation of monolignols, specifically forming coniferyl alcohol 4-*O*-glycoside and sinapyl alcohol 4-*O*-glycoside in light grown tissues (Lanot *et al.*, 2006). Phenylpropanoid glycosides may be involved in lignin monomer transport, although proof for this role is currently lacking (Vanholme *et al.*, 2008). Over-expression of a *P. trichocarpa* GT1 domain containing enzyme in tobacco caused decreased lignin content and early flowering, indicating that these enzymes are important for xylem formation, although further studies into the biological mechanism involved are needed (Wang *et al.*, 2012). GT41 family proteins catalyze the transfer of an *N*-acetyglucosamine (*O*-GlcNAc) residue from UDP-GlcNAc to serine and threonine residues on proteins in a signaling system similar to phosphorylation (Henrissat *et al.*, 2008). Glucosylation of proteins in signaling is a dynamic, sensitive system that has cross talk with other post-translational modification mechanisms (Breton *et al.*, 2012). In *Arabidopsis* the spindly (SPY) protein contains the GT41 domain, and is an *O*-GlcNAc-transferase that causes *O*-GlcNAc modification the of DELLA proteins, causing them to be activated (Zhang *et al.,*

2010). DELLA proteins in plants are inhibitors of gibberellic acid, as such; *O*-GlcNAc activation of DELLA proteins by SPY causes a myriad of developmental effects (Silverstone *et al.*, 2007).

1.4.2. Glycosyl hydrolases (GHs)

Glycosyl hydrolases, or glycoside hydrolases (GHs), are enzymes that catalyze the hydrolysis of glucosidic bonds between molecules. GHs can be responsible for the degradation of the cell wall polysaccharides into their composite structures, and make their energy available for use by microorganisms and industrial extraction, or for the modification of glycosidic bonds in polysaccharide biopolymers (Taylor *et al.*, 2008; Minic, 2008). GHs are expressed endogenously in plants, and are known to function in remodeling the cell wall polysaccharides in tissue development, fruit ripening and leaf abscission (Roberts *et al.*, 2002). Some GHs are regulated in response to sugar starvation, and serve to allow the plant to utilize polysaccharides as an energy source in response to a sugar sensing mechanism that detects low levels of sucrose (Koch, 2004).

The GH class has the most families of all the CAZyme classes (132 families as of 07/2013, http://www.cazy.org/Glycoside-Hydrolases.html). GHs are also the best characterized class of CAZymes, due to their importance in bacterial and fungal biology as GHs allow fungi and bacteria to utilize the energy of complex carbohydrates. The complement of GHs in a lignocellulolytic microorganism can give insights into its nutritional lifestyle and potential pathogenicity. The fungus *Rhizopus oryzae*, which causes Rhizopus rot, has a complement of cell wall degrading enzymes that only allow it to digest simple sugars (Battaglia *et al.*, 2011). Through comparative analysis of the genomes of *Rhizopus spp.* and *Basidiomycota* and *Ascomycota*, it was shown that *R. oryzae* does not have the GHs that are found in other fungal species, which are necessary to digest complex carbohydrates (Battaglia *et al.*, 2011). *R. oryzae* was found to have more CAZymes for the degradation of chitosan

than other fungal species, and this result was functionally verified using growth analysis on different media (Battaglia *et al.*, 2011). In lignocellulolytic organisms, the genomic arsenal of GHs can therefore give insight to the metabolic lifestyle.

In plants, many GH families have been shown to be important in development and wood formation. Most notably, the GH9 domain family contains the *Korrigan* gene, which encodes for a membrane-bound endo-1,4-β-endoglucanase (Maloney *et al.*, 2011). The highly conserved *Korrigan* gene in monocots and dicots is essential for normal xylem development, as mutants have a dwarf phenotype and perturbed cell wall architecture (Nicol *et al.*, 1998). *Korrigan* (*kor*) shows very high coexpression correlation with the secondary cell wall *CesA* genes, and *kor* mutants have disruptions in cellulose crystallinity in secondary cell walls (Szyjanowicz *et al.*, 2004). Although the exact mechanism of *Korrigan* is unknown, it is believed to be involved in microfibril processing after synthesis (Maloney & Mansfield, 2010; Maloney *et al.*, 2011). Other GHs known to affect primary and secondary cell wall architecture are the chitinase-like (CTL) proteins, belonging to GH family 19. In *Gossypium hirsutum CTL1* and its homologue *CTL2* are expressed in elongating fiber cells, and mutants are deficient in cellulose (Zhang *et al.*, 2004). CTL1 and CTL2 have been shown to cooperatively participate in the normal assembly and interactions of cellulose microfibrils and hemicelluloses via glucan polymer binding and modification (Sánchez-Rodríguez *et al.*, 2012).

Genome wide comparisons of GHs in plant genomes have shown that they contain the same set of GH domains, with lineage specific differences in the frequency of these domains between species (Tyler *et al.*, 2010). The lineage specific differences in GH frequency can be linked to the structure and composition of cell walls in different plant lineages. As an example, the GH28 domain family has many more members in *A. thaliana*, compared to grasses such as *B. distachyon*, *Z. mays* and *Oryza sativa.* GH28 domains have

polygalacturonase activity, which acts to remodel pectin, and the type II cell walls found in grasses have lower levels of pectin than the type I cell walls found in *Arabidopsis* (Tyler et al*., 2010*).

1.4.3. Polysaccharide lyases (PLs)

Polysaccharide lyases (PLs) are involved in the degradation and modification of pectin via cleavage of uronic acid containing polysaccharides via a β-elimination mechanism, as opposed to hydrolysis by GHs (Cao, 2012). PLs are important in plants for a variety of functions, including plant defense, growth and development. The *A. thaliana* gene *PMR6* is a PL gene that affects the pectin composition of cell walls and consequently influences resistance to powdery mildew (Vogel *et al.*, 2002). In cotton, the pectate lyase gene *GhPEL* performs an essential function in the degradation of de-esterified pectin, as determined by knockdowns of the gene, which retarded cell wall elongation by preventing primary cell wall loosening (Wang *et al.*, 2010). The known functions of PLs are related to primary cell wall metabolism, and a role for PLs in secondary cell wall biosynthesis is not yet known (Caffall & Mohnen, 2009). Pectin is less abundant in the SCW of dicots (Ishii, 1997), and the remodeling of pectin is likely to be less important in SCW deposition than in the PCW (Palin & Geitmann, 2012).

1.4.4. Carbohydrate esterases (CEs)

Carbohydrate esterases (CEs) are responsible for the degradation and modification of O-acetylated and methylesterified sidechains of cell wall polysaccharides (Pawar *et al.*, 2013). Xylan acetylation promotes crosslinking of xylan with lignin in the wood cell wall, which strengthens the secondary cell wall (Tsai *et al.*, 2012). The acetyl groups present on xylan are known to reduce the hydrolytic capability of enzymes used to break down lignocellulosic biomass for biofuel production, thus CEs have been studied for their potential in increasing the saccharification efficiency. CE familes involved in xylan de-acetylation (acetyl xylan esterases- AXEs) are CE1, 2, 3, 4, 5, 6, 7 and 16, pectin acetyl esterases are found in family

CE13 and Rhamnogalacturonan esterases are found in family CE12 (Pawar *et al.*, 2013). CE family 15 glucuronyl-esterase genes have been shown to remove the ester link between lignin alcohols and methylglucuronic acid side chains of glucuronyl xylan. When these genes are over expressed in *A. thaliana*, the extractability of xylan is increased and there is an increase in soluble lignin (Tsai *et al.*, 2012).

1.4.5. Carbohydrate-binding modules (CBMs)

Carbohydrate-binding modules (CBMs) function to mediate protein-carbohydrate associations, and as such play an important role in the enzymatic degradation and formation of the plant cell wall polysaccharides. CBMs were first studied in cellobiohydrolase I from *Trichoderma reesei*, which was shown to have a catalytic domain, and a binding domain. The experiment determining the domains present in cellobiohydrolase I from *Trichoderma reesei* was done by Van Tilbeurgh et al. in 1986 by digesting the enzyme with papain which separated it into 56 kDa and ~10 kDa sections (van Tilbeurgh *et al.*, 1986). The native cellobiohydrolase I enzyme digests the insoluble microcrystalline cellulose substrate Avicel to give cellobiose as a product. In addition, cellobiohydrolase I hydrolyses soluble glycosides, thus displaying promiscuity in ligand binding. The 56 kDa CBHI was determined to be the catalytic subunit of the CBHI enzyme as no Avicelase activity was detected with heavily papain-digested CBHI, however, the soluble glycoside activity was still intact. Thus it was determined that the 10kDa fragment was specific for binding to insoluble cellulose. Subsequent studies into the thermodynamic and structural properties of cellobiohydrolases from *Trichoderma reesei* have since confirmed the modular nature of CAZymes (van Tilbeurgh *et al.*, 1989).

Since the advent of DNA sequencing technologies, new CBMs have been identified by examining the sequences of CAZymes with polysaccharide binding for modular domains which have sequence similarity to known CBMs (Bolam *et al.*, 2004). In the years since the

confirmation of the existence of CBMs in 1986, numerous studies have examined the diverse structural and functional relationships between the CBMs and the catalytic domains with which they co-occur (Guillén & Sanchez, 2010; Xie *et al*., 2001; Yin *et al*., 2011). The function of CBMs can be separated into three main processes, or effects. The first is the targeting effect, which describes the CBM's affinity and specificity for the ligand. The next is the proximity effect, which describes how the CBM allows the catalytic module that it is appended to remain in contact with the target biopolymer for an extended period of time. This is important, as the cell wall biopolymers are highly recalcitrant to enzymatic action. The third is the disruptive function, and describes the way in which the CBM disrupts the hydrogen bonds between biopolymers, facilitating the phase change between the soluble catalytic module and the insoluble cell wall biopolymers (Arantes & Saddler, 2010).

Hall et al. illustrate the significance of understanding the mechanism of action of CBMs in the pulp and paper industry, and for future biofuels applications in the study in 2010. The study showed that CBMs can be used to increase the efficiency of cellulases (GHs) in industry by reducing the crystallinity of cellulose (Hall *et al.*, 2011). As CBMs disrupt the hydrogen bonds between cellulose microfibrils and decrease the recalcitrance of cellulose to enzymatic digestion, the authors showed that pretreatment of crystalline cellulose with CBMs from family I isolated from *Trichoderma reesei* cellulases decreased the crystallinity index of Avicel and fibrous cellulose, making the release of glucose up to 25% more efficient once cellulases were added. Furthermore, the CBMs were found to be thermostable at up to 50°C, which is in contrast to the full length enzyme, and thus makes CBMs suitable for decreasing cellulose crystallinity in pretreatment protocols (Hall *et al.*, 2011).

An exciting application of CBMs is the engineering of tagged CBMs that can be used to visualize the cell wall polysaccharides *in situ* (McCartney *et al.*, 2004). Due to the specific targeting, recognition and modular nature of CBMs, they can be fluorescently tagged to allow

for the location of a very specific cell wall polysaccharide to be identified even when they are present in low concentrations (Filonova *et al.*, 2007). The knowledge of the localization of each of the polysaccharides in the SCW in relation to each other will greatly improve understanding of how the SCW is synthesized and how it might be modified.

1.4.6. Genome-wide analyses of CAZymes in plants

CAZymes have been examined from a genome-wide perspective in a number of studies. In plants, the first of these was done in 2001, directly after the release of the *A. thaliana* genome. The study analyzed the frequency and diversity of CAZymes in the *A. thaliana* genome (Henrissat *et al.*, 2001). CAZymes in *A. thaliana* were found to be a major class of genes in the genome, with 730 of the 25,000 genes in *A. thaliana* being GTs or GHs, the highest number found in any organism sequenced to date. The authors highlighted the large number of CAZymes in GT families responsible for the synthesis of cell wall polysaccharides, including the GT2s responsible for cellulose biosynthesis (Henrissat *et al.*, 2001). The number of GHs was also interesting, in that the *Arabidopsis* families with cellulase activity had fewer members than had previously been found in bacteria. The authors proposed that this decrease was due to the very specific modifications that GHs make during cell wall synthesis in plants versus the complete degradation of polysaccharides that occurs via bacterial GHs (Henrissat *et al.*, 2001). The GT1 CAZyme family was found to contain 116 potential members in *A. thaliana*, the largest of all the CAZymes (Henrissat *et al.*, 2001). This highlights the importance of this family in plant metabolite glycosylation, a regulatory mechanism that is still not fully understood to this day (Caputi *et al.*, 2012).

Comparative analysis between *A. thaliana* and *O. sativa* cell wall related genes in a previous study have observed differences in the sizes of a few GH and GT gene families between the two species (Yokoyama & Nishitani, 2004). Despite those few exceptions, the gene family sizes of GHs and GTs between these evolutionarily distinct species, the monocot

O. sativa and the dicot *A. thaliana*, the GT and GH gene family size and diversity was well conserved (Yokoyama & Nishitani, 2004). Differences in cell wall gene family diversity between the two species were observed for α-fucosyltransferase genes, which were not found in *O. sativa*, in concordance with the observation that *O. sativa* xyloglucan does not possess fucosyl residues on the galactosyl side chains (Yokoyama & Nishitani, 2004). Thus comparative genomic approaches to can confirm the differences in cell wall biology between two divergent species, along with the striking similarities in core cell wall structure.

GT gene families have been analyzed in the genomes of species basal to the land plants. In *Selaginella moellendorffii*, a basal vascular land plant, and *Physomitrella patens*, a non-vascular land plant, the inventory of GTs was found to be similar to that of seed plants *A. thaliana* and *O. sativa* (Harholt *et al.*, 2012). The most distinct differences between the lycophyte and bryophyte, and the seed plants was in the GT gene family sizes, with the basal species having reduced family sizes in most cases (Harholt *et al.*, 2012). Some GT family members have been lost in the seed plant lineage, mainly GT51 and GT78, indicating retention of ancestral genes in the basal species (Harholt *et al.*, 2012). Comparisons in the cell wall polysaccharides of the basal plant species and the seed plants showed similarities in composition and differences in abundance, and the authors propose that the strategies for cell wall formation between basal land plants and seed plants were different to optimize cell wall architectures for each species using common genes (Harholt *et al.*, 2012).

GHs in the grasses *Brachypodium distachyon* and *Sorghum bicolor* have been analyzed and compared to each other and sequenced plant genomes to improve understanding of GH function (Tyler *et al.*, 2010). Analysis of GHs in grasses in addition to comparative analysis with *A. thaliana*, *P. trichocarpa* and *O. sativa* showed that lineage specific expansions/ contractions of specific GH families have occurred in GH families throughout plant evolution. However, the families present in the five genomes analyzed comprised of the

same 34 GH families (Tyler *et al.*, 2010). GH18 and GH19 domain families were analyzed as they have a shared chitinolytic activity, but have very different sequence and structural identity (Tyler *et al.*, 2010). GH18 and GH19 domain families both contain genes that are classified as pathogenesis related (PR) proteins, which are integral players in plant defense strategies. The authors found expansions in eudicot class V chitinases in GH18 that are the result of duplications in this lineage, however, class III GH18s were indistinguishable between eudicots and monocots (Tyler *et al.*, 2010). This, along with other data presented in the analysis, indicates a shared evolutionary history of plant GHs with lineage-specific functional diversification.

Geisler-Lee and colleagues performed the only comprehensive analysis of woody perennial CAZymes in *P. trichocarpa* in 2006, soon after the release of the *Populus* genome (Geisler-Lee *et al.*, 2006). They found that *P. trichocarpa* had more CAZyme genes than *A. thaliana* or any other sequenced organisms with 1, 647 CAZyme genes, including expansins. The authors noted the proportional increase in most *P. trichocarpa* CAZyme families compared to *A. thaliana* in concordance with known genome-wide duplications in *P.trichocarpa* (Djerbi *et al.*, 2005; Geisler-Lee *et al.*, 2006). The transcriptomes of *A. thaliana* and *P. trichocarpa* in different tissues and developmental conditions were analyzed using expressed sequence tag data. The analysis revealed that the wood forming tissues of *P. trichocarpa* had a higher abundance and greater diversity of expressed CAZymes compared to that of *A. thaliana*, leading the authors to highlight the role that CAZymes play in wood formation. Amongst the differences they found in the transcriptomic comparison was that *P. trichocarpa* also had greater frequency and diversity of expression of CAZymes related to secondary metabolite glycosylation, which is indicative of a more sophisticated defense arsenal (Geisler-Lee *et al.*, 2006).

1.4.7. CAZymes and cell wall evolution

Plant and algal cell walls have large lineage specific diversity that has arisen throughout their evolution (Popper & Fry, 2004). The diversity in cell walls can be related to the evolution CAZymes and the evolution of regulatory mechanisms, especially after genome-wide duplications (Fangel *et al.*, 2012; Rodgers-Melnick *et al.*, 2012). The conservation of core functionality of cell wall metabolic genes can be seen in numerous examples from different plant lineages. The predicted proteins of the unicellular *Chlamydomonas reinhardtii* and the multicellular *Volvox carteri* are remarkably similar. The green algae diverged from the land plant lineage 725-1200 million years ago (Becker & Marin, 2009), and their cell walls are similar in the composition of cellulose and hemicelluloses (Popper *et al.*, 2011). The green algae are non-vascular, so they do not possess the biosynthetic genes for lignin production (Vanholme *et al.*, 2010), which appeared first in the mosses (Weng & Chapple, 2010; Delaux *et al.*, 2012). Interestingly, the brown algae *Ectocarpus siliculosis*, which has diverged considerably from the plant lineage, shares many CAZyme gene families with plants, although the number of genes in each family is reduced compared to land plants (Michel *et al.*, 2010). *E. siliculosis* does however lack GT43 and GH16 (XTHs), the genes necessary for xyloglucan synthesis (Popper *et al.*, 2011).

The diversity of land plant cell wall polysaccharides is predominantly due to the inherent heterogeneity of cell walls and the enzymatic mechanisms that produce them (Burton *et al.*, 2010). Between land plant lineages there is considerable diversity in polysaccharide composition. Monocots contain less pectin and xylans in their cell walls than dicots do, and they have more heteroxylans (glucuronoarabinoxylans), as well as mixed-linked glucans (MLG) in the grasses specifically (Gibeaut & Carpita, 1994). The genes responsible for MLG

production belong to the CslF and CslH (GT2) family of glycosyltransferases and is found only in the grasses (Burton *et al.*, 2006; Doblin *et al.*, 2009). In terms of plant cell wall diversity, an important example is the evolution of the secondary cell wall *cellulose synthase* (*CESA*) genes. It has been hypothesized that genome-wide duplication and subsequent tissue/organ specific specialization and transcriptional changes are responsible for the diversity of gene families in the GT2 *CESA* gene family (Popper *et al.*, 2011).

1.5. *Eucalyptus*

One of the major global concerns for the future is the need to acquire sources of renewable, energy and bio-based materials. Scientists and policy makers have been looking to plantation forestry as a source of lignocellulosic biomass as the alternative fuel of the future (Hinchee *et al.*, 2009). This is due to the fact that a considerable amount of research has already gone into increasing the economic viability of woody plantation species in the past, as they are a major renewable resource for many industries worldwide, such as the paper and pulp industry. Short rotation woody, purpose grown tree species which can be harvested within 3 to 15 years are the ideal biomass feedstock, as they can sustainably replace at least 30% of dependence on fossil fuels according to a sustainability study done in the United States of America (Buford & Neary, 2010).

In order to meet the requirements for lignocellulosic biomass to be a viable option as a feedstock for these applications, the forestry industry needs to increase the annual yield of its crop species. This can be done by cell wall modification to increase the extractability of biopolymers in the present yield (Mansfield, 2009). The US Department of Energy project that an annual yield of 8-10 tons/acre of dry mass will be needed to sustain renewable energy production from forest plantations (Hinchee *et al.*, 2009). There are several genetic strategies that have been identified for achieving that aim. The first is traditional selective breeding, where two parental strains with desired qualities are crossbred and superior progeny are

selected and carried forward to the next generation, an extension of which is molecular marker assisted breeding technology. The second is to modify existing genes to improve the growth and productivity of the species, and the third is to add genes through genetic transformation. All three methods have been utilized in different wood species with varying degrees of success (for a recent review see Hinchee et al, 2009). The focus of our research is the hardwood species *Eucalyptus* and how it compares to other crop species such as *Populus* and *Pinus* in terms of its suitability as a crop species for bioethanol production and genetic modification. The availability of newly sequenced plant species data has allowed for comparative analyses to complement *A. .thaliana* as a model species for cell development (Jansson & Douglas, 2007). With data from *Populus* (Tuskan *et al.*, 2006) and *Eucalyptus* (Mizrachi *et al.*, 2010, Myburg *et al.,* in revision) genome and transcriptome sequences available, the genes responsible for xylogenesis can be studied in the appropriate manner, aided by the extensive data collected by the *Arabidopsis* research community. A major question that needs to be addressed with this data is how the genomic and transcriptomic diversity of CAZyme domain containing proteins is linked to the woody habit. Is the genomic frequency and diversity of CAZymes in woody plants different to those of non-woody plants, or have woody plants utilized existing CAZymes to form wood via regulatory mechanisms not found in non-woody plants?

Eucalyptus is a fast growing hardwood crop species, which is currently the most valuable and widely planted fiber crop species (Grattapaglia & Kirst, 2008). The main advantages of *Eucalyptus* and its hybrids as a plantation crop species in the forestry industry is fast rotation time compared to *Pinus* species (Hinchee *et al.*, 2009). *Eucalypts* are also more adaptable to biotic and abiotic stressors than *Populus*, which is the current model genus for wood development, and the first woody perennial genome sequenced (Jansson & Douglas, 2007). *Eucalyptus* is an attractive genus in which to study CAZymes in a woody perennial for a number of reasons. The genome of *E. grandis* has recently been sequenced, allowing for

genome-wide identification and annotation of CAZymes in a woody plant genus with unique wood producing properties (Myburg et al., 2011).

There is a large amount of transcriptome data available for *E. grandis* which allows for more in depth analysis of the data generated from the genome, as genes coexpressed with known secondary cell wall developmental genes can be identified (Myburg *et al.*, 2010). This coexpression data can be cross-referenced with the same data for *Arabidopsis* genes in databases such as ATTED-II (atted.jp) (Obayashi *et al*., 2009) and Genevestigator (www.genevestigator.com) (Zimmerman *et al*., 2004) for validation. *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid clones have recently been used to produce a high density genetic map of eQTL's in *Eucalyptus* (Myburg *et al.*, unpublished), which in tandem with the reference genome sequence will further aid the selection of candidate genes involved in secondary cell wall biosynthesis.

1.6. Conclusion

The synthesis, modification and degradation of the plant secondary cell wall is an important area of research due to the significance and potential impact this research has to industry (Mansfield, 2009; Hinchee *et al.*, 2009) (www.esa.org/biofuelsreports/). By proper utilization of *in silico* comparative genomics for analyzing predicted proteins that putatively modify cell wall polysaccharides, target proteins for biochemical analysis can be effectively selected. Due to the modular nature of the carbohydrate active enzymes that modify cell wall polysaccharides, the known CAZyme domains present in the gene models and transcriptomes of xylogenic tissue can be identified. The identification of known CAZyme domain containing genes in a newly sequenced genome not only aids annotation efforts, but may also lead to interesting biotechnology applications in terms of the identification of lineage- and species- specific CAZymes and novel SCW biosynthetic genes. Analysis of the literature shows that there are questions as to what the contribution of CAZyme domain frequency and

diversity across plant genomes is to the physiochemical properties of plant cell wall polysaccharides. This study aims to address these questions by analyzing the genomic frequency and diversity of CAZyme domains across twenty-two phylogenetically diverse plant species. Furthermore, the expression investment of CAZyme domains in woody and non-woody tissues in two distinct woody genera, *Eucalyptus* and *Populus*, will be compared to identify the patterns of CAZyme domain expression that define xylogenic tissue. This study will aid in expanding the understanding of the evolution of the woody habit by analyzing the enzymatic building blocks that synthesize, degrade and modify plant cell wall polysaccharides.

1.8. References

Abril N, Gion J, Kerner R, Müller-Starck G, Navarro RM, Plomion C, Renaut J, Valledor L, Jorrin-novo J V, Cerrillo RMN. **2011**. Proteomics research on forest trees, the most recalcitrant and orphan plant species. *Phytochemistry* **72**: 1219–42.

The Arabidopsis Genome Initiative. **2000**. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **1990**. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**: 403–410.

Arantes V, Saddler JN. **2010**. Access to cellulose limits the efficiency of enzymatic hydrolysis : the role of amorphogenesis. *Journal of Biological Chemistry* **3**: 1–11.

Arioli T. **1998**. Molecular analysis of cellulose biosynthesis in *Arabidopsis*. *Science* **279**: 717–720.

Atmodjo MA, Hao Z, Mohnen D. **2013**. Evolving views of pectin biosynthesis. *Annual Review of Plant Biology* **64**: 747–79.

Atmodjo MA, Sakuragi Y, Zhu X, Burrell AJ, Mohanty SS, Atwood J a, Orlando R, Scheller H V, Mohnen D. **2011**. Galacturonosyltransferase (GAUT)1 and GAUT7 are the core of a plant cell wall pectin biosynthetic homogalacturonan:galacturonosyltransferase complex. *Proceedings of the National Academy of Sciences* **108**: 20225–30.

Auerbach D, Thaminy S, Hottiger MO, Stagljar I. **2002**. The post-genomic era of interactive proteomics: Facts and perspectives. *Proteomics* **2**: 611–623.

Battaglia E, Benoit I, Brink J Van Den, Wiebenga A, Coutinho PM, Henrissat B, Vries RP De, van den Brink J, de Vries RP. **2011**. Carbohydrate-active enzymes from the zygomycete fungus *Rhizopus oryzae*: a highly specialized approach to carbohydrate degradation depicted at genome level. *BMC Genomics* **12**: 38–50.

Becker B, Marin B. **2009**. Streptophyte algae and the origin of embryophytes. *Annals of Botany* **103**: 999–1004.

Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A, Björklund AK, Frey-Skött J, Bjo K, Frey-sko J. **2005**. Domain rearrangements in protein evolution. *Proteins* **353**: 911–923.

Bolam DN, Xie H, Pell G, Hogg D, Galbraith G, Henrissat B, Gilbert HJ. **2004**. X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *The Journal of Biological Chemistry* **279**: 22953–22963.

Bradshaw CR, Surendranath V, Henschel R, Mueller MS, Habermann BH. **2011**. HMMerThread: detecting remote, functional conserved domains in entire genomes by combining relaxed sequence-database searches with fold recognition. *PLoS One* **6**: 1–17.

Breton C, Fournel-Gigleux S, Palcic MM. **2012**. Recent structures, evolution and mechanisms of glycosyltransferases. *Current Opinion in Structural Biology* **22**: 540–549.

Brodribb TJ. **2009**. Xylem hydraulic physiology: The functional backbone of terrestrial plant productivity. *Plant Science* **177**: 245–251.

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR. **2009**. Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in *Arabidopsis*. *The Plant Journal* **57**: 732–746.

Buford MA, Neary DG. **2010**. *Sustainable biofuels from forests: Meeting the challenge*.

Buljan M, Bateman A. **2009**. The evolution of protein domain families. *Biochemical Society Transactions* **37**: 751–755.

Burton RA, Gidley MJ, Fincher GB. **2010**. Heterogeneity in the chemistry, structure and function of plant cell walls. *Nature Chemical Biology* **6**: 724–732.

Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB. **2006**. Cellulose Synthase–Like *CslF* genes mediate the synthesis of cell wall (1,3;1,4)-beta-D-glucans. *Science* **311**: 1940–1942.

Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. **2009**. The origin, evolution and structure of the protein world. *The Biochemical Journal* **417**: 621–637.

Caffall KH, Mohnen D. **2009**. The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research* **344**: 1879–1900.

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**: 233–238.

Cao J. **2012**. The pectin lyases in *Arabidopsis thaliana*: evolution, selection and expression profiles. *PLoS One* **7**: 1–15.

Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S. **2012**. A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. *The Plant Journal* **69**: 1030–1042.

Carroll A, Specht CD. **2011**. Understanding plant cellulose synthases through a comprehensive investigation of the cellulose synthase family sequences. *Frontiers in Plant Science* **2**: 1–11.

Choi D, Kim JH, Lee Y. **2008**. Expansins in plant development. *Advances in Botanical Research* **47**: 47–98.

Chothia C, Levitt M, Richardson D. **1977**. Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences* **74**: 4130–4134.

Cosgrove DJ. **2000**. New genes and new biological roles for expansins. *Current Opinion in Plant Biology* **3**: 73–78.

Cosgrove DJ. **2005**. Growth of the plant cell wall. *Nature Reviews: Molecular Cell Biology* **6**: 850–861.

Coutinho PM, Stam M, Blanc E, Henrissat B. **2003**. Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends in Plant Science* **8**: 563–565.

Crowell EF, Gonneau M, Stierhof Y-D, Höfte H, Vernhettes S. **2010**. Regulated trafficking of cellulose synthases. *Current Opinion in Plant Biology* **13**: 700–705.

Van Damme EJM, Lannoo N, Peumans WJ. **2008**. Plant Lectins. *Advances in Botanical Research* **48**: 107–210.

Delaux P-M, Nanda AK, Mathé C, Sejalon-Delmas N, Dunand C. **2012**. Molecular and biochemical aspects of plant terrestrialization. *Perspectives in Plant Ecology, Evolution and Systematics* **14**: 49–59.

Déjardin A, Leplé J-C, Lesage-Descauses M-C, Costa G, Pilate G. **2004**. Expressed sequence tags from poplar wood tissues- a comparative analysis from multiple libraries. *Plant Biology* **6**: 55–64.

Dhugga KS. **2012**. Biosynthesis of non-cellulosic polysaccharides of plant cell walls. *Phytochemistry* **74**: 8–19.

Djerbi S, Lindskog M, Arvestad L, Sterky F, Teeri TT. **2005**. The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (CesA) genes. *Planta* **221**: 739–746.

Doblin MS, Pettolino FA, Wilson SM, Campbell R, Burton RA, Fincher GB, Newbigin E, Bacic A. **2009**. A barley cellulose synthase-like CSLH mediates (1,3;1,4)-beta-D-glucan synthesis in transgenic *Arabidopsis*. *Proceedings of the National Academy of Sciences* **106**: 5996–6001.

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, DePamphilis CW. **2006**. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution* **23**: 469–478.

Dutt S, Singh VK, Marla SS, Kumar A. **2010**. *In silico* analysis of sequential, structural and functional diversity of wheat cystatins and its implication in plant defense. *Genomics, Proteomics & Bioinformatics* **8**: 42–56.

Eddy SR. **1998**. Profile hidden Markov models. *Bioinformatics Review* **14**: 755–763.

Eddy SR. **2001**. HMMER User's Guide: 1–93.

Egelund J, Skjøt M, Geshi N, Ulvskov P, Petersen BL. **2004**. A complementary bioinformatics approach to identify potential plant cell wall glycosyltransferase-encoding genes. *Plant Physiology* **136**: 2609–2620.

Fangel JU, Ulvskov P, Knox JP, Mikkelsen MD, Harholt J, Popper Z a, Willats WGT. **2012**. Cell wall evolution and diversity. *Frontiers in Plant Science* **3**: 1–8.

Filipski A, Kumar S. **2005**. Comparative genomics in eukaryotes. *The Evolution of the Genome*. Academic Press, Burlington: 521–583.

Filonova L, Gunnarsson LC, Daniel G. **2007**. Synthetic xylan-binding modules for mapping of pulp fibres and wood sections. *BMC Plant Biology* **10**: 1–10.

Fry SC. **2004**. Primary cell wall metabolism : tracking the careers of wall polymers in living plant cells. *New Phytologist* **161**: 641–675.

Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, Huala E, Knee E, Lambrecht M, Miller N, Mueller L a, *et al.* 2002. TAIR: a resource for integrated *Arabidopsis* data. *Functional & Integrative Genomics* **2**: 239–253.

Gardiner JC, Taylor NG, Turner SR. **2003**. Control of cellulose synthase complex localization in developing xylem. *The Plant Cell* **15**: 1740–1748.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, *et al.* 2006. Poplar Carbohydrate-Active Enzymes. Gene identification and expression analyses. *Plant Physiology* **140**: 946–962.

Gentry MS, Dixon JE, Worby CA. **2009**. Lafora disease: insights into neurodegeneration from plant metabolism. *Trends in Biochemical Sciences* **34**: 628–639.

Gentry MS, Dowen RH, Worby CA, Mattoo S, Ecker JR, Dixon JE. **2007**. The phosphatase laforin crosses evolutionary boundaries and links carbohydrate metabolism to neuronal disease. *The Journal of Cell Biology* **178**: 477–488.

Gibeaut DM, Carpita NC. **1994**. Biosynthesis of plant cell wall polysaccharides. *The FASEB Journal* **8**: 904–915.

Grattapaglia D, Kirst M. **2008**. *Eucalyptus* applied genomics: from gene sequences to breeding tools. *New Phytologist* **179**: 911–929.

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, Potts BM, Myburg AA. **2012**. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genetics & Genomes* **8**: 463–508.

Guillén D, Sánchez S. **2010**. Carbohydrate-binding domains: multiplicity of biological roles. *Journal of Molecular Biology* **85**: 1241–1249.

Haigler CH, Ivanova-Datcheva M, Hogan PS, Salnikov V V, Hwang S, Martin K, Delmer DP. **2001**. Carbon partitioning to cellulose synthesis. *Plant Molecular Biology* **47**: 29–51.

Hall M, Bansal P, Lee JH, Realff MJ, Bommarius AS. **2011**. Biological pretreatment of cellulose: enhancing enzymatic hydrolysis rate using cellulose-binding domains from cellulases. *Bioresource Technology* **102**: 2910–2915.

Hansen SF, Bettler E, Rinnan A, Engelsen SB, Breton C. **2010**. Exploring genomes for glycosyltransferases. *Molecular Biosystems* **6**: 1773–1781.

Harholt J, Sørensen I, Fangel J, Roberts A, Willats WGT, Scheller HV, Petersen BL, Banks JA, Ulvskov P, Scheller V. **2012**. The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall. *PLoS One* **7**: 1–15.

Harris D, DeBolt S. **2010**. Synthesis, regulation and utilization of lignocellulosic biomass. *Plant Biotechnology Journal* **8**: 244–262.

Hefer C, Mizrachi E, Joubert F, Myburg A. **2011**. The *Eucalyptus* genome integrative explorer (EucGenIE): a resource for *Eucalyptus* genomics and transcriptomics. *BMC Proceedings* **5**: 1.

Henrissat B, Coutinho PM, Davies GJ. **2001**. A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Molecular Biology*: 55–72.

Henrissat B, Davies GJ. **2000**. Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiology* **124**: 1515–1519.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlén M, Teeri TT, Lundeberg J, *et al.* 2001. A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* **98**: 14732–14737.

Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N. **2009**. Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cellular & Developmental Biology - Plant* **45**: 619–629.

Horan K, Shelton CR, Girke T. **2010**. Predicting conserved protein motifs with sub-HMMs. *BMC Bioinformatics* **11**: 205–220.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al.* 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**: 211–215.

Jamet E, Albenne C, Boudart G, Irshad M, Canut H, Pont-Lezica R. **2008**. Recent advances in plant cell wall proteomics. *Proteomics* **8**: 893–908.

Jamet E, Canut H, Boudart G, Pont-Lezica RF. **2006**. Cell wall proteins: a new insight through proteomics. *Trends in Plant Science* **11**: 33–39.

Jansson S, Douglas CJ. **2007**. *Populus*: A model system for plant biology. *Annual Review of Plant Biology* **58**: 435–458.

Jin J, Riechmann L, Ferrier T. **2010**. *Arabidopsis* paves the way: genomic and network analyses in crops. *Current Opinion in Biotechnology* **22**: 1–11.

Joshi CP, Mansfield SD. **2007**. The cellulose paradox- simple molecule, complex biosynthesis. *Current Opinion in Plant Biology* **10**: 220–226.

Joshi CP, Thammannagowda S, Fujino T, Gou J-Q, Avci U, Harris D, Debolt S, Peter GF, Haigler CH, McDonnell LM, *et al.* 2011. Perturbation of wood cellulose synthesis causes pleiotropic effects in transgenic aspen. *Molecular Plant* **4**: 331–345.

Jung J, Kim S, Seo P, Park C. **2008**. Molecular mechanisms underlying vascular development. *Advances in Botanical Research* **48**: 1–68.

Keegstra K. **2010**. Plant cell walls. *Plant Physiology* **154**: 483–486.

Keegstra K, Talmadge KW, Bauer WD, Albersheim P. **1973**. The structure of plant cell walls. *Plant Physiology* **51**: 188–196.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. **2012**. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Evolution and Biology* **4**: 316–329.

Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, Ackerveken G Van Den, Snoek LB, Peeters AJM, Vreugdenhil D, Koornneef M, Jansen RC. **2007**. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences* **104**: 1708–1713.

Koch K. **2004**. Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology* **7**: 235–246.

Koestler T, Haeseler A Von, Ebersberger I, von Haeseler A. **2010**. FACT : Functional annotation transfer between proteins with similar feature architectures. *BMC Bioinformatics* **11**: 417–430.

Krogh A, Larsson È, Heijne G Von, Sonnhammer ELL. **2001**. Predicting transmembrane protein topology with a Hidden Markov Model: Application to complete genomes. *Journal of Molecular Biology* **305**: 567–580.

Krol J, Loedige I, Filipowicz W. **2010**. The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics* **11**: 597–610.

Kryvych S, Kleessen S, Ebert B, Kersten B, Fisahn J. **2010**. Proteomics- The key to understanding systems biology of *Arabidopsis* trichomes. *Phytochemistry*: 1–10.

Lairson LL, Henrissat B, Davies GJ, Withers SG. **2008**. Glycosyltransferases: structures, functions, and mechanisms. *Annual Review of Biochemistry* **77**: 521–555.

Lam BC-H, Blumwald E. **2002**. Domains as functional building blocks of plant proteins. *Trends in Plant Science* **7**: 544–549.

Lanot A, Hodge D, Jackson RG, George GL, Elias L, Lim E-K, Vaistij FE, Bowles DJ. **2006**. The glucosyltransferase UGT72E2 is responsible for monolignol 4-*O*-glucoside production in *Arabidopsis thaliana*. *The Plant Journal* **48**: 286–295.

Lee C, Teng Q, Huang W, Zhong R, Ye Z. **2009**. The Poplar GT8E and GT8F glycosyltransferases are functional orthologs of *Arabidopsis* PARVUS involved in glucuronoxylan biosynthesis. *Plant and Cell Physiology* **50**: 1982–1987.

Lee C, Zhong R, Ye Z-H. **2012**. Biochemical characterization of xylan xylosyltransferases involved in wood formation in Poplar. *Plant Signaling and Behaviour* **7**: 332–337.

Lerouxel O, Cavalier DM, Liepman AH, Keegstra K. **2006**. Biosynthesis of plant cell wall polysaccharides- a complex process. *Current Opinion in Plant Biology* **9**: 621–630.

Li H, Chiu C-C. **2010**. Protein transport into chloroplasts. *Annual Review of Plant Biology* **61**: 157–180.

Li X, Wu HX, Southerton SG. **2010**. Comparative genomics reveals conservative evolution of the xylem transcriptome in vascular plants. *BMC Evolutionary Biology* **10**: 1–14.

Littler SJ, Hubbard SJ. **2005**. Conservation of orientation and sequence in protein domain-domain interactions. *Journal of Molecular Biology* **345**: 1265–1279.

Lombard V, Ramulu HG, Drula E, Coutinho PM, Henrissat B. **2014**. The Carbohydrate-Active enZymes database (CAZy) in 2013. *Nucleic Acids Research* **42**: 490–495

Maloney VJ, Samuels AL, Mansfield SD. **2011**. The endo-1,4-β-glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytologist* **10**: 1–12.

Mansfield SD. **2009**. Solutions for dissolution- engineering cell walls for deconstruction. *Current Opinion in Biotechnology* **20**: 286–294.

McCartney L, Gilbert HJ, Bolam DN, Boraston AB, Knox JP. **2004**. Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers. *Analytical Biochemistry* **326**: 49–54.

Michel G, Tonon T, Scornet D, Cock JM, Kloareg B. **2010**. Central and storage carbon metabolism of the brown alga *Ectocarpus siliculosus*: insights into the origin and evolution of storage carbohydrates in Eukaryotes. *The New Phytologist* **188**: 67–81.

Minic Z. **2008**. Physiological roles of plant glycoside hydrolases. *Planta* **227**: 723–740.

Minic Z, Jamet E, San-Clemente H, Pelletier S, Renou J-P, Rihouey C, Okinyo DPO, Proux C, Lerouge P, Jouanin L. **2009**. Transcriptomic analysis of *Arabidopsis* developing stems: a close-up on cell wall genes. *BMC Plant Biology* **9**: 1–17.

Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA. **2010**. De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* **11**: 681–693.

Mizrachi E, Mansfield SD, Myburg AA. **2011**. Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54–62.

Mohnen D. **2007**. Pectin structure and biosynthesis. *Current Opinion in Plant Biology* **11**: 266–277.

Moorhead GBG, Wever VDE, Templeton G, Kerk D. **2009**. Evolution of protein phosphatases in plants and animals. *Biochemistry Journal* **417**: 401–409.

Myburg AA, Bradfield J, Cowley E, Creux N, De Castro M, Hatherell T, Mphahlele M, O Neill M, Ranik M. **2010**. Forest and fibre genomics: biotechnology tools for applied tree improvement. *Southern Forests* **70**: 59–68.

Myburg AA, Grattapaglia D, Tuskan G, Jenkins J, Schmutz J, Mizrachi E, Hefer C, Pappas G, Sterck L, Van De Peer Y, Hayes R, Rokhsar D. 2011. The *Eucalyptus grandis* Genome Project: Genome and transcriptome resources for comparative analysis of woody plant biology. *BMC Proceedings* **5** (Suppl7) :I20

Nicol F, His I, Jauneau A, Vernhettes S, Canut H, Höfte H. **1998**. A plasma membrane-bound putative endo-1,4-beta-D-glucanase is required for normal wall assembly and cell elongation in *Arabidopsis*. *The EMBO Journal* **17**: 5563–5576.

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K. **2009**. ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Research* **37**: 987–991.

Oikawa A, Lund CH, Sakuragi Y, Scheller H V. **2012**. Golgi-localized enzyme complexes for plant cell wall biosynthesis. *Trends in Plant Science* **18**: 49–48.

Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. **1997**. CATH- a hierarchic classification of protein domain structures. *Structure* **5**: 1093–10108.

Paiva JAP, Prat E, Vautrin S, Santos MD, San-Clemente H, Brommonschenkel S, Fonseca PGS, Grattapaglia D, Song X, Ammiraju JSS, *et al.* 2011. Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. *BMC Genomics* **12**: 137–150.

Palcic MM. **2011**. Glycosyltransferases as biocatalysts. *Current Opinion in Chemical Biology* **15**: 226–233.

Pauly M, Keegstra K. **2010**. Plant cell wall polymers as precursors for biofuels. *Current Opinion in Plant Biology* **13**: 305–312.

Pawar PM-A, Koutaniemi S, Tenkanen M, Mellerowicz EJ. **2013**. Acetylation of woody lignocellulose: significance and regulation. *Frontiers in Plant Science* **4**: 118–126.

Plomion C, Stokes A, Leprovost G. **2001**. Wood formation in trees. *Plant Physiology* **127**: 1513–1523.

Popper ZA, Fry SC. **2004**. Primary cell wall composition of pteridophytes and spermatophytes. *New Phytologist* **164**: 165–174.

Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB. **2011**. Evolution and diversity of plant cell walls: from algae to flowering plants. *Annual Review of Plant Biology* **62**: 567–90.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al.* 2012. The Pfam protein families database. *Nucleic Acids Research* **40**: 290–301.

Ray PM. **1967**. Radioautographic study of cell wall deposition in growing plant cells. *Journal Of Cell Biology* **35**: 659–675.

Reiter W-D. **2002**. Biosynthesis and properties of the plant cell wall. *Current Opinion in Plant Biology* **5**: 536–542.

Reker D, Katzenbeisser S, Hamacher K. **2010**. Computation of mutual information from Hidden Markov Models. *Computational Biology and Chemistry* **34**: 328–333.

Remmerie N, De Vijlder T, Laukens K, Dang TH, Lemière F, Mertens I, Valkenborg D, Blust R, Witters E, Vijlder T De, *et al.* 2011. Next generation functional proteomics in non-model plants: A survey on techniques and applications for the analysis of protein complexes and post-translational modifications. *Phytochemistry* **72**: 1192–1218.

Renuse S, Chaerkady R, Pandey A. **2011**. Proteogenomics. *Proteomics* **11**: 620–630.

Richmond TA, Somerville CR. **2000**. The cellulose synthase superfamily. *Plant Physiology* **124**: 495–498.

Roberts JA, Elliott KA, Gonzalez-Carranza ZH. **2002**. Abscission, dehiscence, and other cell separation processes. *Annual Review of Plant Biology* **53**: 131–158.

Roberts K, McCann MC. **2000**. Xylogenesis: the birth of a corpse. *Current Opinion in Plant Biology* **3**: 517–522.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP. **2012**. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research* **22**: 95–105.

Russell RB, Sasieni PD, Sternberg MJ. **1998**. Supersites within superfolds. Binding site similarity in the absence of homology. *Journal of Molecular Biology* **282**: 903–918.

Samuel MA, Salt JN, Shiu SH, Goring DR. **2006**. Multifunctional ARM repeat domains in plants. *International Review of Cytology* **253**: 1–26.

Sánchez-Rodríguez C, Bauer S, Hématy K, Saxe F, Ibáñez AB, Vodermaier V, Konlechner C, Sampathkumar A, Rüggeberg M, Aichinger E, *et al.* 2012. CHITINASE-LIKE1/POM-POM1 and its homolog CTL2 are glucan-interacting proteins important for cellulose biosynthesis in *Arabidopsis*. *The Plant Cell* **24**: 589–607.

Schaeffer RD, Daggett V. **2011**. Protein folds and protein folding. *Protein Engineering, Design & Selection* **24**: 11–19.

Scheible W-R, Pauly M. **2004**. Glycosyltransferases and cell wall biosynthesis: novel players and insights. *Current Opinion in Plant Biology* **7**: 285–295.

Scheller HV, Ulvskov P. **2010**. Hemicelluloses. *Annual Review of Plant Biology* **61**: 263–289.

Shakhnovich BE, Shakhnovich EI. **2008**. Improvisation in evolution of genes and genomes: whose structure is it anyway? *Current Opinion in Structural Biology* **18**: 375–381.

Silverstone AL, Tseng T-S, Swain SM, Dill A, Jeong SY, Olszewski NE, Sun T-P. **2007**. Functional analysis of SPINDLY in gibberellin signaling in *Arabidopsis*. *Plant Physiology* **143**: 987–1000.

Smith AM, Stitt M. **2007**. Coordination of carbon supply and plant growth. *Plant, Cell & Environment* **30**: 1126–1149.

Somerville C. **2006**. Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology* **22**: 53–78.

Somerville C, Youngs H, Taylor C, Davis SC, Long SP. **2010**. Feedstocks for lignocellulosic biofuels. *Science* **329**: 790–794.

Szyjanowicz PMJ, McKinnon I, Taylor NG, Gardiner J, Jarvis MC, Turner SR. **2004**. The *irregular xylem 2* mutant is an allele of korrigan that affects the secondary cell wall of *Arabidopsis thaliana*. *The Plant Journal* **37**: 730–740.

Taylor LE, Dai Z, Decker SR, Brunecky R, Adney WS, Ding S-Y, Himmel ME. **2008**. Heterologous expression of glycosyl hydrolases in planta: a new departure for biofuels. *Trends in Biotechnology* **26**: 413–425.

Van Tilbeurgh H, Loontiens FG, Engelborgs Y, Claeyssens M. **1989**. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414 cellobiohydrolase II and influence of glucose on their affinity. *European Journal of Biochemistry* **559**: 553–559.

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

Tsai AY-L, 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.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–15604.

Tyler L, Bragg JN, Wu J, Yang X, Tuskan GA, Vogel JP. **2010**. Annotation and comparative analysis of the glycoside hydrolase genes in *Brachypodium distachyon*. *BMC Genomics* **11**: 1–21.

Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. **2010**. Lignin biosynthesis and structure. *Plant Physiology* **153**: 895–905.

Vanholme R, Morreel K, Ralph J, Boerjan W. **2008**. Lignin engineering. *Current Opinion in Plant Biology* **11**: 278–285.

Verwoerd WS. **2011**. A new computational method to split large biochemical networks into coherent subnets. *BMC Systems Biology* **5**: 22.

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. **2004**. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology* **14**: 208–216.

Vogel JP, Raab TK, Schiff C, Somerville SC. **2002**. *PMR6* , a pectate lyase–like gene required for powdery mildew susceptibility in *Arabidopsis*. *The Plant Cell* **14**: 2095–2106.

Wang H, Guo Y, Lv F, Zhu H, Wu S, Jiang Y, Li F, Zhou B, Guo W, Zhang T. **2010**. The essential role of *GhPEL* gene, encoding a pectate lyase, in cell wall loosening by depolymerization of the de-esterified pectin during fiber elongation in cotton. *Plant Molecular Biology* **72**: 397–406.

Wang Y-W, Wang W-C, Jin S-H, Wang J, Wang B, Hou B-K. **2012**. Over-expression of a putative Poplar glycosyltransferase gene, *PtGT1*, in tobacco increases lignin content and causes early flowering. *Journal of Experimental Botany* **63**: 2799–808.

Ward JA, Ponnala L, Weber CA. **2012**. Strategies for transcriptome analysis in nonmodel plants. *American Journal of Botany* **99**: 267–276.

Weng J-K, Chapple C. **2010**. The origin and evolution of lignin biosynthesis. *New Phytologist* **187**: 273–285.

Wienkoop S, Baginsky S, Weckwerth W. **2010**. *Arabidopsis thaliana* as a model organism for plant proteome research. *Journal of Proteomics* **73**: 2239–2248.

Wu A-M, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, Marchant A. **2010**. Analysis of the *Arabidopsis IRX9/IRX9-L* and *IRX14/IRX14-L* pairs of glycosyltransferase

genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiology* **153**: 542–554.

Xie H, Gilbert HJ, Charnock SJ, Davies GJ, Williamson MP, Simpson PJ, Raghothama S, Fontes CMGA, Dias FM V, Ferreira LMA, *et al.* 2001. *Clostridium thermocellum* Xyn10B carbohydrate-binding module 22-2: The role of conserved amino acids in ligand binding. *Biochemistry* **40**: 9167–9176.

Yin Y, Huang J, Xu Y. **2009**. The cellulase synthase superfamily in fully sequenced plants and algae. *BMC Plant Biology* **14**: 1–14.

Yin Q, Teng Y, Ding M, Zhao F. **2011**. Site-directed mutagenesis of aromatic residues in the carbohydrate-binding module of Bacillus endoglucanase EGA decreases enzyme thermostability. *Biotechnology Letters* **33**: 2209–2216.

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* **479**: 1–7.

Yokoyama R, Nishitani K. **2004**. Genomic basis for cell-wall diversity in plants. A comparative approach to gene families in rice and *Arabidopsis*. *Plant & Cell Physiology* **45**: 1111–1121.

Yonekura-Sakakibara K, Hanada K. **2011**. An evolutionary view of functional diversity in family 1 glycosyltransferases. *The Plant Journal* **66**: 182–193.

Zhang D, Hrmova M, Wan C-H, Wu C, Balzen J, Cai W, Wang J, Densmore LD, Fincher GB, Zhang H, *et al.* 2004. Members of a new group of chitinase-like genes are expressed preferentially in cotton cells with secondary walls. *Plant Molecular Biology* **54**: 353–372.

Zimmermann P, Hirsch-hoffmann M, Hennig L, Gruissem W. **2004**. GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiology* **136**: 2621–2632.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H. **2009**. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genetics* **5**: 1–13.

Chapter 2

Comparative analysis of Carbohydrate Active enZyme domains in plants - Carbohydrate metabolism and wood formation in *Eucalyptus grandis*

D. Pinard[1], E. Mizrachi[1], C. Hefer[2, 4], A. Kersting[3], F. Joubert[2], A. A. Myburg[1]

[1] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20 Hatfield, Pretoria 0028, South Africa

[2] Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Private bag X20 Hatfield, Pretoria 0028, South Africa.

[3]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, Hufferstr. 1, D48149 Munster, Germany

[4]Department of Botany, University of British Columbia, 6270 University Blvd, Vancouver, BC V6T 1Z4 Canada

This chapter has been prepared and formatted as a research article manuscript for the journal *New Phytologist*. I performed all data analysis and manuscript preparation. Dr. E. Mizrachi provided supervision, direction and advice throughout the data analysis and manuscript preparation, and gave invaluable manuscript revisions. Prof. F. Joubert assisted in technical support and project supervision, and assisted in manuscript revisions. Dr. A. Kersting was involved in the initial domain analysis of the project. Dr. C. Hefer provided a large amount of technical assistance, knowledge, and the raw transcriptomic data used in this study. Prof. A. A. Myburg conceived of the study and provided supervision and manuscript revisions.

2.1. Summary

• Previous studies have suggested that the woody perennial *Populus trichocarpa* has greater Carbohydrate Active enZyme (CAZyme) diversity and expression in carbohydrate metabolism than the herbaceous annual *Arabidopsis thaliana*. A comprehensive study of CAZymes across more plant species, including a second woody perennial, *Eucalyptus grandis* is now possible.

• We investigated the genome-wide frequency of CAZyme domain families in plant species available on Phytozome. In addition, the diversity of CAZyme domains was analyzed for certain sequenced monocots, dicots and algae. Expressed CAZymes across six tissues in *E. grandis* were analyzed and the xylem and leaf CAZyme expression profiles in *E. grandis* and *P. trichocarpa* were compared.

• The relative abundance of CAZyme domains within plant species were similar, as was the diversity of CAZyme domain families in land plants, indicating that the retention of a basic suite of CAZymes is a feature of land plant evolution. Of the 2,542 predicted CAZyme domain-containing proteins in *E. grandis*, those related to cell wall biopolymer synthesis and cellular signaling showed higher expression in the immature xylem, a pattern that is conserved in *P. trichocarpa* xylem.

• Our results suggest that genomic potential to metabolize carbohydrates is similar among land plants. The expression level of cell wall biosynthetic CAZyme domain families within different tissues is what allows secondary cell wall biopolymers to be deposited in large amounts in woody perennials *E. grandis* and *P. trichocarpa*.

2.2. Introduction

Carbohydrate metabolism in plants is responsible for a diverse array of products involved in energy metabolism, signaling, defense and cell wall structure (Coutinho *et al.*, 2003) as well carbohydrate-related post-translational modifications (Wilson, 2002). Carbohydrate biopolymers in the secondary cell walls (SCWs) of fiber cells form the bulk of woody biomass, a valuable natural resource with a variety of industrial applications, including potential for biofuel production (Grattapaglia *et al.*, 2009; Hinchee *et al.*, 2009). The vessel and fiber cells in angiosperm wood have large amounts of cellulose, hemicellulose and lignin in their SCWs (Plomion *et al.*, 2001; Cosgrove, 2005). Cellulose and hemicelluloses are synthesized, modified, and degraded by Carbohydrate Active enZymes (CAZymes), a group comprising of modular protein domains that are ubiquitous across all living organisms (Hansen et al. 2010; Cantarel et al. 2009; Henrissat & Davies 2000). CAZymes have been classified into four classes of enzymatic domains, namely glycosyl transferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs) and carbohydrate esterases (CEs) (Henrissat & Davies 1997; Henrissat & Davies 2000) as well as the non-enzymatic class of carbohydrate-binding modules (CBMs) (Henrissat et al. 2001; van Tilbeurgh et al. 1986). Currently, the five CAZyme classes are collected and organized into families based on amino acid sequence similarity in the CAZy database (http://www.cazy.org/) (Cantarel *et al.*, 2009).

GTs catalyze glycosyl bonds between a donor sugar substrate and another molecule, typically another sugar (Lairson *et al.*, 2008). Along with defense, signaling and storage carbohydrate biosynthesis, plant GTs are responsible for the production of cellulose (GT2 domain family- *CESA* gene superfamily) (Dhugga, 2001) and hemicelluloses (GT2, GT8, 43, 47, and 61 families, among others) (Djerbi *et al.*, 2005; Lee *et al.*, 2009; Li *et al.*, 2011; Serapiglia *et al.*, 2011; Chiniquy *et al.*, 2012; Dhugga, 2012). GH domains hydrolyze the glycosyl bonds between sugars in carbohydrate biopolymers (Henrissat & Davies 2000). They

play an important role in the modification of biopolymers to be introduced into the cell wall, as well as abscission and dehiscence (Minic, 2008). PLs are implicated in non-hydrolytic cleavage of activated glycosidic bonds in pectin modification and degradation (Linhardt *et al.*, 1986; Garron & Cygler, 2010). CEs de-acetylate polysaccharide side-chains, and are thought to modify the ability of hemicellulose to cross-link with lignin (Cantarel *et al.*, 2009; Tsai *et al.*, 2012). CBMs facilitate specific binding to different carbohydrate biopolymers, allowing for the precise modification of these biopolymers by enzymatic domains as they are added to the cell wall (Boraston *et al.*, 2004; Hervé *et al.*, 2010). Due to their ability to disrupt the secondary cell wall network by binding to cell wall polymers, CBMs have been used in industry to increase the efficiency of cell wall degradation during the pulping process (Levy & Shoseyov, 2002).

Previous studies have shown that the genome of *P. trichocarpa* has a higher frequency and diversity of CAZyme genes than that of *A. thaliana* (Geisler-Lee *et al.*, 2006), which in 2001, had the most CAZymes in its genome compared to sequenced fungi and bacteria (Henrissat *et al.*, 2001). Furthermore, the CAZymes expressed in wood forming tissues of *P. trichocarpa*, specifically those involved in cellulose and hemicellulose biosynthesis, were more abundant and diverse than those in non-wood forming tissue such as the young leaves (Geisler-Lee *et al.*, 2006). Based on these findings, the authors noted the importance of CAZymes to the woody habit.

Protein domains, as the functional and evolutionary building blocks of plant proteins, are informative of the functional capacity of the genome (Nasir *et al.*, 2011; Kersting *et al.*, 2012). A recently published database of CAZyme domains, dbCAN (http://csbl.bmb.uga.edu-/dbCANdev/index.php) (Yin et al. 2012), can be utilized to identify the frequency and diversity of CAZyme domains in plant genomes available on Phytozome (http://www.phytozome.net/). dbCAN utilizes Hidden Markov Models (HMMs), based on the

seminal CAZyme family sequence data available (www.cazy.org), to accurately and reproducibly identify CAZyme domains (Yin *et al.*, 2012). Using the dbCAN database, protein coding genes containing CAZyme domains in plant species can be compared to analyze their CAZyme domain repertoire. In this study, twenty plant and two algal species (grouped together and hereafter referred to as twenty-two plant species) available from Phytozome v8.0 (http://www.phytozome.net/) with well annotated genes and domains (Dr A. Kersting, personal communication) were chosen to represent the diversity of Viridiplantae leading to the angiosperm lineage and wood-forming plant species.

With the availability of the genome of a second hardwood species, that of *E. grandis,* along with mRNA-Seq data for *E. grandis* and *P. trichocarpa* (Myburg et al., 2011, Mizrachi *et al*., 2010) (https://eucgenie.bi.up.ac.za/), we aimed to characterize the CAZyme domain frequency and diversity in plant species, and their expression levels in *P. trichocarpa* and *E. grandis* woody and non-woody tissues. By comparing the xylem and leaf transcriptomes of *E. grandis* and *P. trichocarpa*, we could identify the common expressed CAZyme repertoires involved in carbohydrate metabolism in wood forming tissues of two evolutionary divergent tree genera. Specifically, we asked: Does the frequency and diversity of CAZyme domains between plants reflect their evolution and developmental complexity? Is the expression of CAZyme domains related to wood formation in *E. grandis* and *P. trichocarpa*? We hypothesized that expression investment of CAZyme domain-containing genes expressed in the developing xylem would be higher than in non-xylogenic tissue as a reflection of focused carbohydrate metabolism in this sink tissue. This study is the most comprehensive analysis of genomic and expressed CAZyme domains in plant species, with a focus on the newly sequenced *E. grandis* genome. The identification of CAZyme domains involved in wood formation will aid the identification of target genes for increasing yield, and modifying carbohydrate structure and composition in trees for industrial purposes.

2.3. Materials and Methods

2.3.1. Genome-wide analysis of CAZyme domains in plant species

All CAZyme domains for twenty two plant species (Table 2.1) in Phytozome v8.0 (www.phytozome.net) available on dbCAN were obtained from the dbCAN database (http://csbl.bmb.uga.edu/dbCAN/) (Yin *et al*., 2012). The plant species examined for the genome-wide analysis of CAZyme domains were chosen in order to encompass the Viriplantae lineage (see Table 2.1 for all species and abreviations), including Chlorophyta (in cluding only *C. reinhardtii* and *V. carteri*), Embryophyta, encompassing *P. patens* onwards, Tracheophyta, encompassing *S. moellendorffii* onwards and monocot and dicot representatives of the Magniliophyta.

Analysis was performed using custom Python scripts (Python v2.6, Additional file 2.5 and Additional file 2.6) and Galaxy text manipulation tools (http://galaxyproject.org). Python is a programming language (Lutz, 2008) used in this study to write and execute custom scripts to rapidly, reproducibly and accurately analyze large tables of data. The primary applications of these scripts included basic data manipulation of the text files obtained from the dbCAN database, firstly extracting all the CAZyme domains present in each genome and collating them by domain family, and secondly, counting all the CAZyme domains in each family per genome. These collated and counted values of domain frequency per CAZyme domain family per genome were analyzed further in Excel. We classified three parameters, namely i. Frequency- the absolute numbers and relative frequencies of annotated genes within each of the five CAZyme domain classes, and the families assigned to these classes in the genomes of all twenty-two species, ii. Diversity- the number and type of individual CAZyme domain families within and between species and iii. Complexity- occurance, frequency and diversity of CAZyme domains *within* annotated genes. Covariance

analysis to determine within and between species CAZyme domain class relative frequency variation was done using SAS v9.3 (Statistical Analysis Software- SAS Institute Inc.).

Diversity of CAZyme domains were analyzed by grouping and counting all the individual domains present in each genome (including each domain in multidomain proteins) into their families based on dbCAN annotations. Complexity analysis was performed on a subset of ten species representing major lineages of land plant evolution (Table 2.1). The analysis of CAZyme domain complexity with annotated genes in each genome involved identifying all annotated genes that contained multiple CAZyme domains. Firstly, the number of annotated domains per gene in each of the ten genomes was calculated and visualized in Excel. Secondly, all genes containing multiple annotated CAZyme domains were separated based on whether they consisted soley of repeat CAZyme domains, or contained unique CAZyme domain families. These two different categories of multiple CAZyme domain containing annotated genes were then analyzed either by the frequency of the domain repeats, or by the combinations of unique domains they contained, and subsequently compared across species.

2.3.2. Gene expression analysis of CAZyme-coding genes in *E. grandis* and *P. trichocarpa*

In previous studies, next generation deep mRNA-sequencing using the Illumina platform was used to quantify the genome-wide expression in the transcriptomes of multiple tissues in *E. grandis* and *P. trichocharpa* (Hefer *et* al., 2011, Hefer *et al*., in preparation; Myburg *et al*., 2011; https://eucgenie.bi.up.ac.za/). Genome-wide transcriptome data for six tissues in *E. grandis* from Dr. C. Hefer was obtained for analysis of the transcript abundance and tissue specificity of all expressed genes (https://eucgenie.bi.up.ac.za/). The tissues analyzed in this study were: Young leaves, mature leaves, immature xylem, phoem, shoot tips, and flowers of *E. grandis* (Hefer *et al*., 2011; https://eucgenie.bi.up.ac.za/), as well as young leaves and immature xylem of *P. trichocarpa* (Hefer *et al*., in preparation). The expression levels of

every gene in each tissue/organ were averaged across three biological replicates, and filtered for genes containing CAZyme domains in *E. grandis* and *P. trichocarpa* from the dbCAN database (http://csbl.bmb.uga.edu/dbCAN/) for further analysis.

The transcript abundance of genes from mRNA-Seq can be quantified as Fragments Per Kilobase of exon per Million fragments mapped (FPKM) (Trapnell *et al*., 2009). FPKM parameters K and M are optimized to individual experiments in the software used to assemble the transcriptome, in this case Cufflinks (http://cufflinks.cbcb.umd.edu), was used (for more detail, refer to Trapnell *et al*., 2009). was used to infer the investment of expression of CAZyme domain families in each tissue. This was done by adding up the total transcript abundance for all genes in each CAZyme domain family and comparing that total to the FPKM expression investment values for the other tissues, using Excel for numerical comparisons and visualization. When calculating total expression investment of domain families, genes annotated with multiple CAZyme domain families were treated differently: If the gene was annotated as consisting soley of repeats of the same CAZyme domain, the total transcript abundance of the entire gene was added once to the CAZyme domain family total transcript abundance. Therefore repeat domains of the same CAZyme family were ignored when calculating CAZyme domain family specific transcript abundance. If the gene was annotated as having multiple domains from different CAZyme domain families, the transcript abundance of that gene was added seperately to each domain family once. For example, a gene annotated as having domains "X-X-Y", would have the FPKM value of the gene added once to "family X expression investment total", and once to "family Y expression investment total".

2.4. Results

2.4.1. Genome-wide analysis of CAZyme classes in plants

To determine the CAZyme domain content of twenty-two plant species that have been annotated in dbCAN from Phytozome, we examined the number of genes containing CAZyme domains and the frequency of CAZyme domains in these genes within each plant genome (Table 2.1). The frequencies of the five CAZyme domain classes were compared to give insight into the evolution of CAZyme genes in these plants. The average frequency of CAZyme domains is highest in dicot genomes (2,230) and lowest in the green algal lineages considered (693).

The absolute frequency of genes from each CAZyme class per genome shows that seed producing plants (except *Carica papaya*- 1, 341 CAZyme domains) have more CAZyme containing genes and CAZyme domains than non-seed organisms such as the bryophyte *Physcomitrella patens* (1, 519 CAZyme domains) and the lycophyte *Selaginella moellendorffii* (1, 476 CAZyme domains) and almost double that of the green algae species *Volvox carteri* and *Chlamydomonas reinhardtii* (654 and 731 CAZyme domains respectively) (Figure 2.1a., Table 2.1). However, the absolute frequencies of these genes in angiosperms can be deceptive, as some plant genomes have undergone whole genome duplications (WGD) and experienced extensive gene loss in the past (Freeling, 2009; Proost *et al.*, 2011). The absolute gene frequency may reflect the age of the genome since the last WGD and the rate of gene loss in the lineage, as well as other mechanisms such as tandem gene duplication (Maere *et al.*, 2005; Freeling, 2009; Proost *et al.*, 2011).

Although the absolute frequencies of CAZy domains vary between plant genomes, the proportions of the five functional classes of CAZymes are remarkably similar between species (Figure 2.1 a. and b). Coefficient of variance analysis was performed to determine if the ratios of CAZyme classes between monocots, eudicots, lycophytes and bryophytes, and green algae

varied significantly. For all CAZyme domain classes except PLs, the variance between the frequency ratios in the land plant (excluding the green algae) classes is negligible (Supplementary table 2.1). In land plants, GTs comprise roughly 40% of the CAZyme domain content in the genome, with GHs having a relative frequency of 30%. CEs, CBMs, and PLs have relative frequencies of 18, 10 and 2 percent respectively. In contrast, the green algae have frequency ratios of 57% for GTs, 14.5% for GHs, 0.5% for PLs, 10% for CEs and 17% for CBMs.

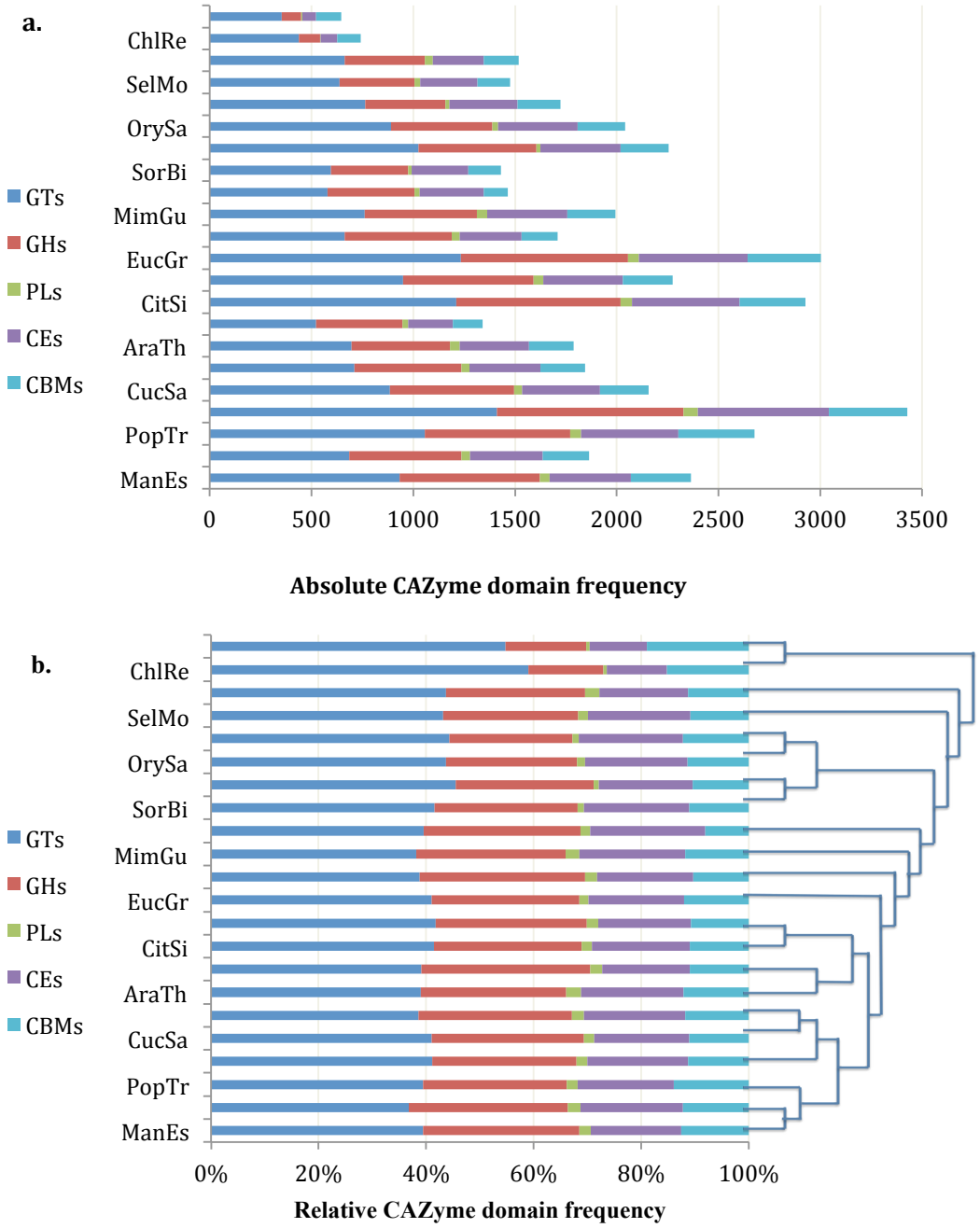


Figure 2.1 Absolute and relative frequency of CAZyme domain class frequency across twenty-two plant species. (a) Absolute frequency of CAZyme domains in five classes across twenty-two plant species. Plant species are on the y-axis, and the absolute frequency of CAZyme domains within all CAZyme genes is shown on the x- axis. The glycosyl transferase (GT) domain class is represented in blue, glycosyl hydrolase (GH) domain class in red, polysaccharide lyase (PL) domain class in green, carbohydrate esterase (CE) domain class in purple and carbohydrate binding module (CBM) domain class in light blue. (b) Relative frequency of CAZyme domain classes in twenty-two plant species. The relative frequency of carbohydrate active enzyme (CAZyme) domain classes in CAZyme genes, as a percentage, is shown on the x-axis. The species of plant is shown on the y-axis. For species abbreviation, refer to Table 2.1.

Table 2.1 Genome- wide CAZyme gene and domain content for twenty-two plant species.

Organism*	Genome Size (Mbp)	#Genes	#CAZyme genes	%CAZyme genes	#CAZyme domains	# GTs	# GHs	# PLs	# CEs	# CBMs	Reference
***Volvox carteri* (VolCa)**	138	14,520	490	3.37	645	283	85	4	65	102	(Prochnik *et al.*, 2010)
***Chlamydomonas reinhardtii* (ChlRe)**	121	15,143	574	3.79	741	367	85	4	76	87	(Merchant *et al.*, 2007)
***Physcomitrella patens* (PhyPa)**	480	35,938	1,236	3.44	1,519	664	392	41	250	172	(Rensing *et al.*, 2008)
***Selaginella moellendorffii* (SelMo)**	212	22,285	1,224	5.49	1,476	637	370	27	281	161	(Banks *et al.*, 2011)
***Brachypodium distachyon* (BraDi)**	272	25,532	1,418	5.55	1,723	764	394	20	334	211	(The International Brachypodium Initiative, 2010)
***Oryza sativa* (OrySa)**	420	42,109	1,724	4.09	2,040	891	498	29	389	233	(Goff *et al.*, 2002)
***Zea mays* (ZeaMa)**	2,300	30,579	1,920	6.28	2,256	1,026	578	22	394	236	(Schnable *et al.*, 2009)
***Sorghum bicolor* (SorBi)**	730	34,496	1,751	5.08	1,431	784	474	23	288	171	(Paterson *et al.*, 2009)
***Aquilegia coerulea* (AquCo)**	302	24,823	1,554	6.26	1,464	657	471	32	360	141	http://www.phytozome.net/aquilegia.php
***Mimulus guttatus* (MimGu)**	312	26,718	1,671	6.25	1,992	680	503	43	363	201	http://www.phytozome.org/mimulus.php
***Vitis vinifera* (VitVi)**	490	30,434	1,424	4.68	1,710	664	525	39	305	177	(Jaillon *et al.*, 2007)
***Eucalyptus grandis* (EucGr)**	641	36,376	2,542	6.99	3,334	1,233	823	54	534	360	(Myburg *et al.*, in prep)
***Citrus clementina* (CitCl)**	301	24,533	1,971	8.03	2,328	862	572	41	413	205	http://www.phytozome.net/clementine.php
***Citrus sinensis* (CitSi)**	319	25,376	2,439	9.61	2,927	1,049	735	52	498	267	http://www.phytozome.net/citrus.php
***Carica papaya* (CarPa)**	372	27,873	1,131	4.06	1,341	466	380	25	209	124	(Ming *et al.*, 2008)
***Arabidopsis thaliana* (AraTh)**	119	27,400	1,505	5.49	1,787	697	483	49	341	217	(The Arabidopsis Genome Initiative, 2000)
***Prunus persica* (PruPe)**	227	27,852	1,591	5.71	1,843	654	491	34	337	180	IPGI 2010. http://www.phytozome.org/peach
***Cucumis sativus* (CucSa)**	243	26,682	2,157	8.08	2,157	779	555	36	355	184	(Huang *et al.*, 2009)
***Glycine max* (GlyMa)**	975	54,175	2,839	5.24	3,429	1,412	917	69	645	386	(Schmutz *et al.*, 2010)
***Populus trichocarpa* (PopTr)**	422	41,335	2,252	5.45	2,677	1,057	713	55	479	373	(Tuskan *et al.*, 2006)
***Ricinus communis* (RicCo)**	350	31,237	1,540	4.93	1,864	605	486	36	328	186	(Chan *et al.*, 2010)
***Manihot esculenta* (ManEs)**	533	30,666	1,957	6.38	2,365	825	616	42	377	239	http://www.phytozome.net/cassava.php

*The first column shows the plants analysed with the abbreviations used in this study.

2.4.2. Genome-wide comparison of CAZy domain diversity and complexity

Next, we asked what the diversity of CAZyme families within each class was between plant species. The objectives for this part of the study were to analyze the CAZyme domain families within each broader functional class present in each genome to determine whether the presence of unique domains families contributed to the organismal complexity of seed producing vascular plants. All twenty-two species from the previous analysis were analyzed to determine the diversity of the domain families present in each species (Additional file 2.1). There are 231 different CAZyme domains present across the plant lineages analyzed (72 GT, 92 GH, 13 PL, 16 CE and 38 CBM families). *R. communis* (157) has a greater diversity of CAZyme domain families in its genome than any of the other species, followed by *P. patens* (148). There are 15 domain families that are unique to *R. communis* compared to the other 21 species analyzed, including 7 GHs, 3 CBMs, 3 GTs and 2 PLs (Additional file 2.1)

There are no unique domains in *A. thaliana* (Additional file 2.1), and there were no domains that were unique to the two woody perennials, *E. grandis* and *P. trichocarpa* compared to the other plant species analyzed. Of the 233 CAZyme domain families found across all 22 species, 65 are common to all, leaving 166 that are not present in all. The 166 domain families show varying levels of presence/absence between genomes that is more at the species-specific level rather than at the lineage specific level (Additional file 2.1). Of these, 108 (46% of total, 65% of domain families that are not ubiquitous) are not present in the green algae, *Chlamydomonas reinhardtii* and *Volvox carteri* (Additional file 2.1). CBM16, which binds cellulose and glucomannan, is only detected in vascular plants in the twenty-two species analyzed, but is also found in Archaea and Bacteria (http://www.cazy.org/CBM16.html).

The distribution of CAZy domain-containing multi-domain proteins in ten representative land plant species (five dicots, three monocots, lycophyte and bryophytes) followed the power law of gene complexity and gene number (Tordai *et al.*, 2005)

(Supplementary figure 2.1). The composition of complex CAZyme proteins was considered in terms of whether they consisted solely of repeat domains, or of combinations of different domains. CAZyme proteins containing repeat domains were found in all 10 genomes, representing the majority of complex proteins considered (60% of complex CAZyme proteins are repeats of a single domain) (Additional file 2.3). All CAZy domain-containing proteins that have five or more domains in all examined species contain GT41 domains, which are involved in glycan biosynthesis and signaling (Martinez-Fleites *et al.*, 2010; Breton *et al.*, 2012).

Annotated genes containing more than one CAZyme domain family show lineage specific combinations across the ten plant genomes analyzed. Supplementary figure 2.2 is a Venn diagram depicting the CAZyme domain family combinations that occur within the same annotated genes in the five eudicots studied. Of these, 15 CAZyme domain combinations are common to all five eudicots, as shown in the central over-lapping region of Supplementary figure 2.2. *Glycine max* has the most unique combinations between the eudicots at six (shown in pink on the figure), with *P. trichocarpa*, *E. grandis* and *V. vinifera* having two unique CAZyme domain family combinations each (shown in blue, yellow and green respectively).. *A. thaliana* has only one unique CAZyme domain family combination compared to the other five eudicots (shown in beige on Supplementary figure 2.2). Of the 28 CAZyme domain combinations that occur in *E. grandis*, the six that have genomic frequency >10 are: CBM43-GH17 (38), GH28-GH55 (21), CBM18-GH19 (16), CBM22-GH10 (16), CE1-CE10 (12) and CE1-CE7 (12) (Additional file 2.4). It is interesting to note that in *E. grandis*, CE domains only occur in combination with other CEs, and PLs are only found as repeats, never in combination. CBMs in combination are thought to act as enhancers and mediators of the enzymatic action of their appended domains. In the *E. grandis* dataset, this cooperative relationship is evident in the activity of the enzymatic domain and the specificity of the attached CBM. CBM43 protein domains bind to β-1,3-glucan, and the complementary GH17 protein domain is a β-1,3-endoglucanase (http://www.cazy.org/CBM43.html, http://www.cazy.org/GH17.html), similarly, the binding specificity of CBM22 is to xylan and

GH10 is a xylanase. Combinations with CBM domains are prevalent in the *E. grandis* genome, with CBM domain- enzymatic CAZyme domain combinations accounting for 11 of the 28 combinations.

2.4.3. Expression of CAZyme domain containing genes in *E. grandis*

mRNA-Seq expression profiling across six tissues in *E. grandis* showed that of the 2, 542 CAZyme domain containing genes in the *E. grandis* genome, 80.5% (2, 044) are expressed in at least one tissue (Additional file 2.2). The proportion of transcript abundance for each CAZyme domain class is very similar across tissues (Figure 2.2), although the expression of GH and GT domain classes are proportionally higher in the immature xylem. GTs constitute 44.5% of expression investment of CAZyme domain containing genes in the immature xylem vs. 35.9% in the young leaf, similarly, GHs account for 39.8% of transcript abundance of CAZyme domain containing genes in the immature xylem and 29.7% in the young leaf (Figure 2.2). CE domain family expression is proportionally lower in the phloem and immature xylem compared to the young leaf, mature leaf, flowers and shoot tips, making up 7% of the total CAZyme expression investment in the immature xylem and 20.3% in the young leaf. Variation at the level of individual CAZyme domain families was observed, and is discussed below.

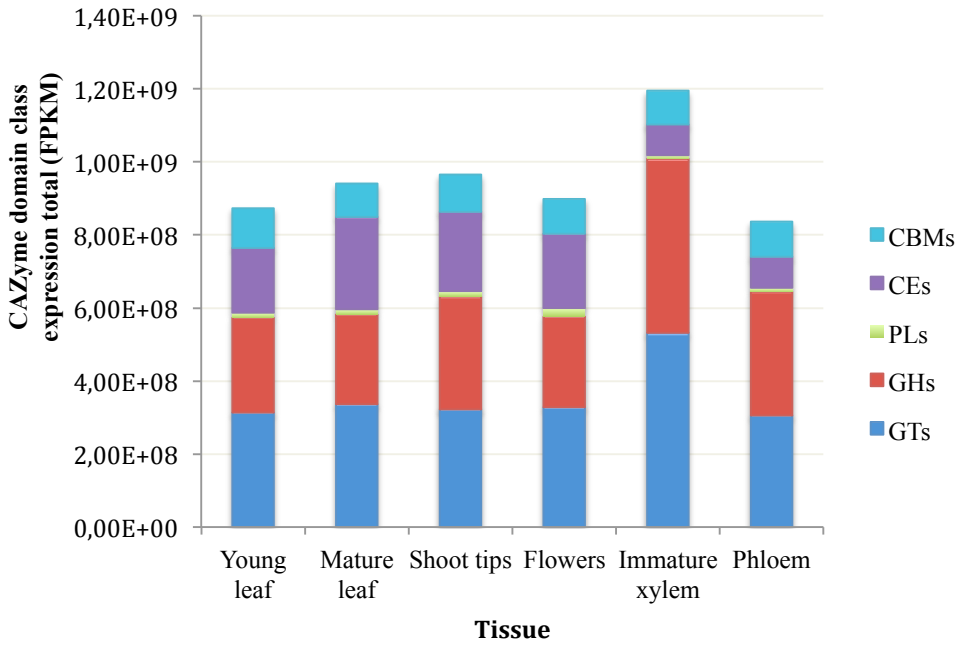


Figure 2.2 Total gene expression levels of five CAZyme domain classes across six tissues in *E. grandis*. The y-axis represents the mRNA-Seq expression data in FPKM, and the x-axis the tissue type analyzed. The average expression data in FPKM for each gene can be found in Additional data file 2.2.

The majority of GT domain families show fairly low levels of ubiquitous expression across six tissues in *E. grandis* tissues (Figure 2.3). Of the forty-seven GT domain families present in the *E. grandis* genome, eleven GT families are not expressed at the same level in all tissues. Together, the eleven families that have differential investment across tissues contribute 81% of the total average expression of GTs across all six tissues. GT1, GT2, GT4, GT8, GT14, GT31, GT41, GT43, GT47, GT65 and GT68 have total tissue specific expression investment that differs between tissue types (Figure 2.3). Of these, all except GT1 have greater expression investment in the immature xylem than in the other five tissues. GT41 has the highest expression investment across all tissues (Figure 2.3). GT41 proteins often contain repeats of the GT41 domain, and in *E. grandis* the gene Eucgr.L00641 contains 7 GT41 repeats and has the highest expression of all GT41 containing proteins at >6 million FPKM in the xylem (Additional file 2.2). The GT41 domain occurs 241 times in the *E. grandis* genome, of which 120 genes containing this domain are expressed in at least one tissue. In comparison,

GT1 occurs 511 times in the genome, of which 332 are expressed in at least one tissue and has lower expression investment in the immature xylem compared to the other five tissues analyzed. Thus GT1 domains are more prevalent in the genome, and more genes containing this domain are expressed, but the magnitude of expression of these genes is considerably lower than the less abundant GT41 domain-containing genes.

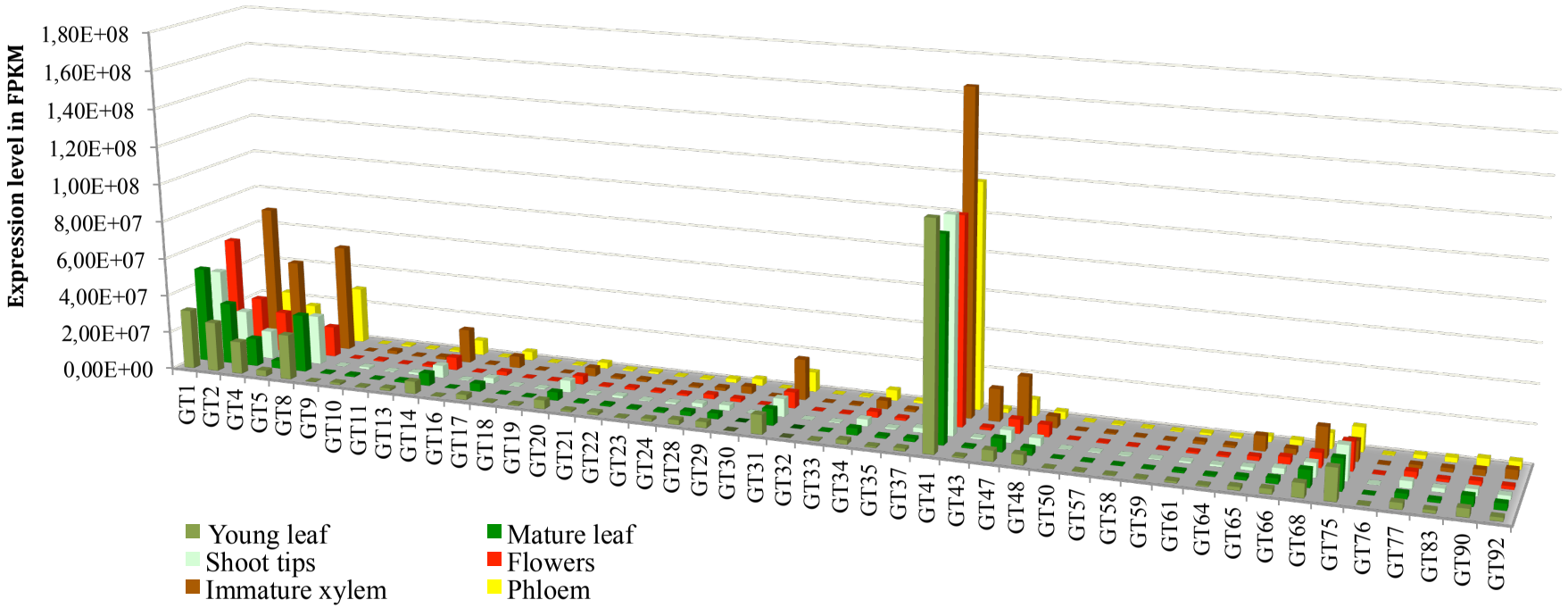


Figure 2.3 Total gene expression levels of GT domain families across six tissue types in *E. grandis*. The y-axis shows the total expression investment in FPKM from raw mRNA-Seq data summed across expressed glycosyl transferase (GT) genes, while the x-axis shows the GT domain family. The depth axis is the tissue type in *E. grandis* for which each domain family expression level in FPKM was calculated. Light green- young leaf, dark green- mature leaf, mint green- shoot tips, red- flowers, brown-immature xylem and yellow- phloem. The expression level in FPKM for each gene can be found in Additional file 2.2.

GH domain family expression investment across six tissues in *E.grandis* (Supplementary figure 2.3) showed that three GH domain families (GH9, GH16 and GH19) have relatively high levels of expression in different tissues. GH16 domain containing genes are highly expressed in the immature xylem and phloem, and GH19 domain containing genes are highly expressed in the immature xylem and shoot tips. The GH16 domain family is present in the xyloglucan endotransglycosylase/transferase (XTH) gene family, which can be involved in side chain hydrolysis or side chain rearrangement without hydrolysis (Eklöf *et al.*, 2013). GH9 domain families are highly expressed in the immature xylem compared to the other tissues, the overall higher expression investment in the immature xylem is due to fewer genes (18) being expressed at higher levels than in the other tissues, similar to the GT41 domain containing genes (Supplementary figure 2.3, Additional file 2.2). The most highly expressed CAZyme gene in *E. grandis* xylem is a GH19 family gene, *Eucgr.H04034* at 1,01E+08 FPKM (Additional file 1- Table 4). The *A. thaliana* ortholog *AT3G16920.1* is a chitinase-like (*CTL2*) gene which is known to be involved in cellulose synthesis (Sánchez-Rodríguez *et al.*, 2012).

PL families have very few CAZyme domain families (13) across all species, including the four expressed in *E. grandis*. The expression investment of PL domain families across six tissues in *E. grandis* shows that all four PL domain families are expressed at diverse levels in all tissues (Figure 2.3, Supplementary figure 2.4). PL1 and PL10 show high expression investment in the flowers compared to the other four tissues. There are no PL families that show high expression in woody tissues compared to non-woody tissues. CEs show interesting expression investment across six tissues in *E. grandis* in that they are expressed fairly ubiquitously in the same level across all tissues (Supplementary figure 2.5), leading to their lower proportional expression investment level in the immature xylem compared to GT and GH expression investment. Of the 12 CE domain families that are

expressed in the *E. grandis* genome, the exception to this pattern is CE16, which has low relative expression in the immature xylem and phloem, despite having the highest level of expression investment across the remaining four tissues. CE16 domain-containing genes are acetyl xylan esterases, which de-acetylate preferentially at the *O-3* and *O-4* positions of the backbone xylopyranosyl residues (Pawar *et al.*, 2013).

Most CBM domain families do not show tissue specific expression investment, they are expressed at the same (relative) level in 3 or more tissues (Supplementary figure 2.6). There are two exceptions to this expression pattern: CBM18, which is highly expressed in young leaf and shoot tips compared to the other four tissues, and CBM22, which is highly expressed in the immature xylem compared to the other five tissues. CBM18 has the highest expression investment in young leaf of all the CBM domains expressed in *E. grandis*. CBM57 has the highest expression investment of all the CBM domains expressed in the mature leaf, immature xylem and the shoot tips, while CBM43 has the highest expression investment in the flowers and phloem. CBM43 and CBM57 together contribute 51% of the total average expression investment out of 17 CBM domain families in all tissues.

2.4.4. Comparative expression investment of CAZyme domains in *E. grandis* and *P. trichocarpa*

To analyze the expression investment of CAZyme domain families in two divergent tree species, we compared the transcript abundance of CAZyme domain families in xylem and leaf tissues of *E. grandis* and *P. trichocarpa* (Additional file 2.2, Hefer *et al*., 2011; Hefer *et al*., in prep). The pattern of CAZyme family expression was said to be similar if the expression level of the genes containing that CAZyme domain family was comparatively higher or lower between the two tissue types in both species. i.e. if the *E. grandis* expression level for GTX is visibly higher/lower in the xylem than in the leaf, and the *P. trichocarpa* expression level for GTX is also visibly higher/lower in the xylem than in the leaf, the expression pattern was said to be the same. The absolute transcript abundance in FPKM

cannot be directly compared between these analyses as the experiments were conducted independently, with gene length (K) and sequence depth (M) parameters normalized within the individual transcriptomes for each species (Charles Hefer, personal correspondence).

For the GT family of CAZyme genes, the expression pattern is similar in *E. grandis* and *P. trichocarpa*. The majority of GT domain families are expressed at a low level in *E. grandis* and *P. trichocarpa* xylem and leaf tissue, which indicates that they are involved in other aspects of carbohydrate metabolism, rather than cell wall biosynthesis (Figure 2.4). GT1 family shows higher expression investment in the leaf tissue as opposed to the xylem tissue in both *E. grandis* and *P. trichocarpa*. The GT domain families identified in this study as having higher expression investment in the immature xylem compared to the other five tissues in *E. grandis* (Figure 2.3) show greater expression investment in both *E. grandis* and *P. trichocarpa* xylem compared to leaf (Figure 2.4). These mainly include the domain families that have been implicated in cellulose and hemicellulose biosynthesis, namely GT2, GT4, GT8, GT14, GT31, GT43, GT47, GT65 and GT68. The conservation of these expression investment patterns between xylogenic and non-xylogenic tissues of divergent species indicates a conserved mechanism for cell wall biosynthesis at a functional domain level.

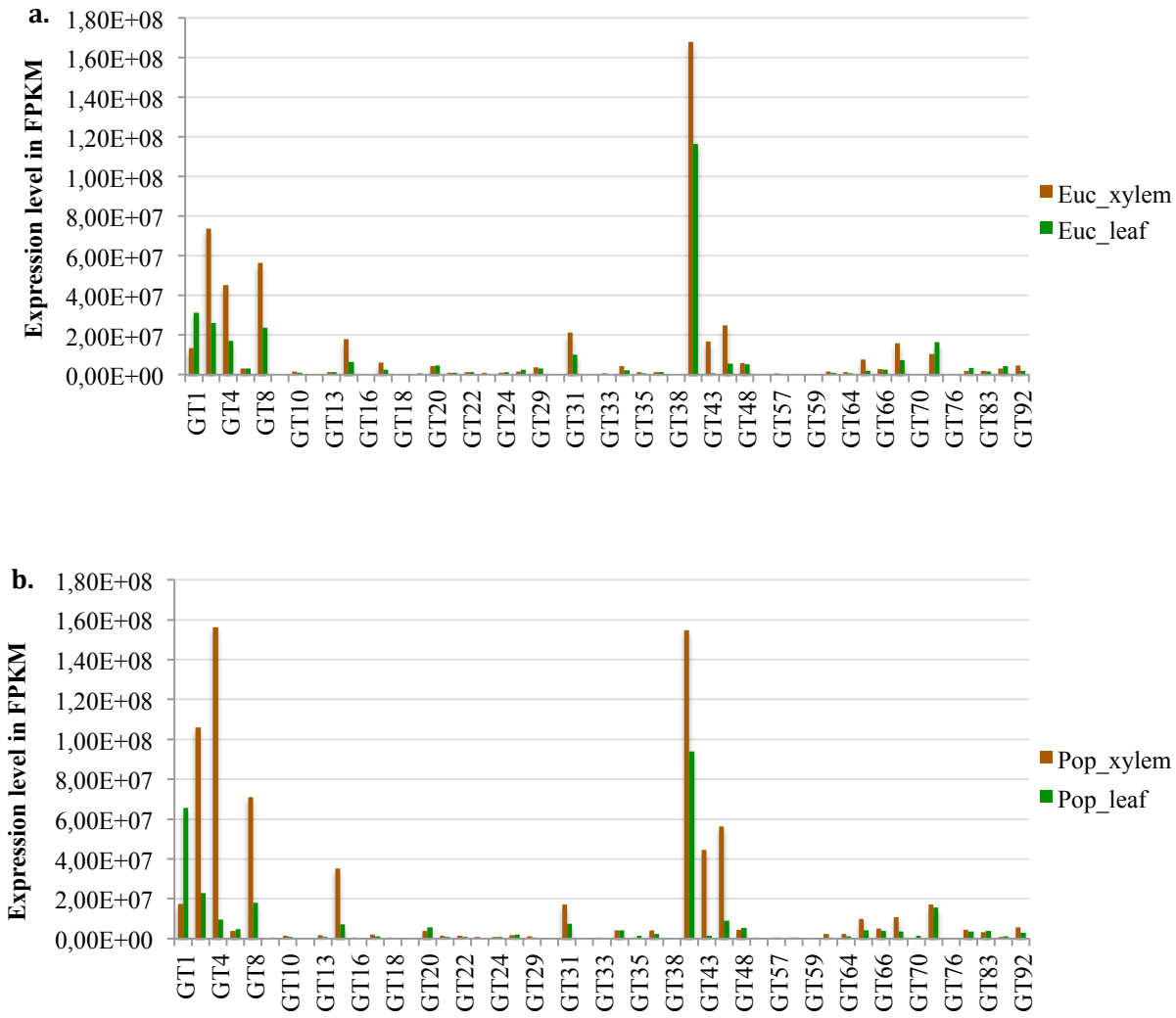


Figure 2.4 Total gene expression levels of GT domain families in *E. grandis* and *P. trichocarpa* xylem and leaf tissues. (a) Expression level per glycosyl transferase (GT) family in *E. grandis* in xylem and leaf tissues. The y-axis shows the transcript abundance in FPKM, the x-axis shows the GT family with xylem in brown, and leaf in green. (b) GT domain family (x-axis) expression level for *P. trichocarpa* xylem and leaf tissues in FPKM (y-axis).

The expression pattern of GH domains in *E. grandis and* P. trichocarpa is similar for most of the GH domain families (Supplementary figure 2.7). GH4 family is not expressed in the *E. grandis* tissues studied, including the xylem, while it is expressed at relatively low levels in the xylem and leaf tissue of *P. trichocarpa*. Furthermore, GH57, 62 and 80 are not expressed in *P. trichocarpa* xylem and leaf, while they are expressed at low levels in *E. grandis*. GHs that are expressed at low levels in one species and not in the other may be involved in specific defense or response to abiotic factors, and are thus not captured in a tissue transcriptome of either species. As with *E. grandis*, the most highly expressed CAZyme in the immature xylem of *P. trichocarpa* is a CTL2 homolog, POPTR_0010s15150 (Additional file 1- Table 3).

For PL family expression between *E. grandis* and *P. trichocarpa* xylem and leaf tissue (Supplementary figure 2.8), PL1 has a higher expression investment in the xylem as compared to the leaf in both species, with X:L ratios of 1.9 and 7.1 in *E. grandis* and *P. trichocarpa* respectively. The same four PL domain families are present in the genome and expressed in both *E. grandis* and *P. trichocarpa*. The CE domain family shows variable expression investment patterns between xylem and leaf tissues for *E. grandis* and *P. trichocarpa* (Supplementary figure 2.9). CE2, 3 and 5 are not expressed in *E. grandis* and are expressed at relatively low levels in *P. trichocarpa* xylem and leaf. CE15 is expressed in *E. grandis* and not in *P. trichocarpa*. The CE8 domain family contains pectin methylesterase genes (Jolie *et al.*, 2010), and is more highly expressed in *P. trichocarpa* xylem than in the leaf, while *E. grandis* shows the opposite trend (Supplementary figure 2.9). CE16 has the highest expression investment in the leaf tissue of all the CE domain families expressed in both species. CBMs show different expression investment patterns between *P. trichocarpa* and *E. grandis* xylem and leaf tissues in a number of families (Supplementary figure 2.10). CBM18 shows the highest expression investment in the leaf of all the CBMs expressed in *E. grandis* xylem and leaf, while CBM57 has the highest expression level in xylem compared to the expressed CBMs in *E. grandis* xylem and leaf. In *P. trichocarpa*, CBM47 shows the

highest expression in both xylem and leaf of all the CBM domain families expressed (Supplementary figure 2.10 a and b).

2.5. Discussion

We find that the genomic content of CAZyme domains in evolutionary diverse plant genomes is conserved with respect to the ratios of GTs, GHs, CEs, PLs and CBMs, although the absolute frequencies vary (Figure 2.1). This result is surprising in that our hypothesis, based on previous findings (Geisler-Lee *et al.*, 2006), was that woody perennials would have a larger proportion of GTs for the carbohydrate metabolism needed for wood production. The reality is that the land plants analyzed show a genomic ratio of 40:30 percent of GTs to GHs, regardless of their relative investment in different types of carbohydrate metabolism. This result is explained when the literature regarding functional domain conservation across species is examined. A study that compared the main functional domain classes across thousands of species across all forms of life found that the ratio of functional domain classes was highly conserved, indicating that the integrity of cellular functionality is maintained by a ratio of functional domains within each organism (Nasir *et al.*, 2011).

Importantly, when considering the maintenance of the ratios of the different CAZyme classes within the genomes between plant species, we observe that the ratio of functional enzymatic domains is maintained, despite high levels of tandem and segmental duplications in plant genomes due to the gene dosage balance model (Veitia *et al.*, 2008; De Smet & Van de Peer, 2012).The retention of duplicated genes after polyploidization or tandem duplication is based on the position of the protein product of that gene within a network where the interaction of proteins is dosage sensitive. Genes within large biochemical networks where stoichiometry needs to be maintained are rarely retained if they do not undergo neofunctionalization or become pseudogenes (Arrigo & Barker, 2012). As CAZymes involved in polysaccharide biosynthesis form parts of complex interacting networks that control the carbon flux within plants, encompassing primary and secondary metabolic

networks, gene dosage is a likely explanation for the maintenance of functional CAZyme domain class ratios.

The genome of *Glycine max* has undergone multiple WGD events, and as such almost 75% of genes in this species are duplicated (Schmutz et al,. 2010). The large amount of recently duplicated genes in *G. max* (Cannon & Shoemaker, 2012) is the likely reason why it has 2839 CAZyme containing genes, compared to the average of 1784 CAZyme genes in angiosperm genomes. There is evidence of a correlation between the number of genes in a genome and the number of CAZyme domain containing genes (specifically GHs and GTs) (Coutinho *et al.*, 2003), which is supported by the results obtained in this study, the number of CAZyme containing genes per total number of genes in the genome of angiosperms is between 4% and 9% (Table 2.1).

From examining the presence of unique domains in the sequenced genomes of twenty-two plant species, we can conclude that the genomic potential to metabolize carbohydrates to form wood is apparently not associated with to the emergence of unique CAZyme domain families (Kersting *et al*., in preparation). The fact that *P. patens*, despite having a relatively low amount of CAZyme genes, has a larger diversity of CAZyme domains in its genome than vascular plant species, indicates that primary and secondary cell wall metabolism utilizes a standard set of CAZyme domains between different tissue types in land plant species. The proportionally higher frequency of GT domains in the green algal genomes compared to GH, PL and CE domains reflects the minimal need of non-vascular green algae to modify cell wall biopolymers after synthesis compared to land plants (Popper *et al*., 2011; Leliaert *et al*., 2012). Within woody species, the existing CAZyme domain family diversity of vascular plants may contribute to wood formation via unique combinations and regulatory mechanisms of ancestral domains within the genomic and transcriptomic context. We have found that unique combinations of CAZyme domains does not differentiate woody plants from non-woody plants, as the majority of the types of combinations that CAZymes make in

complex proteins are common between lineages, with low promiscuity of domains (Additional file 2.4). This result is expected as protein domain promiscuity has been found to be highest in proteins involved in protein-protein interactions and chromatin and ubiquitin signaling (Basu *et al.*, 2008).

An evolutionary mechanism that may have played an important role in the emergence of wood development is the sub-functionalization of CAZyme domains after WGD. A study by Yokoyama (2010), suggests that vascular plants and bryophytes have undergone independent diversification of ancestral gene families. The study examined the frequency and diversity of the GH class xyloglucan endo-transglycosylase/hydrolase (XTH) genes in *P. patens* compared to *A. thaliana* and *O. sativa*. They found that *P. patens* had similar diversity and frequency of XTHs compared to the angiosperms, although further biochemical analysis uncovered functional diversity between bryophyte specific XTHs compared to angiosperm XTHs (Yokoyama *et al.*, 2010). The bryophyte XTHs demonstrated distinct spatial and temporal distribution patterns and different hormone responsiveness compared to those in angiosperms (Yokoyama *et al.*, 2010), highlighting the importance of understanding the functional differences in CAZyme families that have diverged between lineages in response to unique evolutionary pressures. Another interesting example of this maintenance of cell wall biosynthetic mechanisms can be seen in the green algae, where the multicellular *V. carteri* and the unicellular *C. reinhardtii* share similar domains and domain combinations, including those for cell wall biosynthesis (Prochnik *et al.*, 2010).

PL domain family diversity was shown to be highly variable across plant species, with *E. grandis*, *P. trichocarpa* and *S. bicolor* having the same four PL families present in their genomes (Additional file 2.1), all of which were expressed in *P. trichocarpa* and *E. grandis*. The PL families cannot be clustered by their presence or absence across plant evolution or their frequency within those genomes. Of the four PL families expressed in *E. grandis* and *P. trichocarpa*, PL1 is a pectin lyase that is known to be highly expressed in

wood development in *Populus* and *Eucalyptus* (Mellerowicz & Sundberg, 2008; Goulao *et al.*, 2011) (Supplementary figure 2.8). This expression is likely related to primary cell wall pectin remodeling (Palin & Geitmann, 2012), rather than SCW deposition, as SCWs are known to have less pectin than PCWs (Ishii, 1997; Cosgrove & Jarvis, 2012).

The results of the expression analysis show that the CAZymes in the *E. grandis* genome are almost always expressed in at least one tissue, and across tissues at a low level (80.5%) (Additional file 2.2). This suggests that either the remaining 19.5% of CAZyme genes in the *E. grandis* genome are pseudogenes, or the more likely option; they are expressed in response to biotic and abiotic stresses, or at other stages of development, not sampled in this study. As an example, GT domain containing gene UGT74E2 in *A. thaliana* is expressed in response to oxidative stress, and mediates drought response and plant architecture via auxin glycosylation (Tognetti *et al.*, 2010). GH domains and CBM domains have been implicated in pathogen response in plants, especially GHs and CBMs that recognize and respond to chitin (Kawabata *et al.*, 2000; Minic, 2008).

We identify domains that are known to be crucial in cell wall polysaccharide biosynthesis by examining the expression investment patterns across and between species, highlighting the capacity to identify functional domain families from a global comparative approach using next generation mRNA-Seq technology. CE domain containing genes have low transcript abundance in the immature xylem compared to the other five tissues studied in contrast to GT and GHs (Figure 2.2). CE16 domain family has a leaf to xylem (L:X) expression ration of 15.8 in *E. grandis* and 13 in *P. trichocarpa*, contributing to the decreased proportion of CE expression in the xylem of both species (Supplementary figure 2.9). CE16 domain containing acetylesterases have been shown to remove acetyl groups from xylo-, gluco- and manno-oligosaccharides in *Hypocrea jecorina* (Li et al. 2008). Given their function in removing side chains from hemicelluloses and pectin, decreasing cross linking to lignin; the relatively stable expression of CEs in the immature xylem with increased

expression investment of GTs and GHs may contribute to the physiochemical properties of vessel and fiber cells of wood. CE8 has a differential expression pattern in *P. trichocarpa* and *E. grandis* xylem and leaf tissues (Supplementary figure 2.5). Pectin methylesterase (PME) genes are found in this CAZyme family, and they function to remove methyl ester groups from the homogalacturonan (HG) backbone of pectin, affecting the gel properties of the cell wall (Jolie *et al.*, 2010). The difference in expression is a possible contributing factor to the differences in wood properties between *Populus* and *Eucalyptus*.

The most highly expressed genes in *E. grandis* immature xylem and *P. trichocarpa* xylem are GH19 domain containing genes, Eucgr.H04034 (Additional file 2.1- Table 4) and POPTR_0010s15150 (Additional file 2.2, Hefer *et al*., 2011; Hefer *et al*., in prep). *CTL2*; along with its homolog *CTL1,* modify cellulose microfibrils as they are extruded, shown by the reduced levels of crystalline cellulose in double knockdown mutants of *ctl1/ctl2* (Sánchez-Rodríguez *et al.*, 2012). *CTL2* has previously been shown to be a part of the SCW regulatory network in *E. grandis* (Hussey *et al.*, 2011), highlighting the importance of domain families responsible for degradation acting as modifiers to the synthesis of the SCW.

GTs known to synthesize cellulose and hemicellulose show greater expression investment in the immature xylem compared to the other tissues in *E. grandis* (Figure 2.3). Furthermore, these GT domain families show conserved expression patterns in *E. grandis* and *P. trichocarpa* xylem and leaf (Figure 2.4), identifying the importance of conserved biosynthetic mechanisms at the domain level. This pattern of conserved domain expression investment in xylem is seen in GT2, GT4, GT8, GT14, GT31, GT41, GT43, GT47, GT65 and GT68 domain families. GT41 family genes are GlcNAc transferases, involved in a multitude of functions, predominantly intracellular signaling (Breton *et al.*, 2012). Signaling is an important cellular mechanism, providing the sensitive feedback necessary to coordinate the deposition of cell wall polysaccharides. GT41 mediated modification of proteins can be compared to phosphorylation, as it is dynamic method of post-translational modification for

cytoplasmic and nuclear proteins. GT41 domain containing proteins are also the most complex CAZymes, with greater than four GT41 domain repeats within a single gene found across all plant species analyzed. GT4 domain containing genes include the Sucrose synthase (SuSy) genes, which are involved in the synthesis of UDP-glucose to cellulose synthase complexes (Haigler *et al.*, 2001). As expected, GT2 domain family showed higher expression in the xylem than in non-xylogenic tissues and conservation between *E. grandis* and *P. trichocarpa*.

GT43 (*IRX*9 and *IRX*14) (Lee *et al.*, 2012a) and GT47 (e.g. *fragile fiber 8*) (Doering *et al.*, 2012) are known to be involved in xylan biosynthesis. GT43 gene family members responsible for xylan backbone biosynthesis have been shown to have conserved biochemical functions across vascular plants (Lee, Zhong & Z.-H. Ye 2012b). GT8 domain family containing genes have high expression investment in the immature xylem compared to the other tissues analyzed, members of this gene family have been characterized as xylan glucuronyl transferases, including PARVUS and GAUT/GATL genes (galacturonosyltransferase1) (Yin *et al.*, 2010; Rennie *et al.*, 2012). GT31 domain containing gene *At4g21060* in *A. thaliana* has recently been shown to be a galactosyltransferase that is responsible for the galactosylation of arabinogalactan proteins during backbone formation (Basu *et al.*, 2013). GT65 and GT68 are fucosyl and oligosaccharide transferases respectively (www.cazy.org). GT and GT-like enzymes accounted for 20% of the proteome of *Arabidopsis* Golgi apparatus in seven day old protoplasts, including GT14, GT8 and GT31 domain containing genes, showing that these CAZyme domains are translated after being expressed during primary cell wall biosynthesis (Parsons *et al.*, 2012), strengthening the argument for using expression investment as a proxy for functional importance.

This study provides a functional overview of CAZyme domains in the genomes of twenty-two plant species. There are many CAZyme domain families that have gene members with unknown functions, and domain analysis can give insight into their primary function

(synthesis, degradation, modification). Along with comparative genomic data and transcriptomic data, we can identify domain families that are important for wood formation. Further analysis into CAZyme domain containing genes that differ in their expression pattern and level of expression between *E. grandis* and *P. trichocarpa* could elucidate the enzymes that contribute to the micro-heterogeneity that exists between carbohydrate biopolymers in the two species (Burton *et al.*, 2010).

The main result of this study is that the CAZyme containing genes in plant genomes have a conserved ratio between species, regardless of their organizational complexity. Although we find evidence for lineage specific diversity of CAZyme families in plant genomes, the domain family diversity of CAZymes cannot be used to discriminate the eudicot and monocot lineages. The expression investment pattern of the CAZyme domains responsible for cellulose and xylan biosynthesis are conserved between two divergent woody species; comparisons between other plant species transcriptomes may reveal that this is a defining characteristic of plant evolution. This study highlights the importance of transcriptional regulation in wood development as opposed to genomic innovations in the enzymatic domains responsible for carbohydrate metabolism.

2.6. References

Arrigo N, Barker MS. **2012**. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* **15**: 140–146.

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, DePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, *et al.* 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**: 960–963.

Basu MK, Carmel L, Rogozin IB, Koonin E V. **2008**. Evolution of protein domain promiscuity in eukaryotes. *Genome Research* **18**: 449–461.

Basu D, Liang Y, Liu X, Himmeldirk K, Faik A, Kieliszewski M, Held M, Showalter AM. **2013**. Functional identification of a hydroxyproline-*O*-galactosyltransferase specific for arabinogalactan protein biosynthesis in *Arabidopsis*. *The Journal of Biological Chemistry*: 1–25.

Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. **2004**. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemistry* **382**: 769–781.

Breton C, Fournel-Gigleux S, Palcic MM. **2012**. Recent structures, evolution and mechanisms of glycosyltransferases. *Current Opinion in Structural Biology* **22**: 540–549.

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR. **2009**. Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in *Arabidopsis*. *The Plant Journal* **57**: 732–746.

Burton RA, Gidley MJ, Fincher GB. **2010**. Heterogeneity in the chemistry, structure and function of plant cell walls. *Nature Chemical Biology* **6**: 724–732.

Caetano-Anollés G, Nasir A. **2012**. Benefits of using molecular structure and abundance in phylogenomic analysis. *Frontiers in Genetics* **3**: 172.

Cannon SB, Shoemaker RC. **2012**. Evolutionary and comparative analyses of the soybean genome. *Breeding Science* **61**: 437–444.

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**: 233–238.

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, *et al.* 2010. Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology* **28**: 951–6.

Chiniquy D, Sharma V, Schultink A, Baidoo EE, Rautengarten C, Cheng K. **2012**. XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan. *Proceedings of the National Academy of Sciences* **109**: 17117–17122.

Cosgrove DJ. **2005**. Growth of the plant cell wall. *Nature* **6**: 850–861.

Cosgrove DJ, Jarvis MC. **2012**. Comparative structure and biomechanics of plant primary and secondary cell walls. *Frontiers in Plant Science* **3**: 1–6.

Coutinho PM, Stam M, Blanc E, Henrissat B. **2003**. Why are there so many carbohydrate-active enzyme-related genes in plants ? *Trends in Plant Science* **8**: 563–565.

Darley CP, Forrester AM, McQueen-Mason SJ. **2001**. The molecular basis of plant cell wall extension. *Plant Molecular Biology* **47**: 179–195.

Demura T, Fukuda H. **2007**. Transcriptional regulation in wood formation. *Trends in Plant Science* **12**: 64–70.

Dhugga KS. **2001**. Building the wall: genes and enzyme complexes for polysaccharide synthases. *Current Opinion in Plant Biology* **4**: 488–493.

Dhugga KS. **2012**. Biosynthesis of non-cellulosic polysaccharides of plant cell walls. *Phytochemistry* **74**: 8–19.

Djerbi S, Aspeborg H, Schrader J, Coutinho PM, Stam M, Nilsson P, Denman S, Amini B, Sterky F, Master E, *et al.* 2005. Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid Aspen. *Plant Physiology* **137**: 983–997.

Doering A, Lathe R, Persson S. **2012**. An update on xylan synthesis. *Molecular Plant* **5**: 769–771.

Eklöf JM, Shojania S, Okon M, McIntosh LP, Brumer H. **2013**. Structure-function analysis of a broad specificity Populus trichocarpa endo-β-glucanase reveals an evolutionary link between bacterial licheninases and plant XTH gene products. *The Journal of Biological Chemistry* **288**: 15786–15799.

Freeling M. **2009**. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60**: 433–453.

Garron M-L, Cygler M. **2010**. Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. *Glycobiology* **20**: 1547–1573.

Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, *et al.* 2006. Poplar Carbohydrate-Active Enzymes. Gene identification and expression analyses. *Plant Physiology* **140**: 946–962.

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science* **296**: 92–100.

Goulao LF, Vieira-Silva S, Jackson PA. **2011**. Association of hemicellulose- and pectin-modifying gene expression with *Eucalyptus globulus* secondary growth. *Plant Physiology and Biochemistry* **10**: 1–9.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR. **2009**. Genomics of growth traits in forest trees. *Current Opinion in Plant Biology* **12**: 148–156.

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, Potts BM, Myburg AA. **2012**. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genetics & Genomes* **8**: 463–508.

Haigler CH, Ivanova-Datcheva M, Hogan PS, Salnikov V V, Hwang S, Martin K, Delmer DP. **2001**. Carbon partitioning to cellulose synthesis. *Plant Molecular Biology* **47**: 29–51.

Hansen SF, Bettler E, Rinnan A, Engelsen SB, Breton C. **2010**. Exploring genomes for glycosyltransferases. *Molecular Biosystems* **6**: 1773–1781.

Harholt J, Sørensen I, Fangel J, Roberts A, Willats WGT, Scheller HV, Petersen BL, Banks JA, Ulvskov P, Scheller V. **2012**. The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall. *PLoS One* **7**: 1–15.

Henrissat B, Coutinho PM, Davies GJ. **2001**. A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Molecular Biology* **47**: 55–72.

Henrissat B, Davies G. **1997**. Structural and sequence-based classification of glycoside hydrolases. *Carbohydrates and Glycoconjugates* **7**: 637–644.

Henrissat B, Davies GJ. **2000**. Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiology* **124**: 1515–1519.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlén M, Teeri TT, Lundeberg J, *et al.* 2001. A transcriptional roadmap to wood formation *Proceedings*. *of the National Academy of Sciences* **98**: 14732–14737.

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.

Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N. **2009**. Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cellular & Developmental Biology - Plant* **45**: 619–629.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, *et al.* 2009. The genome of the cucumber, *Cucumis sativus L*. *Nature Genetics* **41**: 1275–1281.

Hussey SG, Mizrachi E, Spokevicius A V, Bossinger G, Berger DK, Myburg AA. **2011**. SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in *Arabidopsis* fibres and increases fibre cell area in Eucalyptus. *BMC Plant Biology* **11**: 173–190.

Ishii T. **1997**. Structure and functions of feruloylated polysaccharides. *Plant Science* **127**: 111–127.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Jolie RP, Duvetter T, Loey AM Van, Hendrickx ME. **2010**. Pectin methylesterase and its proteinaceous inhibitor: a review. *Carbohydrate Research* **345**: 2583–2595.

Jung J, Kim S, Seo P, Park C. **2008**. Molecular mechanisms underlying vascular development. *Advances in Botanical Research* **48**: 1–68.

Kawabata S, Miura K, Nitta K, Kawano K. **2000**. Chitin-binding proteins in invertebrates and plants comprise a common chitin- binding structural motif. *The Journal of Biological Chemistry* **275**: 17929–17932.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. **2012**. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Evolution and Biology* **4**: 316–329.

Lairson LL, Henrissat B, Davies GJ, Withers SG. **2008**. Glycosyltransferases: structures, functions, and mechanisms. *Annual Review of Biochemistry* **77**: 521–555.

Lee C, Teng Q, Huang W, Zhong R, Ye Z. **2009**. The Poplar GT8E and GT8F glycosyltransferases are functional orthologs of *Arabidopsis* PARVUS involved in glucuronoxylan biosynthesis. *Plant and Cell Physiology* **50**: 1982–1987.

Lee C, Zhong R, Ye Z-H. **2012a**. *Arabidopsis* family GT43 members are xylan xylosyltransferases required for the elongation of the xylan backbone. *Plant & Cell Physiology* **53**: 135–143.

Lee C, Zhong R, Ye Z-H. **2012b**. Biochemical characterization of xylan xylosyltransferases involved in wood formation in poplar. *Plant Signaling and Behaviour* **7**: 332–337.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O. **2012**. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31**: 1–46.

Levy I, Shoseyov O. **2002**. Cellulose-binding domains: Biotechnological applications. *Biotechnology Advances* **20**: 191 – 213.

Li Q, Min D, Wang JP-Y, Peszlen I, Horvath L, Horvath B, Nishimura Y, Jameel H, Chang H-M, Chiang VL. **2011**. Down-regulation of glycosyltransferase 8D genes in *Populus trichocarpa* caused reduced mechanical strength and xylan content in wood. *Tree Physiology* **31**: 226–236.

Li X-L, Skory CD, Cotta M a, Puchart V, Biely P. **2008**. Novel family of carbohydrate esterases, based on identification of the *Hypocrea jecorina* acetyl esterase gene. *Applied and Environmental Microbiology* **74**: 7482–7489.

Linhardt RJ, Galliher PM, Cooney CL. **1986**. Polysaccharide lyases. *Applied Biochemistry and Biotechnology* **12**: 135–176.

Lutz M. **2008**. *Learning Python* (T Apandi, Ed.). O'Reilly Meidia, Inc.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. **2005**. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences* **102**: 5454–5459.

Maloney VJ, Mansfield SD. **2010**. Characterization and varied expression of a membrane-bound endo-1,4-β-glucanase in hybrid poplar. *Plant Biotechnology Journal* **8**: 294–307.

Maloney VJ, Samuels AL, Mansfield SD. **2011**. The endo-1,4-β-glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytologist* **10**: 1–12.

Martinez-Fleites C, He Y, Davies GJ. **2010**. Structural analyses of enzymes involved in the *O*-GlcNAc modification. *Biochimica et Biophysica Acta* **1800**: 122–133.

Mellerowicz EJ, Baucher M, Sundberg B, Boerjan W. **2001**. Unravelling cell wall formation in the woody dicot stem. *Plant Molecular Biology* **47**: 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.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, *et al.* 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**: 245–250.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly B V, Lewis KLT, *et al.* 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). *Nature* **452**: 991–996.

Minic Z. **2008**. Physiological roles of plant glycoside hydrolases. *Planta* **227**: 723–740.

Nasir A, Naeem A, Jawad Khan M, Lopez-NicoraArshan HD, Caetano-Anollés G. **2011**. Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across Superkingdoms. *Gene* **2**: 869–911.

Palin R, Geitmann A. **2012**. The role of pectin in plant morphogenesis. *Bio-Systems* **109**: 397–402.

Parsons HT, Christiansen K, Knierim B, Carroll A, Ito J, Batth TS, Smith-Moritz AM, Morrison S, McInerney P, Hadi MZ, *et al.* 2012. Isolation and proteomic characterization of the *Arabidopsis* Golgi defines functional and novel components involved in plant cell wall biosynthesis. *Plant Physiology* **159**: 12–26.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.

Pawar PM-A, Koutaniemi S, Tenkanen M, Mellerowicz EJ. **2013**. Acetylation of woody lignocellulose: significance and regulation. *Frontiers in Plant Science* **4**: 118–126.

Plomion C, Stokes A, Leprovost G. **2001**. Wood formation in trees. *Plant Physiology* **127**: 1513–1523.

Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB, Herve' C. **2011**. Evolution and diversity of plant cell walls : From algae to flowering plants. *Annual Review of Plant Biology* **62**: 567–590.

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-laylin LK, *et al.* 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**: 223–226.

Proost S, Pattyn P, Gerats T, Van de Peer Y. **2011**. Journey through the past: 150 million years of plant genome evolution. *The Plant Journal* **66**: 58–65.

Rennie EA, Hansen SF, Baidoo EEK, Hadi MZ, Keasling JD, Scheller HV. **2012**. Three members of the *Arabidopsis* glycosyltransferase family Are xylan glucuronosyltransferases. *Plant Physiology* **159**: 1408–1417.

Rensing S a, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist E a, Kamisugi Y, *et al.* 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**: 64–69.

Sánchez-Rodríguez C, Bauer S, Hématy K, Saxe F, Ibáñez AB, Vodermaier V, Konlechner C, Sampathkumar A, Rüggeberg M, Aichinger E, *et al.* 2012. Chitinase-like1/pom-pom1 and its homolog CTL2 are glucan-interacting proteins important for cellulose biosynthesis in *Arabidopsis*. *The Plant Cell* **24**: 589–607.

Schmutz J, Cannon SB, Schlueter J, MA J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Serapiglia MJ, Cameron KD, Stipanovic AJ, Smart LB. **2011**. Correlations of expression of cell wall biosynthesis genes with variation in biomass composition in shrub willow (*Salix* spp.) biomass crops. *Tree Genetics & Genomes* **10**: 1–9.

Showalter AM. **1993**. Structure and function of plant cell wall proteins. *The Plant Cell* **5**: 9–23.

De Smet R, Van de Peer Y. **2012**. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology* **15**: 168–176.

The Arabidopsis Genome Initiative. **2000**. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

The International Brachypodium Initiative. **2010**. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763-768.

Trapnell C, Pachter L, Salzberg SL. **2009**. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105-1111.

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

Tognetti VB, Van Aken O, Morreel K, Vandenbroucke K, van de Cotte B, De Clercq I, Chiwocha S, Fenske R, Prinsen E, Boerjan W, *et al.* 2010. Perturbation of indole-3-butyric

acid homeostasis by the UDP-glucosyltransferase UGT74E2 modulates *Arabidopsis* architecture and water stress tolerance. *The Plant Cell* **22**: 2660–2679.

Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. **2005**. Modules, multidomain proteins and organismic complexity. *The FEBS Journal* **272**: 5064–5078.

Tsai AY-L, 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.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–15604.

Tyler L, Bragg JN, Wu J, Yang X, Tuskan GA, Vogel JP. **2010**. Annotation and comparative analysis of the glycoside hydrolase genes in *Brachypodium distachyon*. *BMC Genomics* **11**: 1–21.

Vanholme R, Morreel K, Ralph J, Boerjan W. **2008**. Lignin engineering. *Current Opinion in Plant Biology* **11**: 278–85.

Vega-Sánchez ME, Ronald PC. **2010**. Genetic and biotechnological approaches for biofuel crop improvement. *Current Opinion in Biotechnology* **21**: 218–224.

Veitia RA, Bottani S, Birchler JA. **2008**. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics* **24**: 390–397.

Wilson IBH. **2002**. Glycosylation of proteins in plants and invertebrates. *Current Opinion in Structural Biology* **12**: 569–577.

Wu A-M, Hörnblad E, Voxeur A, Gerber L, Rihouey C, Lerouge P, Marchant A. **2010**. Analysis of the *Arabidopsis* IRX9/IRX9-L and IRX14/IRX14-L pairs of glycosyltransferase

genes reveals critical contributions to biosynthesis of the hemicellulose glucuronoxylan. *Plant Physiology* **153**: 542–554.

Yin Y, Chen H, Hahn MG, Mohnen D, Xu Y. **2010**. Evolution and function of the plant cell wall synthesis-related glycosyltransferase family 8. *Plant Physiology* **153**: 1729–1746.

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* **479**: 1–7.

Yokoyama R, Nishitani K. **2004**. Genomic basis for cell-wall diversity in plants. A comparative approach to gene families in rice and *Arabidopsis*. *Plant & Cell Physiology* **45**: 1111–1121.

Yokoyama R, Uwagaki Y, Sasaki H, Harada T, Hiwatashi Y, Hasebe M, Nishitani K. **2010**. Biological implications of the occurrence of 32 members of the XTH (xyloglucan endotransglucosylase/hydrolase) family of proteins in the bryophyte *Physcomitrella patens*. *The Plant Journal* **64**: 645–656.

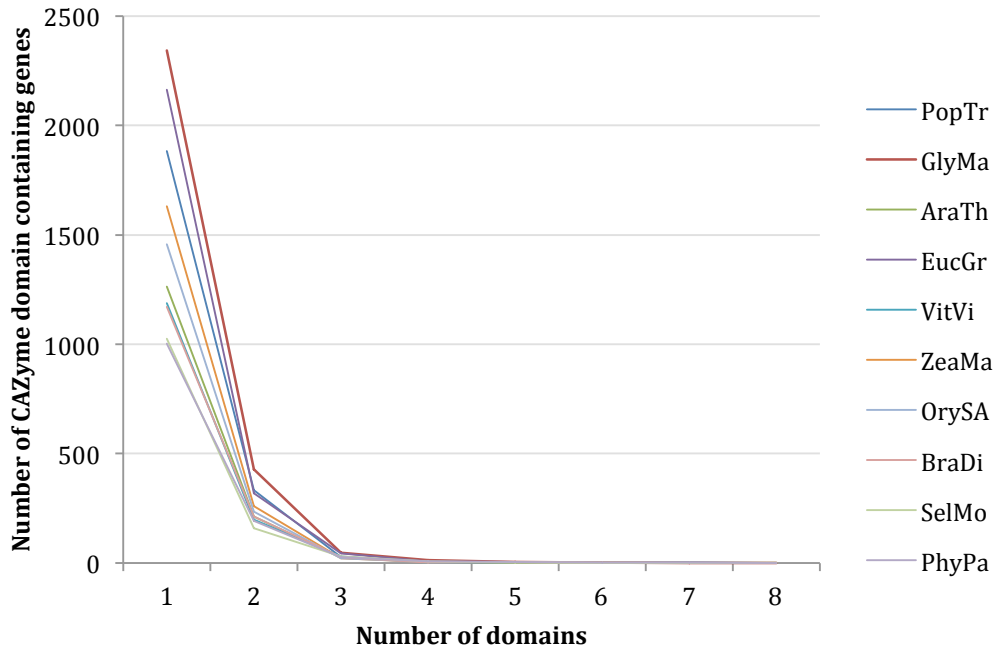
Zabotina OA. **2012**. Xyloglucan and its biosynthesis. *Frontiers in Plant Science* **3**: 134–139.

2.7. Supplementary Tables and Figures

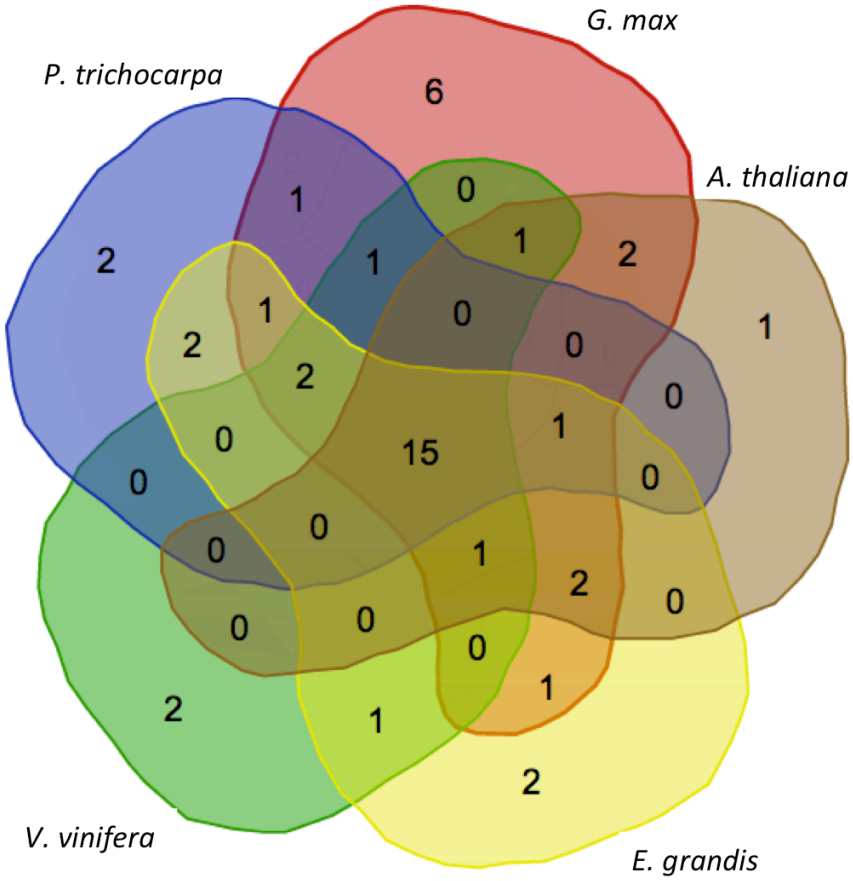
Supplementary Table 2.1 Relative standard deviation (RSD) (absolute co-efficient of variation) between plant species.

	GH RSD%	GT RSD%	PL RSD%	CE RSD%	CBM RSD%
Green algae	4.88	5.28	9.80	3.25	15.17
Lycophytes and bryophytes	05	0.90	27.20	10.28	64
Monocots	8.70	6.78	37.68	6.57	13.82
Eudicots	5.44	3.75	11.85	4.91	8.40

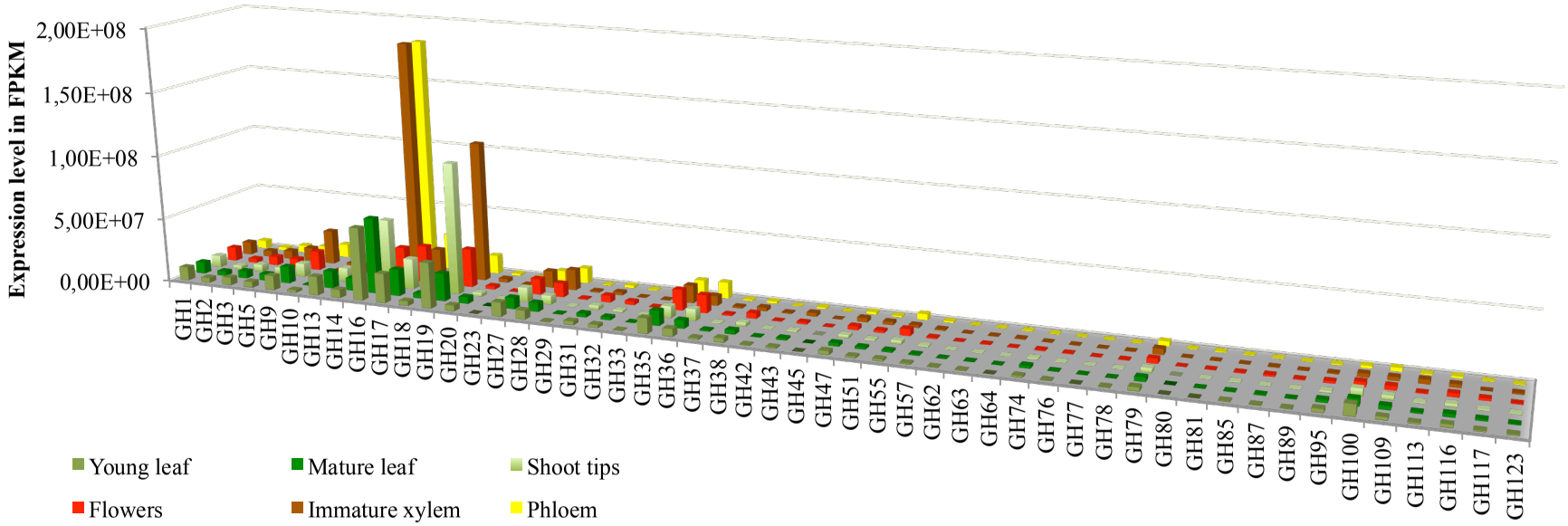
Co-efficient of variation analysis for 5 classes of CAZyme domains within phylogenetic groups, anything below 20% is considered significant (No variation). The null hypothesis for this analysis was that the CAZyme domain class ratios within the groups of plants do not vary significantly**.**



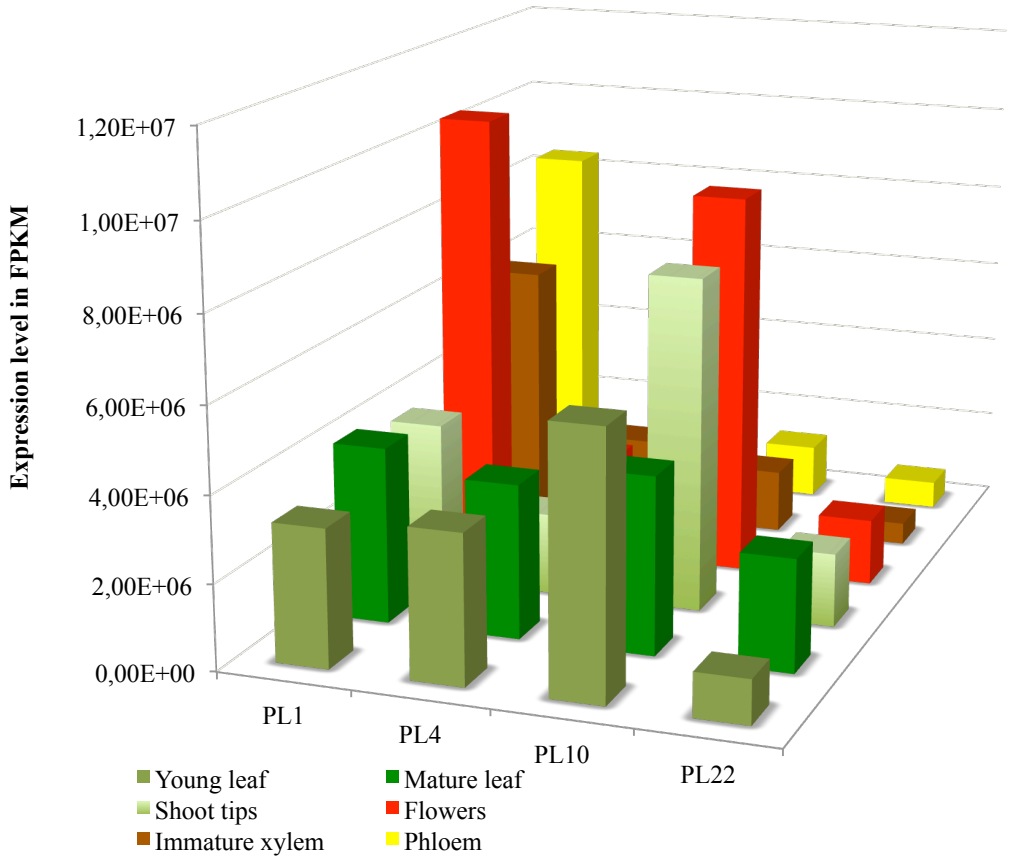
Supplementary Figure 2.1 Number of CAZy domains in complex CAZy domain containing proteins across ten representative plant species. Protein domain complexity in CAZyme domain containing proteins decreases with the number of domains within complex proteins across ten plant species.



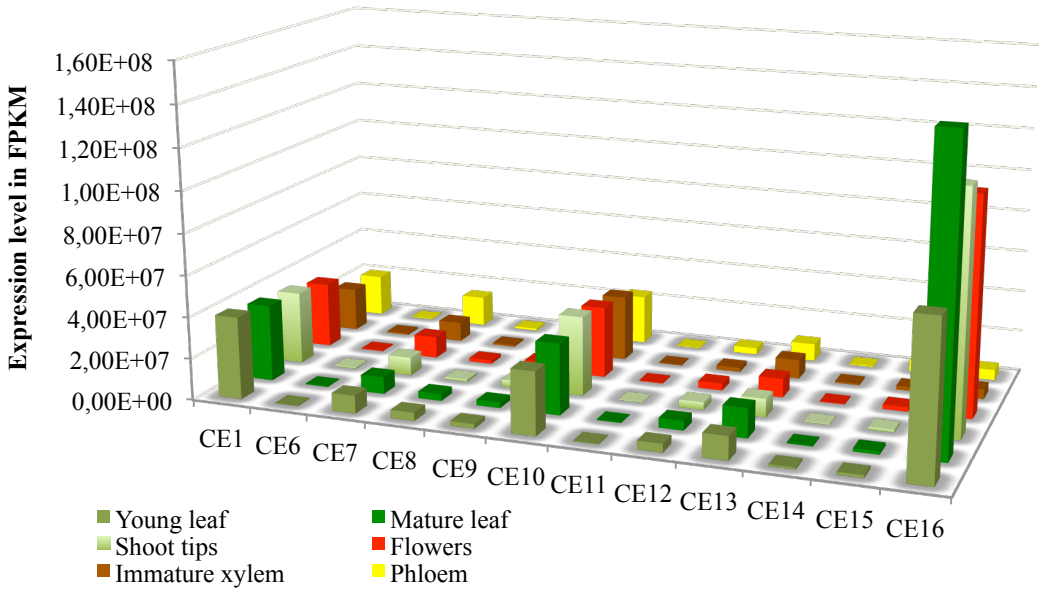
Supplementary Figure 2.2 Venn diagram of CAZyme domain unique combinations within complex proteins in five eudicots. The majority of CAZyme domain unique combinations in complex proteins are shared among the five eudicot species analyzed. The raw data used to generate this Venn diagram can be found in Additional file 2.4.



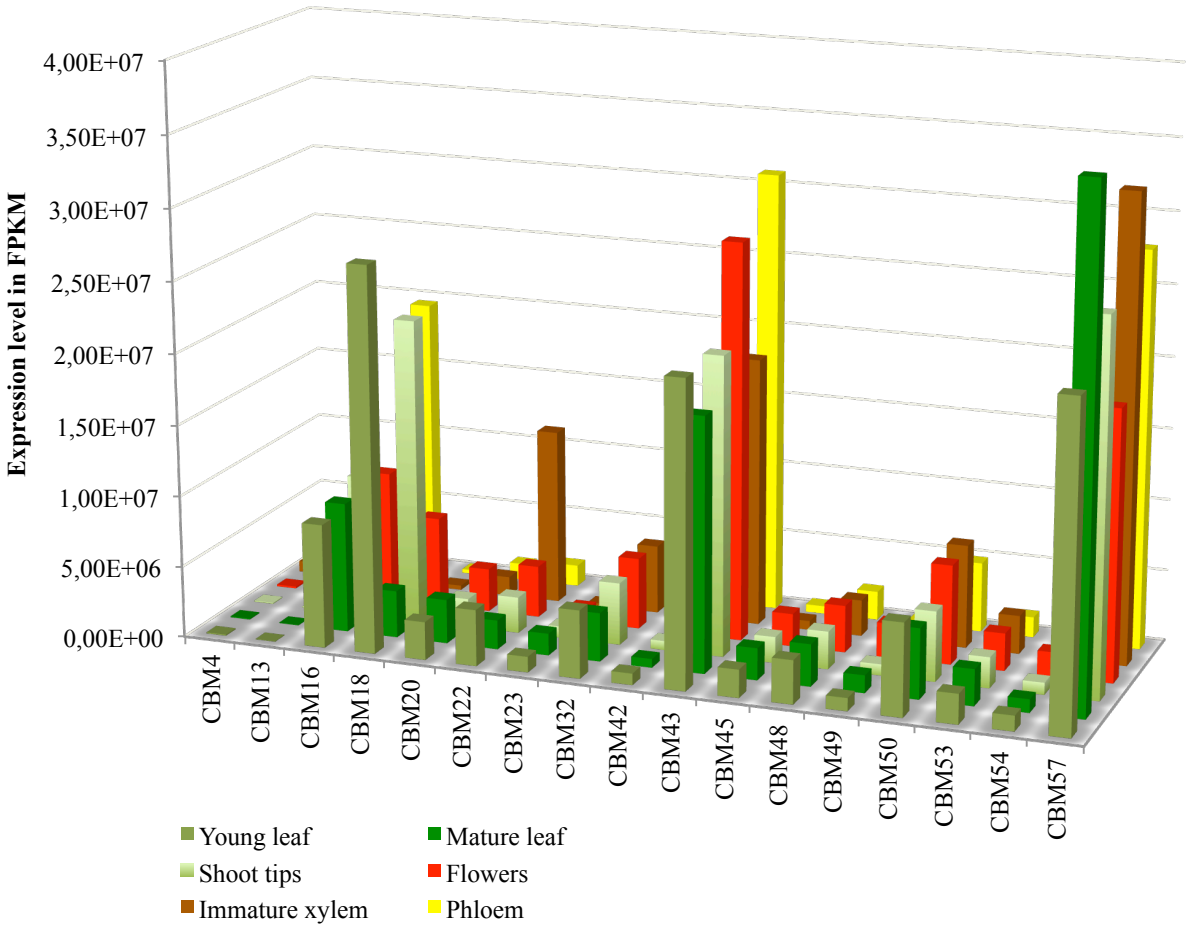
Supplementary figure 2.3 GH domain family expression levels across six tissues in *E. grandis* in FPKM. The y-axis shows the total expression investment in FPKM from raw mRNA-seq data, while the x-axis shows the glycosyl hydrolase (GH) domain family whose gene expression was summed for the analysis. The depth axis is the tissue type in *E. grandis* for which each domain family investment FPKM was calculated. Light green- young leaf, dark green- mature leaf, mint green- shoot tips, red- flowers, brown-immature xylem and yellow- phloem. The raw FPKM data used to generate this figure can be found in Additional file 2.2



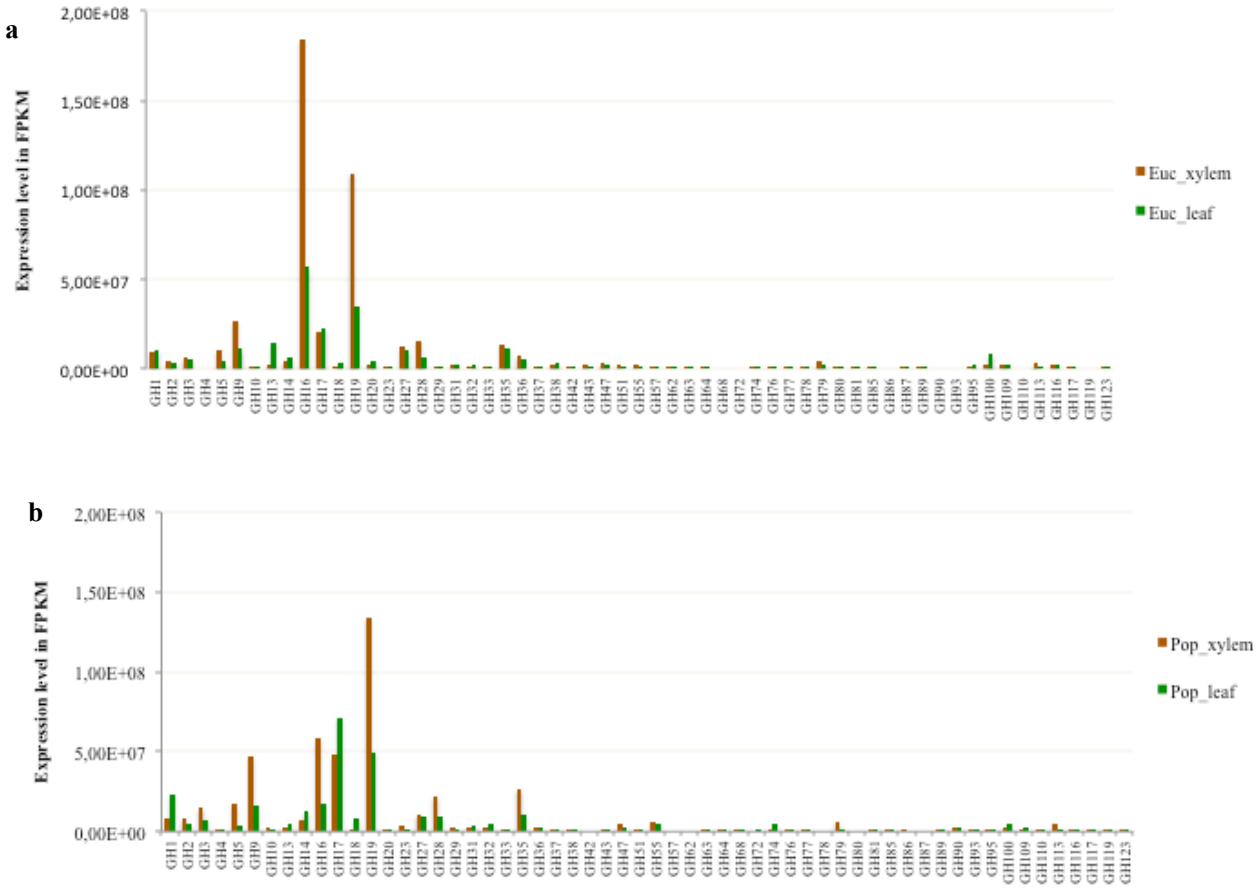
Supplementary Figure 2.4 PL domain family expression levels across six tissues in *E. grandis* in FPKM. The y-axis shows the total expression investment in FPKM from raw mRNA-seq data, while the x-axis shows the polysaccharide lyase (PL) domain family whose gene expression was summed for the analysis. The depth axis is the tissue type in *E. grandis* for which each domain family investment FPKM was calculated. Tissues are colour coded as in Supplementary Figure 2.3.



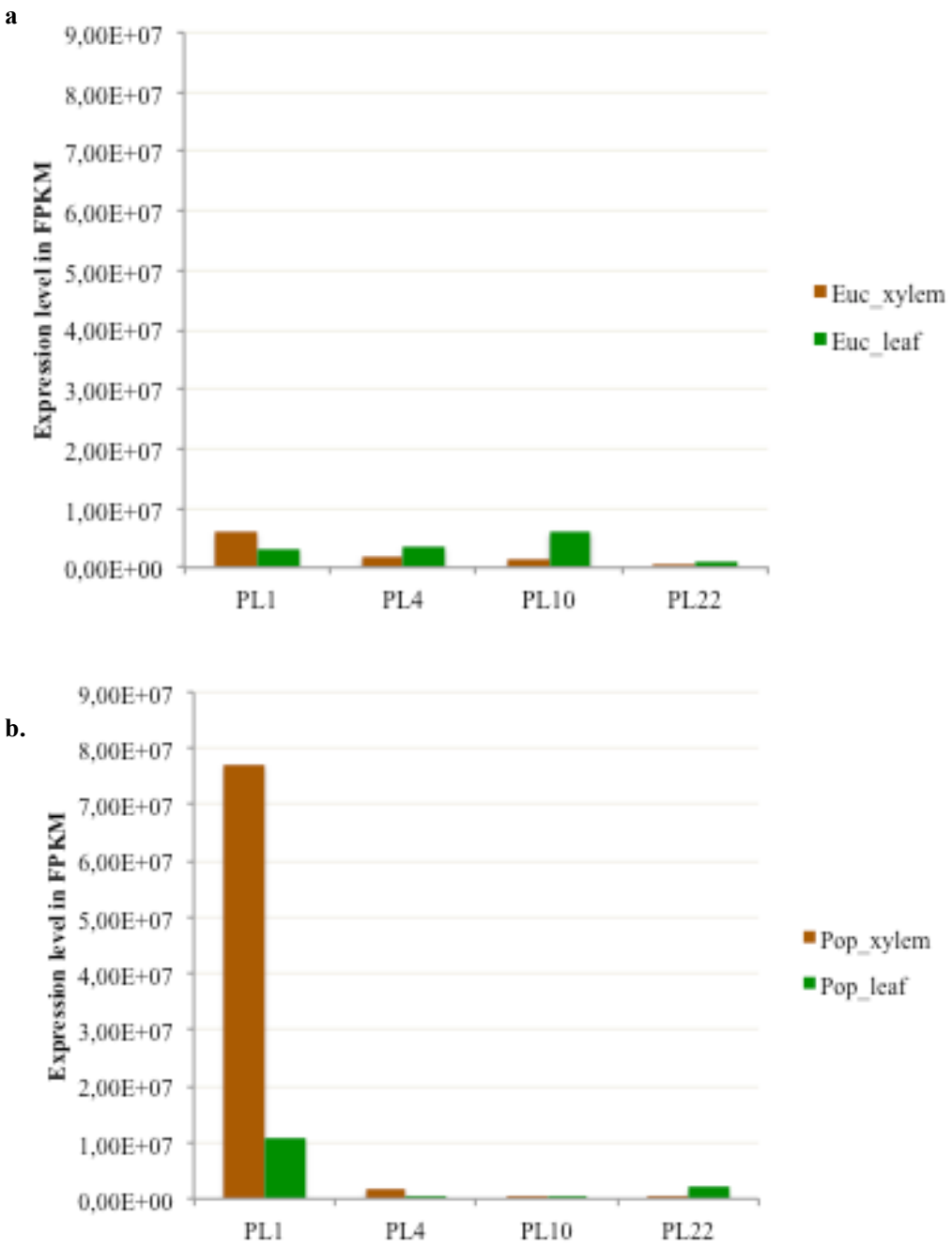
Supplementary Figure 2.5 CE domain family expression level across six tissues in *E. grandis* in FPKM. The y-axis shows the total expression investment in FPKM from raw mRNA-seq data, while the x-axis shows the carbohydrate esterase (CE) domain family whose gene expression was summed for the analysis. The depth axis is the tissue type in *E. grandis* for which each domain family investment FPKM was calculated.



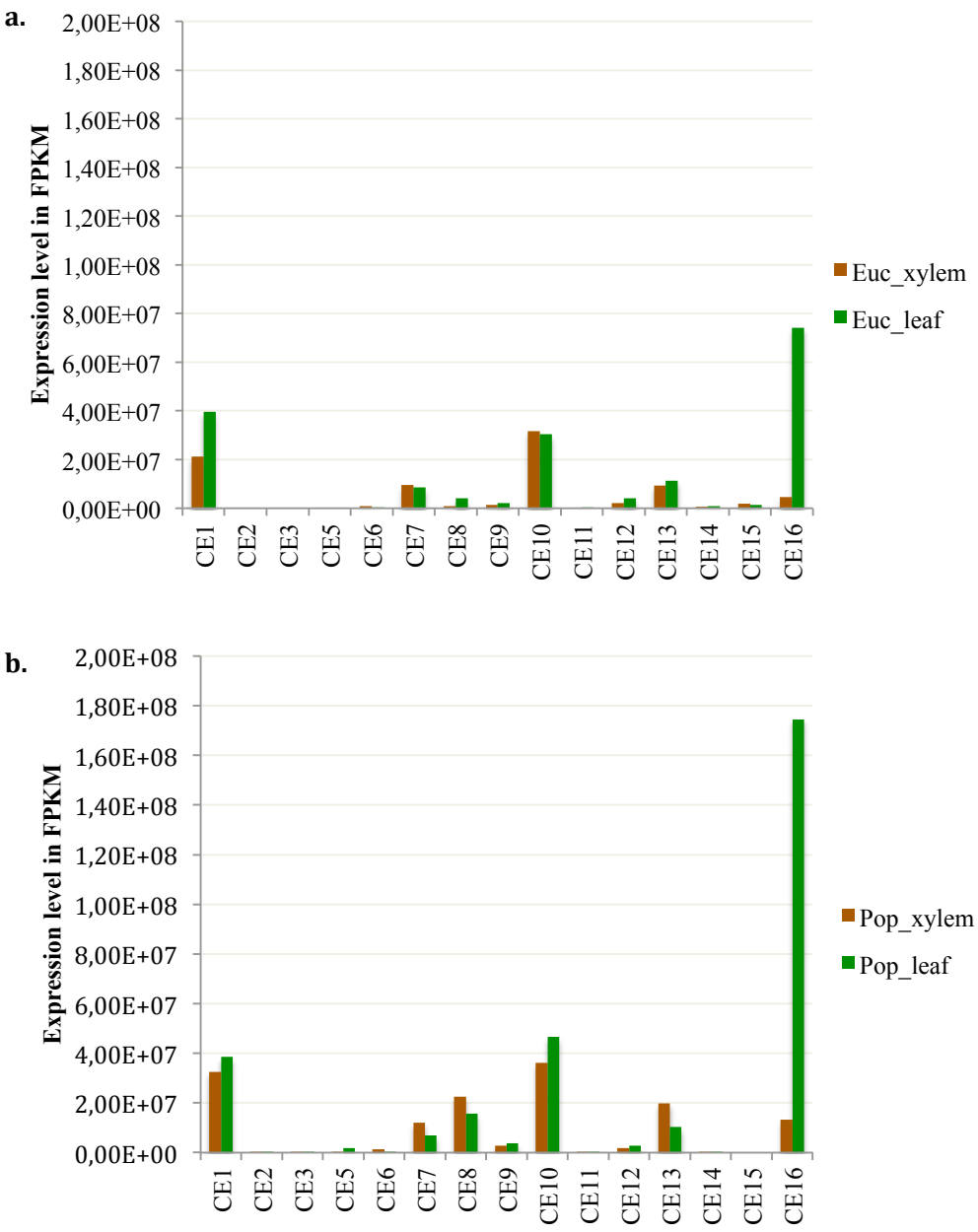
Supplementary Figure 2.6 CBM domain family expression level across six tissues in *E. grandis* in FPKM. The y-axis shows the total expression level in FPKM from raw mRNA-seq data, while the x-axis shows the carbohydrate binding module (CBM) domain family whose gene expression was summed for the analysis. The depth axis is the tissue type in E. grandis for which each domain family investment FPKM was calculated.



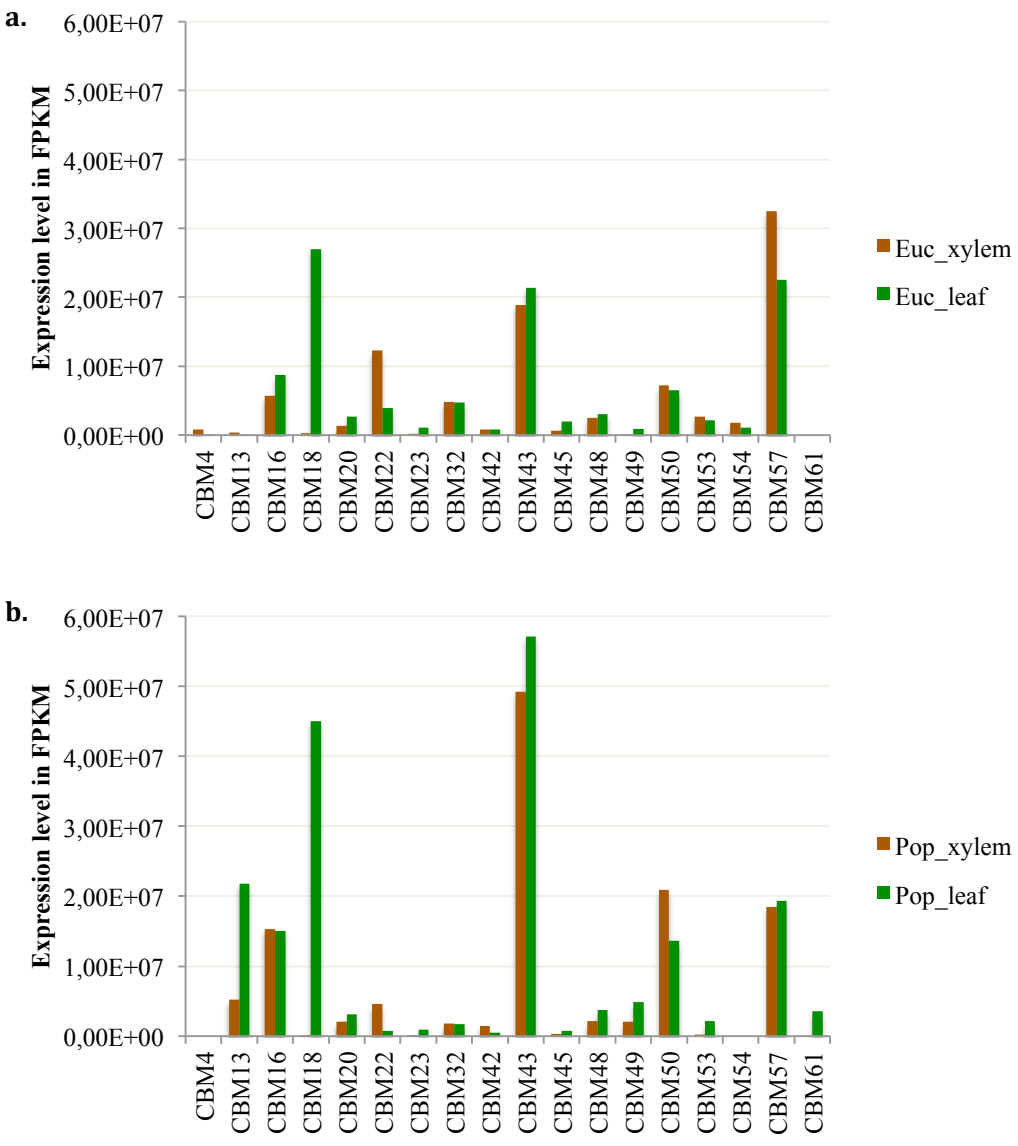
Supplementary Figure 2.7 Comparative expression patterns of GH domain families in E. grandis and P. trichocarpa. The y-axis shows the expression levels in FPKM for each glycosyl hydrolase (GH) domain family on the x-axis in (a) *E. grandis* xylem and leaf tissue and (b) *P. trichocarpa* xylem and leaf tissue. The bars representing the xylem tissue are shown in brown, and the bars representing the leaf tissue are shown in green.



Supplementary Figure 2.8 Comparative expression patterns of PL domain families in *E. grandis* and *P. trichocarpa*. The y-axis shows the expression investment values in FPKM for each polysaccharide lyase (PL) domain family on the x-axis in (a) *E. grandis* xylem and leaf tissue and (b) *P. trichocarpa* xylem and leaf tissue. The bars representing the xylem tissue are shown in brown, and the bars representing the leaf tissue are shown in green.



Supplementary Figure 2.9 Comparative expression patterns of CE domain families in *E. grandis* and *P. trichocarpa*. The y-axis shows the expression level in FPKM for each carbohydrate esterase (CE) domain family on the x-axis in (a) *E. grandis* xylem and leaf tissue and (b) *P. trichocarpa* xylem and leaf tissue. The bars representing the xylem tissue are shown in brown, and the bars representing the leaf tissue are shown in green.



Supplementary Figure 2.10 Comparative expression patterns of CBM domain families in *E. grandis* and *P. trichocarpa*. The y-axis shows the expression investment values in FPKM for each carbohydrate binding module (CBM) domain family on the x-axis in (a) *E. grandis* xylem and leaf tissue and (b) *P. trichocarpa* xylem and leaf tissue. The bars representing the xylem tissue are shown in brown, and the bars representing the leaf tissue are shown in green.

2.8 Additional files

Please see attached disc on back cover

- Additional file 2.1.xls- CAZyme domain family frequency across twenty-two plant species.

- Additional file 2.2.xls- Expressed CAZyme domain containing proteins (FPKM) and domain content in *E. grandis*.

- Additional file 2.3.xls- CAZyme domain containing protein complexity summary in 10 plant species, showing the frequency of complex proteins with unique and repeated CAZyme domains in ten plant species**.**

- Additional file 2.4.xls- Frequency of unique CAZyme domain combinations in complex proteins in 10 plant species

- Additional file 2.5.txt- Python script domain_counter.py: Used to count the frequency of multiple domains in all species for all families across columns.

- Additional file 2.6.txt- Python script domain_pull.py: Used to sort gene frequency based on domain family.

Chapter 3

Concluding remarks

Protein domains are the functional and evolutionary building blocks of proteins, they evolve independently, and in combination and via interactions they constitute the "protein universe" (Chothia *et al.*, 2003; Levitt, 2009). The protein universe consists of a set number of functional domains that combine and interact in a multitude of ways to produce the wide variety of boilogical networks seen across life (Levitt, 2009). The majority of protein domains are conserved across all forms of life, with a relatively minor proportion being unique across kingdoms (Apic *et al.*, 2001). As independently evolving units, they can be used to analyze the functional potential of whole genomes (Parikesit *et al.*, 2012).

A domain-centric analysis can be extremely useful in studying the fundamental metabolic strategies of carbohydrate metabolism without the confounding element of paralogs and homologs that are found in abundance in plant genomes (Forslund *et al.*, 2011; Bradshaw *et al.*, 2011; Zhang *et al.*, 2012). Carbohydrate metabolism in plants is a very economically important process, as the carbohydrates that are deposited in plant cell walls as celluloses, hemicelluloses, and pectins are used in a variety of industrial activities. These range from the production of pulp and paper, high-end cellulose derivatives, food products, and other bioenergy or biomaterial applications. Carbohydrate metabolism strategies vary across plant lineages, with some plants placing an emphasis on secondary cell wall components. The carbon that is sequestered can be stored temporarily, e.g. starch, or permanently, e.g. in the cell wall as cellulose and hemicelluloses. Although the main types of polysaccharides that are deposited in cell walls in plants are common to plant lineages, a lot of heterogeneity in cell wall composition exists within, and between plant species. To that end, it is important to understand the proteins, and indeed, the domains that make up these proteins in plants with different carbohydrate metabolism strategies.

Carbohydrate Active enZyme (CAZyme) domains are the domains responsible for the synthesis, degradation and modification of glycosidic bonds in plants and all other forms of life (Coutinho *et al.*, 2003). In the case of polysaccharide biosynthesis, much work has been done in identifying and classifying these functional building blocks of the enzymatic proteins

responsible (Cantarel *et al.*, 2009; Yin *et al.*, 2012). Due to the complexity of polysaccharides that are found in plants, and the difficulty in determining the structures of complex cell wall polysaccharides such as pectin, much is still unknown about the functions of many of the CAZyme domains beyond broad functional classifications (Popper *et al.*, 2011; Atmodjo *et al.*, 2013). The functions of glycosyl hydrolases (GHs) are the best characterized out of the five different functional classes of CAZyme domains, as they are integral to the ability of fungi and bacteria to degrade plant cell walls and utilize the sugars for energy (Wei *et al.*, 2009). A domain-centric approach has been used to delineate carbohydrate metabolism strategies in a variety of fungi and bacteria (Ospina-Giraldo *et al.*, 2010; Battaglia *et al.*, 2011; Manzo *et al.*, 2011). The functions of other domain class families are less well characterized, the exceptions being the gene members of CAZyme domain families known to be involved in the synthesis of cell wall polysaccharides such as cellulose and hemicelluloses (Coutinho *et al.*, 2003).

This study has highlighted the role of conserved functional CAZyme domain classes in plant evolution. We find that the fundamental building blocks of cell wall formation in plant species is reflected in the conserved ratio of functional CAZyme classes. Recent studies in the evolution of plant biopolymers have shown that organisms such as green algae have the ability to synthesize the same fundamental polysaccharides that are found in woody plants (Leliaert *et al*., 2012). Del Bem and Vincentz reported in 2010 that the enzymatic machinery necessary for the synthesis of xyloglucan, the major hemicellulose in non-graminaceous angiosperms, predates the terrestrialization of plants (Del Bem & Vincentz, 2010). Similarly, we find that the expression profiles of CAZyme domain families are conserved in two divergent tree species. The results of this study highlight the importance of several key factors in the evolution of plant carbohydrate metabolism. Firstly, the fundamental functional building blocks for polysaccharide synthesis, degradation and modification are conserved across plant evolution; and second, the controlled, dynamic differential regulation of these CAZyme domain containing genes is the major contributor to the diversity of plant metabolic

strategies seen throughout their evolution. Furthermore, we highlight CAZyme domain families with unresolved functions in the process of carbohydrate metabolism, such as the GT41 domain family members involved in signaling, provide an attractive avenue for future research in wood formation. This study emphasizes the need to re-examine past comparative genomics studies, as due to next generation sequencing technology, more sequenced land plant genomes are available (Martin *et al*., 2013). Our results have expanded the understanding of the role of the fundamental tools plants utilize synthesize, degrade and modify polysaccharides, CAZyme domains, in wood formation and evolution from the previous findings in *Populus trichocarpa* and *Arabidopsis thaliana* (Geisler-Lee *et al*., 2006).

References

Apic G, Gough J, Teichmann SA. **2001**. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* **310**: 311–325.

Atmodjo MA, Hao Z, Mohnen D. **2013**. Evolving views of pectin biosynthesis. *Annual Review of Plant Biology* **64**: 747–79.

Battaglia E, Benoit I, Brink J Van Den, Wiebenga A, Coutinho PM, Henrissat B, Vries RP De, van den Brink J, de Vries RP. **2011**. Carbohydrate-active enzymes from the zygomycete fungus *Rhizopus oryzae*: a highly specialized approach to carbohydrate degradation depicted at genome level. *BMC Genomics* **12**: 38–50.

Del Bem LE V, Vincentz MG a. **2010**. Evolution of xyloglucan-related genes in green plants. *BMC Evolutionary Biology* **10**: 341-358.

Bradshaw CR, Surendranath V, Henschel R, Mueller MS, Habermann BH. **2011**. HMMerThread: detecting remote, functional conserved domains in entire genomes by combining relaxed sequence-database searches with fold recognition. *PLoS One* **6**: 1–17.

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**: 233–238.

Chothia C, Gough J, Vogel C, Teichmann S a. **2003**. Evolution of the protein repertoire. *Science* **300**: 1701–3.

Coutinho PM, Stam M, Blanc E, Henrissat B. **2003**. Why are there so many carbohydrate-active enzyme-related genes in plants? *Trends in Plant Science* **8**: 563–565.

Forslund K, Pekkari I, Sonnhammer ELL. **2011**. Domain architecture conservation in orthologs. *BMC Bioinformatics* **12**: 326-341.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O. **2012**. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31**: 1–46.

Levitt M. **2009**. Nature of the protein universe. *Proceedings of the National Academy of Sciences* **106**: 11079–84.

Manzo N, D'Apuzzo E, Coutinho PM, Cutting SM, Henrissat B, Ricca E. **2011**. Carbohydrate-active enzymes from pigmented *Bacilli*: a genomic approach to assess carbohydrate utilization and degradation. *BMC Microbiology* **11**: 198.

Martin LBB, Fei Z, Giovannoni JJ, Rose JKC. **2013**. Catalyzing plant science research with RNA-seq. *Frontiers in Plant Science* **4**: 1–10.

Ospina-Giraldo MD, Griffith JG, Laird EW, Mingora C. **2010**. The CAZyome of *Phytophthora spp*.: a comprehensive analysis of the gene complement coding for carbohydrate-active enzymes in species of the genus *Phytophthora. BMC Genomics* **11**: 525-541

Parikesit AA, Stadler PF, Prohaska SJ. **2012**. Large-scale evolutionary patterns of protein domain distributions in Eukaryotes *Open Series in Bioinformatics* **4**: 1-10

Popper Z a ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB, Herve' C. **2011**. Evolution and diversity of plant cell walls: From algae to flowering plants. *Annual Review of Plant Biology* **62**: 567–590.

Wei H, Xu Q, Taylor LE, Baker JO, Tucker MP, Ding S-Y. **2009**. Natural paradigms of plant cell wall degradation. *Current Opinion in Biotechnology* **20**: 330–338.

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* **479**: 1–7.

Zhang X-C, Wang Z, Zhang X, Le MH, Sun J, Xu D, Cheng J, Stacey G. **2012**. Evolutionary dynamics of protein domain architecture in plants. *BMC Evolutionary Biology* **12**: 6.