Functional characterization of cell wall related proteins of unknown function (CW-PUFs) in *Arabidopsis thaliana*

by

Ritesh Mewalal

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

October 2015

Supervisor: Prof. Alexander A. Myburg

Co-supervisors: Prof. Shawn D. Mansfield and Dr Eshchar Mizrachi

Declaration

I, Ritesh Mewalal declare that the thesis, which I hereby submit for the degree Philosophiae Doctor (Ph.D.) at the University of Pretoria, is my own work and has not been submitted by me for a degree at this or any other tertiary institution.

Ritesh Mewalal

October 25, 2015

Date

THESIS SUMMARY

Functional characterization of cell wall-related proteins of unknown function (CW-PUFs) in *Arabidopsis thaliana*

Ritesh Mewalal

Supervised by Prof. A.A. Myburg (University of Pretoria)

Co-supervised by Prof. S.D. Mansfield (University of British Columbia, Canada) and Dr. E. Mizrachi (University of Pretoria)

Department of Genetics, University of Pretoria, Pretoria

Submitted in partial fulfillment of the requirements for the degree Philosophiae Doctor

Lignocellulosic biomass is an important feedstock for the bioeconomy, particularly for biorefinery and biomaterial application. This is due to the characteristic secondary cell walls composed of a matrix of cellulose and hemicellulose intricately linked to and rigidified by lignin. Studies have estimated 10-15% of ~27,000 protein-coding genes in the model herbaceous plant *Arabidopsis thaliana* are dedicated to cell wall development. However, conclusive experimental evidence validating cell wall functionality is only available for approximately 120 genes. This observation highlights a gap in our understanding which, poses a hindrance on biotechnology aimed at optimizing yield and reducing cell wall recalcitrance for efficient release of biopolymers. High-throughput "omics" technologies have generated inventories of cell wall-related genes, many of which are annotated as unknown (**c**ell **w**all-related **p**roteins of **u**nknown **f**unction, CW-PUFs) due to absence of supporting evidence for biological and/or molecular function. The importance of CW-PUFs can be estimated by evolutionary conservation, co-expression with known genes (e.g. cellulose synthase genes) and experimental characterization. CW-PUFs may be classified into two

groups, proteins containing conserved **d**omains of **u**nknown **f**unction (DUFs) and **p**roteins of **o**bscure **f**eatures (POFs) lacking any recognized domains/motifs.

The aim of this study was to identify candidate CW-PUFs in the fast-growing and economically important hardwood crop genus, *Eucalyptus,* and functionally characterize these genes in the model plant *Arabidopsis*. Using bioinformatics tools and approaches, xylem-expressed members of the DUF1218 gene family and a single POF gene were prioritized for functional characterization.

To gain insight into its structure and suggested role in cell wall biology, a comparative genomics analysis of the DUF1218 gene family was performed. Approximately 284 non-redundant DUF1218-encoding genes were identified across 22 plant genomes revealing a land plant origin for the protein domain family. Furthermore, several DUF1218 family members putatively involved in cell wall biology were identified. We characterized the DUF1218 gene At4g27435 and showed that it is specifically expressed in interfascicular and xylem fibers, and loss-of-function results in a perturbation of lignin content in the cell walls of *Arabidopsis*. A second candidate, At1g31720, was also preferentially expressed in secondary cell wall depositing tissues. Due to potential functional redundancy, we included the most closely related family member At4g19370 and showed that both proteins are targeted to the cell periphery. No changes were seen in cell wall lignin content and growth in the loss-of-function and overexpression lines. However, we observed decreased rosette size, rosette fresh weight and stem length for the double homozygous mutant. Cell wall chemistry analysis revealed a decrease in the total lignin content and an increase in the syringyl/guaiacyl (S:G) monolignol ratio in the double mutant relative to the control plants. *At1g31720* and *At4g19370* were subsequently named *MODIFYING WALL LIGNIN-1* (*MWL-1*) and *MWL-2* respectively.

Next, the role of an *Arabidopsis* POF gene (*At*POF1) in secondary cell wall biology was investigated. POF1 is Angiosperm-specific and a singleton in the sequenced plants. In agreement with characteristics of POFs, predicted DNA-, RNA- and protein-binding sites as well as regions of disorder were identified. *At*POF1 co-expressed genes were overrepresented in cell wall-related gene ontology terms including cell wall biogenesis. We showed that *AtPOF1* expression is associated with xylem fibres, while the protein is nuclear-targeted. No changes in growth parameters were observed in loss-of-function or overexpression

lines, however, secondary cell wall chemistry analysis revealed changes in the glucose content of the mutant line and total lignin content in the overexpression lines.

This Ph.D. study identified and ascribed function to four new candidate CW-PUFs. We found a novel function for At4g27435 where previous attempts were inconclusive. Furthermore, we found two additional members of the DUF1218 family that function redundantly as contributors to secondary cell wall biology, specifically a lignin-related role. We provide insight into the role of a POF in secondary cell wall biosynthesis. The studies presented in this dissertation provides further understanding into cell wall biology and reveals new candidate genes for engineering plants with customised cell wall properties.

PREFACE

A defining characteristic of plants is the presence of the cell wall, which functions as a determinant of morphology, structure, transport and defence. These functions are influenced by the biopolymer content and structure of the cell wall and comprise cellulose, pectin and hemicellulose in the primary cell wall and larger amounts of cellulose, hemicellulose and impregnations of lignin in the secondary cell wall. Increase in human population size and corresponding increase in the demand posed on natural resources is driving the field of agro-economic research. In particular, the secondary cell wall of lignocellulose biomass has great value for biorefinery and biomaterial-based utilities. The fast-growing and widely planted hardwood crop genus *Eucalyptus,* is an excellent lignocellulosic feedstock as it can typically contain 39-46% cellulose, 24-28% hemicellulose and 29-32% lignin and can produce an average of over 50 cubic meters of woody biomass per hectare per year in good growing conditions with up to 100 m^3/ha/yr recorded for the best genotypes in the best sites.

Despite the benefits of lignocellulosic biomass as an enriched source of carbohydrates and phenylpropanoids, a fundamental understanding of this complex biomaterial is imperative for experimental design towards optimized biopolymer yield and efficient depolymerisation to extract sugars and other building blocks. Biopolymer content, composition and spatial-temporal deposition is modulated by highly coordinated genetic pathways including biosynthetic, modifying and structural proteins which are in turn, regulated by a network of transcription factors, hormones and signaling proteins, ultimately resulting in the unique cell wall ultrastructure. The number of genes predicted to be involved in cell wall formation vastly exceed those with conclusive experimental characterization (~4000 vs. ~120 respectively) highlighting a gap in our understanding of this biologically and economically important process. Many implicated genes lack biological, cellular and/or molecular annotation and are broadly classified as **c**ell **w**all-related **p**roteins of **u**nknown **f**unction (CW-PUFs). The importance of CW-PUFs can be determined through analysis of evolutionary conservation, co-expression with known cell wall-related genes and reverse genetic approaches. Based on characteristics of the proteins, CW-PUFs can be divided into proteins containing Hidden Markov Model (HMM)-recognized **d**omains of **u**nknown **f**unction

(DUFs), or **p**roteins of **o**bscure **f**eatures (POFs) that do not contain any recognized domains or motifs. While the long-term goal is to contribute to the understanding of secondary cell wall biology, the aim of this Ph.D. study was to identify candidate CW-PUFs in a woody biomass crop of interest, *Eucalyptus,* and functionally characterize these genes in the model plant *Arabidopsis thaliana*. To achieve this aim, I mined *Eucalyptus* RNA-Seq data for genes with preferential expression in xylogenic tissue. A number of candidate genes were further prioritized using across-species comparative analyses and plant bioinformatics tools (discussed in Chapter 1). I subsequently identified three genes from the DUF1218 family (At4g27435, At1g31720 and At4g19370) and a singleton POF for further functional characterization in the context of cell wall biology and chemistry. The results of these studies were synthesized into the following chapters, compiled as manuscripts either published or in review:

In **Chapter 1**, I identify the different types of CW-PUFs from previous cell wall-related studies and review the status of the field with regard to these proteins. I catalog DUFs and POFs identified in fifteen important cell wall related genomic, transcriptomic and proteomic studies in *Arabidopsis,* and identify DUFs and POFs preferentially expressed in *Eucalyptus* and *Populus* xylem tissue. Furthermore, I propose approaches to prioritize candidates based on intrinsic protein features combined with genomic and genetic information to further guide experimental design.

Chapter 2 describes the DUF1218 family in representative model plants, *Arabidopsis*, *Populus* and *Oryza* and the plant of interest, *Eucalyptus*. A number of genomic and bioinformatics tools are used to characterize the domain and the DUF1218 containing proteins. Overall, the study offers a better understanding of the structure and possible functions of the DUF1218 gene family and guided future functional characterization studies of individual members. While attempts have been made in previous studies to understand the functions of the DUF1218-containing protein, At4g27435, its role was still unclear. Therefore, I expanded on previous studies by analysing the spatiotemporal expression and cell wall chemistry and growth parameters of loss-of-function and overexpression lines for the gene.

In **Chapter 3**, I characterize the role of a second DUF1218 gene, At1g31720. I also include the most closely related family member, At4g19370, due to potential functional redundancy. I identify

corresponding loss-of-function mutant lines, generate double homozygous mutant lines and independent overexpression lines to characterize the cell wall chemistry and growth parameters of these two genes. Furthermore, I investigate the subcellular localization of the proteins using transient expression in tobacco. Based on the results, At1g31720 and At4g19370 are named *MODIFYING WALL LIGNIN-1* (*MWL-1*) and *MWL-2* respectively.

In **Chapter 4**, I describe the role of an *Arabidopsis* POF gene that I name *At*POF1. I perform a comprehensive comparative bioinformatics analysis of the protein and generate promoter:β-glucuronidase reporter gene lines and stable N- and C-terminal GFP fusion lines to characterize the spatiotemporal expression and subcellular localization respectively. I further characterize the biological function of the protein in *Arabidopsis* by analysing cell wall chemistry and growth parameters in the identified loss-of-function mutant line and overexpression lines.

Finally, in **Chapter 5**, I provide a summary and concluding remarks on the current thesis. The main findings are discussed in context of the field at present, potential limitations and shortcomings of the research project are discussed and future prospects are proposed.

The studies presented in this Ph.D. dissertation is the result of work undertaken at the Department of Genetics at University of Pretoria, South Africa (2010-2015) and the Department of Wood Science at University of British Columbia, Canada (11/2011-03/2012) under the supervision of Prof. A.A. Myburg and co-supervision of Prof. S.D. Mansfield and Dr. E. Mizrachi. All chapters were prepared as independent manuscripts and have been published or submitted to the relevant journals.

Peer-reviewed publications and conference presentations from this Ph.D. study:

Ritesh Mewalal, Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg. (2015). The A*rabidopsis* domain of unknown function 1218 (DUF-1218) containing proteins, MODIFYING WALL LIGNIN-1 and 2 (At1g31720/MWL-1 and At4g19370/MWL-2 respectively), function redundantly to alter secondary cell wall lignin content. *In review: PLoS ONE*.

Ritesh Mewalal, Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg. (2015). Genome-wide and functional characterization of plant-specific Domain of Unknown Function 1218 (DUF1218) family. The International Union of Forest Research Organizations (IUFRO) Tree Biotechnology (Florence, Italy), June 2015. (Oral presentation).

Ritesh Mewalal, Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg. (2015). The nuclear-localized *Arabidopsis* protein of obscure features1 (*At*POF1, At1g47410) affects secondary cell wall glucose and lignin content. *Manuscript in preparation.*

***Ritesh Mewalal**, Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg. (2014). Cell Wall-related Proteins of Unknown Function: Missing Links in Plant Cell Wall Development. Plant and Cell Physiology* **55** (6):1031-1043.

Ritesh Mewalal, Eshchar Mizrachi, Victoria J. Maloney, Shawn D. Mansfield, Alexander A. Myburg. (2013). Genome-wide and functional characterization of a putative Cell Wall-related Proteins of Unknown Function (CW-PUFs). The International Union of Forest Research Organizations (IUFRO) Tree Biotechnology (North Carolina, United States), May 2013. (Poster presentation).

Ritesh Mewalal, Eshchar Mizrachi, Shawn D. Mansfield, Alexander A. (2012). Functional characterization of a putative secondary cell wall-related Protein of Unknown Function (PUFs). South African Genetics & Bioinformatics Society Conference, "The Data-mining Revolution" (Stellenbosch, South Africa), September 2012. (Oral presentation).

ACKNOWLEDGEMENTS

I have learnt that there is no science to doing science, just method to madness and through this crazy journey and perhaps my favourite part of the dissertation; I would like to acknowledge the following people:

My supervisor, **Zander Myburg** – Thank you for the opportunity to be part of the Forest Molecular Group and the freedom to grow as an independent researcher. I hope that one-day I have as much knowledge of the field and the success you have achieved in such a short space of time. Not forgetting, thank you for all the support and I hope that we can work as collaborators in future.

My co-supervisor, **Shawn Mansfield** – Thank you for your patience and valued insight in my project. The time spent and skills acquired at your laboratory have given me the confidence to thrive in unfamiliar surroundings, a necessary skill for all young researchers.

My co-supervisor, **Eshchar Mizrachi** – It was a pleasure to work alongside you and witness your thought process, not only because you are a talented researcher but a friend.

Henk Huismans – A special thank you to the ultimate mentor for **always** believing in me and starting my postgraduate career. You are an inspiration, that is all!

Sanushka Naidoo – An incredibly successful and the kindest woman I've ever met in science. Thank you for all the opportunities and support.

Mike Wingfield and **Bernard Slippers** – Thank you for showing me that the passion for science should also be enjoyed outside the laboratory.

Mom – your devotion and loyalty to your children continues to amaze me every day. I thank you for all the sacrifices you made throughout your life for me. **Dad** – your hard work and dedication to provide your children with everything they need to be where they are, is greatly appreciated. My sister, **Shivani** – You have walked through so many paths in life and have flourished. I am so proud of you. Thank you for being in my life.

My brother, **Sumanth Mewalal** - I am the luckiest person to have a brother like you and an unconditional friend. I really wish you all the luck in your future as you always do for me.

My best friend, **Ouma** (**Minique de Castro**). I am grateful for all walks of life, good and bad because when you push the boundaries, you realize friendship can be limitless! I cannot express how important you are to me but I know I am grateful I get to call you my best friend! I know that our passion for red wine, travel and science will always pave the road for epic adventures. Sleep when you are dead lady!

Des(re) Pinard, you have become one of the most important people in my life. You are a kind and loyal friend and best of all…I experience all of that, much appreciated lady!

Cool FMG: **Colan**(sworth), **Karen** (KK), (An)**Drew**, **Martin**, **Gabi** (not really FMG), Mr **Jono** and **Caryn Oates -** All you guys have gone out of your way for me, celebrating the good times and supporting me when I was down. I have enjoyed the wine, bands (even though I have no idea who they were) and shop talk. You guys are so smart and I have no doubt that we can expect **big things** in the future!

My amazing extended family, **Charl** and **Tes Hechter** - what a journey! I will never forget passing out on your couch, the eviction, holidays, clubbing…a lifetime. I watch in awe at how successful you are in life and how effortless you make everything seem. I look forward to sharing many future adventures.

Shani, **Joanne**, **Vinet**, **Bronwyn** and **Nicky** - I think everyone passes through our lives for a reason. We had great times and I am sure we will cross paths again but if we do not, Good Luck!

Allan Hall, **Berdine Coetzee**, **Jane Bredenkamp**, **Lizahn Zwart**, **Masipa Mokgadi** (**Sharon**), **Victoria J. Maloney** and **Elna Cowley** for everyday and technical help.

My mentorship students – **Mpho Sekgejane, Donovin W. Coles, Antonie Kloppers, Yorateme Tii-Klizu Jr, Martin Wierzbicki, Bianca Jansen van Rensburg, Sarayna Naidoo, Faith Daniels**. My sincerest thank you! Your hard work and thirst for knowledge reminded me constantly why I wanted to be a researcher.

The **Department of Genetics** (University of Pretoria, South Africa) and the **Forestry and Agricultural Biotechnology Institute** (FABI) (University of Pretoria, South Africa), **University of British Columbia**

(Canada), **Sappi Technology Centre** and **Council for Scientific and Industrial Research** (CSIR) is acknowledged for laboratory space, equipment and facilities to conduct scientific research.

The **University of Pretoria**, **National Research Foundation (NRF)**, **Sappi**, **The Technology and Human Resources for Industry Programme (THRIP)**, **Department of Science and Technology** (**DST)** and **African Initiative** are acknowledged for project funding and Ph.D. scholarships through the **Forest Molecular Genetics (FMG) Programme**.

The **International Union of Forest Research Organizations** (**IUFRO**) (Florence, Italy, June 2015; Asheville, North Carolina, USA, May 2013; Arraial d"Ajuda, Bahia, Brazil, June 2011) and the **South African Genetics & Bioinformatics Society Conference** (**SAGS**) (Stellenbosch, September 2012) are acknowledged for the opportunity to present work from the current dissertation.

"**There is no passion to be found playing small - in settling for a life that is less than the one you are capable of living**"

Nelson Mandela (1918-2013)

Table of Contents

CHAPTER 1

LITERATURE REVIEW

Cell Wall-related Proteins of Unknown Function: Missing Links in Plant Cell Wall Development

Ritesh Mewalal[1], Eshchar Mizrachi[1], Shawn D. Mansfield[2] and Alexander A. Myburg[1,*]

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

[2]Department of Wood Science, Faculty of Forestry, University of British Columbia, Forest Sciences Centre, 4030-2424 Main Mall, Vancouver, BC, V6T 1Z4, Canada

The following review was published in *Plant and Cell Physiology* (*PCP* 55 (6):1031-1043, Mewalal, *et al*. 2014..doi: 10.1093/pcp/pcu050). I drafted the manuscript and figures. I was also involved in the initial conception of the review along with Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg. The drafts of the manuscript were peer reviewed/edited by Eshchar Mizrachi, Shawn D. Mansfield and Alexander A. Myburg.

1.1. Abstract

Lignocellulosic biomass is an important feedstock for the pulp and paper industry as well as emerging biofuels and biomaterials industries. However, recalcitrance of the secondary cell wall to chemical or enzymatic degradation remains a major hurdle for efficient extraction of economically important biopolymers such as cellulose. It has been estimated that approximately 10-15% of ~27,000 protein-coding genes in the *Arabidopsis* genome are dedicated to cell wall development, however only ~130 *Arabidopsis* genes thus far have experimental evidence validating cell wall function. While many genes have been implicated through co-expression analysis with known genes, a large number are broadly classified as proteins of unknown function (PUFs). Recently the functionality of some of these unknown proteins in cell wall development has been revealed using reverse genetic approaches. Given the large number of cell wall-related PUFs, how do we approach and subsequently prioritize the investigation of such unknown genes that may be essential to or influence plant cell wall development and structure? Here, we address the aforementioned question in two parts; we first identify the different kinds of PUFs based on known and predicted features such as protein domains. Knowledge of inherent features of PUFs may allow for functional inference and a concomitant link to biological context. Second, we discuss omics-based technologies and approaches that are helping identify and prioritize cell wall-related PUFs by functional association. In this way, hypothesis-driven experiments can be designed for functional elucidation of many proteins that remain missing links in our understanding of plant cell wall biosynthesis.

Abbreviations: CW, cell wall; CWP, cell wall proteins; DUFs, domains of unknown function; GT, glycosyltransferase; GXM, glucuronoxylan methyltransferase; IRX, irregular xylem; ORFans, orphan open reading frames; POFs, proteins of obscure features; PUFs, proteins of unknown function; QTL, quantitative trait loci; eQTL, expression quantitative trait loci.

Key words: domain of unknown function; plant cell wall; proteins of obscure features; proteins of unknown function.

1.2. Introduction: moving toward the unknown

Worldwide, the growing demand for energy is driving increased fossil fuel consumption and contributing to rising levels of atmospheric carbon dioxide and associated effects on global climate (Asif and Muneer 2007; Quadrelli and Peterson 2007; Solomon et al. 2009). Combined, the situation has stimulated the search for alternative sustainable sources of energy and carbon-based materials. Plant cell wall-derived lignocellulosic biomass is a promising feedstock for cellulosic ethanol, a second generation biofuel (Rubin 2008; Wyman 2007), that does not compete with food production in contrast to starch-based biofuels (Cassman and Liska 2007; Karp and Shield 2008). In secondary tissues, the plant cell wall comprises primary and secondary wall layers. A thin, elastic primary cell wall is synthesized during initial cellular expansion, typically containing cellulose, xyloglucan and pectin (Cosgrove 1997). A thick secondary cell wall is deposited on the inside of the primary cell wall after cell expansion, and has higher amounts of cellulose, heteroxylan and is impregnated with lignin (Reiter 2002). Fast-growing *Populus* and *Eucalyptus* tree species are excellent sources of lignocellulosic biomass, as their secondary cell walls are rich in cellulose comprising 39 - 48% of biomass (Carroll and Somerville 2009).

While there are many advantages to employing woody species as biomass feedstock, the recalcitrance of plant cell walls to degradation confounds efficient extraction of polysaccharides (Himmel et al. 2007; Mansfield 2009). This is mainly due to the physical interactions between major biopolymer constituents of the cell wall namely; cellulose, hemicellulose and lignin (Chang and Holtzapple 2000; Cosgrove 2005; Gilbert 2010). The genetic contribution to recalcitrance is to a large extent the result of an underlying, highly coordinated biosynthetic program comprising biosynthetic, modifying and structural proteins which are regulated by a network of transcription factors, hormones and signaling proteins, ultimately resulting in the cell wall ultrastructure (Keegstra 2010; Plomion et al. 2001; Showalter 1993; Somerville et al. 2004). Functional characterization of genes contributing to the plant cell wall ultrastructure is a prerequisite to the design of strategies for engineering customized phenotypes (Galperin and Koonin 2010; Mansfield 2009). This has been demonstrated *in planta*, for example, transgenic *Populus* targeting known lignin biosynthetic-related genes caused a shift in biomass recalcitrance (Mansfield et al. 2012; Nookaraju et al. 2013; Pilate et al. 2002). Furthermore, appropriate growth conditions can promote

extensive secondary growth in the model herbaceous plant *Arabidopsis* (Chaffey et al. 2002), facilitating functional testing of candidate genes implicated in cell wall biology.

Despite considerable research efforts focused on cell wall development and many genes empirically validated relating to lignin biosynthetic pathways (Boerjan et al. 2003; Vanholme et al. 2012a; Vanholme et al. 2008) and polysaccharide biosynthesis (Atmodjo et al. 2013; Doering et al. 2012; Taylor 2008), it is estimated that approximately 10-15% of ~27 000 protein coding genes in the *Arabidopsis* genome are involved in cell wall biology (Carpita et al. 2001; McCann and Carpita 2008; Yong et al. 2005). A study by Yang et al. (2011) found 121 experimentally validated cell wall-related genes via PubMed text mining. The discrepancy between the number of genes functionally validated and those implicated in cell wall biology highlights a clear gap between what is known and what remains to be discovered. Plant bioinformatics has been applied in meta-analyses of genome-wide expression data, resulting in the identification of genes putatively involved in cell wall biology. In many of these studies, a number of implicated genes lack any functional annotation (biochemical or cellular function) and are broadly characterized as proteins of unknown function (PUFs) (Horan et al. 2008). These genes, in the context of cell wall biology referred to as cell wall-related PUFs (CW-PUFs), show similar expression patterns to known cell wall-related genes such as those encoding cellulose synthase, enzymes essential for cellulose biosynthesis (Brown et al. 2005; Mutwil et al. 2009; Persson et al. 2005; Ruprecht et al. 2011). Although these CW-PUFs may not be directly involved in cell wall biosynthesis, many could be essential for normal cell wall developmental biology. In recent years CW-PUFs have been the focus of many functional studies and have been shown to be involved in various aspects of cell wall biology (Bischoff et al. 2010a; Brown et al. 2011; Gille et al. 2011; Jensen et al. 2011; Ruprecht et al. 2011; Urbanowicz et al. 2012). These studies highlight the importance of understanding the biological roles of other CW-PUFs, and suggest aspects of the biosynthetic pathway and/or alternative pathways of wall development not yet characterized.

Functional analysis of CW-PUFs requires gene prioritization i.e., identification and functional prediction of these genes to guide a hypothesis-driven study. While the rapid accumulation of plant genomic data has revealed a large number of genes encoding PUFs, a challenge is sieving through the plethora of genomic

data to identify PUFs relating to wall function. The problem is further exacerbated by limited background information on PUFs to guide experimental design. In an attempt to address these problems, we identified different types of CW-PUFs based on what is known about the function of the protein and its association with cell wall biology. To assist with experimental design, we also discuss relevant and appropriate plant *omics*-based approaches, which may provide functional insight for CW-PUFs selection, thereby facilitating gene prioritization.

1.3. Classification of CW-PUFs types

CW-PUFs refer to any functionally un-annotated proteins which have some evidence linking them to cell wall biology. These proteins generally only have homology-based annotations in plant genomes. Using this approach, a gene is annotated as a PUF if it lacks significant sequence similarity to any previously characterized protein family within a reference database (Boeckmann et al. 2003; Punta et al. 2012). This annotation approach remains one of the default methodologies of many genome initiatives including the plant comparative genomic resource, Phytozome (Goodstein et al. 2012) (http://www.phytozome.net/). Mining candidate cell wall-related gene lists (Bayer et al. 2006; Brown et al. 2005; Capron et al. 2013; Ko et al. 2006; Mentzen and Wurtele 2008; Minic et al. 2009; Mutwil et al. 2009; Mutwil et al. 2010; Ohashi-Ito et al. 2010; Oikawa et al. 2010; Persson et al. 2005; Ruprecht et al. 2011; Yamaguchi et al. 2011; Yang et al. 2011; Zhong et al. 2010) for PUFs revealed two categories of proteins annotated using the similarity approach, i) proteins of known function, and ii) PUFs (Supplementary Figure S1.1). Some proteins grouped within the proteins of known function category do not have experimental evidence demonstrating function in cell wall biology yet are implicated in this process typically based on co-expression analysis. Such proteins may even have validated roles in other biological processes. We refer to this group as CW-known unknown (abbreviated from *known* functional annotation yet *unknown* cell wall function). CW-known unknown may include proteins with previously annotated domains or lack previously identified domains/motifs (referred to as proteins of obscure features (POFs) (Gollery et al. 2007)). The PUF category may be further subdivided into protein families which contain a Hidden Markov Model (HMM)-recognized domains of unknown function (DUFs), or do not contain any recognized domains i.e. POFs (Table 1.1, Supplementary Figure S1.1 and Appendix A, Supplementary Table S1.1/2). Finally,

proteins of unknown function and possibly some proteins of known function may lack sequence homology to other Open Reading Frames (ORFs) currently available on curated databases and are referred to as ORFans (Fischer and Eisenberg 1999). These proteins may have a phyletic distribution limited to an organism or group of organisms (Siew et al. 2004; Siew and Fischer 2003). This is the most difficult class of PUFs to characterize since there is no sequence homology to other genomes especially in non-model organism where mutants will not be available or that are not amenable to genetic transformation. Since we do not expect a major role for this type of PUF in a conserved biological process such as cell wall formation ORFans will not be discussed.

1.3.1. CW-known unknown proteins

The concept of 'known-unknown' was previously proposed for hypothetical proteins for which there is a predicted biochemical function or cellular function (Galperin and Koonin 2004; Galperin and Koonin 2010). Examples of such proteins include the protein kinase superfamily, methyltransferase-related proteins, small GTP-binding proteins and heat shock proteins to mention only a few. In the context of cell wall development, many proteins can be identified with *known* annotation, but *unknown* roles in wall biology. The benefit of CW-known unknowns is that there are at least functional clues for hypothesis generation and experimental design (Galperin and Koonin 2010). A previous example of a CW-known unknown protein is the eukaryotic translation elongation factor eEF-1Bβ1. eEF-1Bβ1 is a guanine exchange factor, yet gene silencing in *Arabidopsis* caused several wall defects including reduction in both lignin and cellulose (Hossain et al. 2012). Oikawa et al. (2010) also highlighted several other CW-known unknown genes (including calmodulin-binding proteins, putative methyltransferases, protein kinases and transmembrane receptors) where there is a general biochemical functional annotation, yet precise functional role in cell wall biology remains elusive. A generalized biochemical and cellular identity does not necessarily reveal the functions of these genes in cell wall biosynthesis and therefore further investigation in a cell wall context is still required.

1.3.2. Domains of unknown function (DUFs)

DUFs refer to conserved protein domains within the Pfam database that lack functional assignment. The advantage of these HMM-recognized motifs is perhaps associating distantly related proteins by domain presence. This in-turn allows for the functional assignment to a protein family rather than individual members, a suggested robust method of annotating molecular function (Galperin and Koonin 2010). Thus, functional elucidation of a single member of a particular DUF family results in the renaming of the entire family accordingly (Punta et al. 2012). In recent years, several DUFs have been implicated in cell wall development (Appendix A, Supplementary Table S1.1). One of the first of these was a rice brittle culm mutant, *bc10*, isolated by Zhou et al. (2009). Map-based cloning of *BC10* revealed that the gene encodes a protein containing DUF266 which is localized to the Golgi. Relative to the wildtype, *bc10* had reduced cellulose and arabinogalactan protein levels, which is believed to underlie the decreased mechanical strength and retarded growth of the mutant plants. *BC10* showed similarities to β-1,6-N-acetylglucosaminyltransferase and enzymatic assays showed glycosyltransferase (GT) activity. It was postulated that *BC10* may be involved in the glycosylation of cellulose synthase proteins or related proteins and/or the arabinogalactan proteins. Hansen et al. (2009) performed an *in silico* proteomics study, including Hidden Markov Models (HMM) and three-dimensional fold recognition, to identify new plant GTs in *Arabidopsis*. One of their main findings was that DUF266 and DUF246 contain GT signatures found in the CAZyme families, GT14 and 65, respectively. In 2011, Ye et al. proposed that the GT14 and DUF266 be incorporated into a new GT14/GT14-like gene family since GT14 and GT14-like displayed similarities in gene expression, protein domains and tertiary structure. The authors further reported that several GT14/GT14-like genes from *Arabidopsis* and *Populus* have a stem/xylem expression preference respectively which is suggestive of a possible cell wall-related function (Ye et al. 2011). DUF266 and DUF246 have recently been merged into the Pfam Branch domain (PF02485) and O-fucosyltransferase domain (PF10250) families, respectively. Unlike other members of these families, DUF266- and DUF246-containing proteins are plant-specific, implying plant-specific biological roles for these proteins (Hansen et al. 2012). Furthermore, members of the DUF266 and DUF246 family are co-expressed with several cell wall-related GTs (Oikawa et al. 2010) and recently it was further hypothesized

that DUF246-containing proteins may function as rhamnogalacturonan-II fucosyltransferases (Hansen et al. 2012).

Another plant-specific DUF, (DUF231, Bischoff et al. (2010a)) is part of the TRICHROME BIREFRIGENCE (TBR) and TBR-like (TBL) proteins belonging to a large family comprising 46 members in *Arabidopsis*. Proteins containing DUF231 have been shown to be related to several biological processes including pathogenesis (*POWDERY MILDEW RESISTANCE 5*) (Vogel et al. 2004), cold stress (*ESKIMO1*) (Xin et al. 2007) and aspects of cell wall biology (Bischoff et al. 2010a; Gille et al. 2011; Lefebvre et al. 2011; Vogel et al. 2004; Yuan et al. 2013). More specifically, the cell walls of *powdery mildew resistance 5* knockdown mutants had a lower degree of pectin methyl-esterification or O-acetylation relative to the wildtype control (Vogel et al. 2004). In 2010a, Bischoff et al. showed that mutants of DUF231 members, TRICHOME BIREFRINGENCE (TBR) and TBR-like3 (TBL3) contained lower cellulose and altered pectin composition. The DUF231 domain has been renamed to *PC-Esterase* (PF13839); however, Bischoff et al. (2010b) reported that TBL/DUF231 is unlikely to be a catalytically active esterase, despite the presence of a conserved DxxH motif, also present in esterases (e.g. fungal rhamnoglacturonan acetylesterase), and the TBL motif of DUF231-containing proteins (Bischoff et al. 2010b). The authors did however hypothesize that DUF231 proteins may function as pectin binding proteins or as bridging proteins which may bind and crosslink pectin and other cell wall polysaccharides. Recently, DUF231-containing proteins TBL27 and TBL22 were shown to be involved in xyloglucan O-acetylation (Gille et al. 2011), while Yuan et al. (2013) demonstrated that *ESKIMO1* functions in O-acetylation of xylan during biosynthesis of the secondary cell wall. This raises the question of whether DUF231 family members may play a role in the acetylation of other cell wall biopolymers, as suggested by Gille et al. (2011).

In 2011, two members of DUF579 family (referred to as IRREGULAR XYLEM 15 (IRX15) and IRX15-LIKE/IRX15L) were shown to play a role in xylan biosynthesis (Brown et al. 2011; Jensen et al. 2011). The single (*irx15* and *irx15l*) and double knockdown (*irx15 irx15*) mutants showed pleiotropic phenotypes including decrease in xylan content and chain length, almost undetectable glucuronic acid side chain content and methylation of the GlcA side chains. However, this was not consistent with the phenotype of

three additional DUF579-containing proteins, GLUCURONOXYLAN METHYLTRANSFERASE1 (GXM1), GXM2 and GXM3/GXMT. These plants did not display the characteristic IRX15 and IRX15L phenotype and the genes were further shown to be responsible for 4-O-methylation of glucuronic acid to form MeGlcA side chains of xylan (Lee et al. 2012; Urbanowicz et al. 2012). As such, the exact molecular function of IRX15 and IRX15L is still unknown.

In 2006 and 2008, members of the DUF642 were identified in two cell wall proteomic studies (Bayer et al. 2006; Irshad et al. 2008). Later, Gao et al. (2011) showed that DUF642 members, At1g80240 and At5g25460, were up-regulated when *Arabidopsis* ascorbic acid deficient mutants were fed with L-Galactono-1,4-lactone, a precursor in the ascorbic acid biosynthetic pathway (Gao et al. 2011). Gene knockdown mutants of At5g25460 displayed smaller rosettes and shorter roots relative to the wildtype control, while no observable phenotypes were present in the At1g80240 mutants (Gao et al. 2012). The authors hypothesized that At5g25460 may play a role in cell wall pectin dynamics. It was proposed that boron is necessary for the structural integrity of the pectin polysaccharides through the formation of boron-pectin complexes (Hu and Brown 1994; Hu et al. 1996; Loomis and Durst 1992) and it was also shown that At5g25460 was up-regulated in response to boric acid (Gao et al. 2012; Zimmermann et al. 2004). This in turn may influence plant growth through cell wall extensibility and hence explaining the mutant phenotype (Gao et al. 2012). This hypothesis was corroborated by a second study which showed *in vitro* that two DUF642-containing proteins (At5g11420 and At4g32460) interact with a pectin methylesterase protein (Zúñiga-Sánchez and Gamboa-de Buen 2012). This study also identified two additional interactors, including a leucine-rich repeat protein (FLOR1) and a vegetative storage protein. Concurrently, Vázquez-Lobo et al. (2012) showed *in vitro* that an *Arabidopsis* DUF642-containing protein (At3g08030) interacts with cellulose, specifically cellobiose and to a lesser extent hemicellulose. The study showed that DUF642-containing proteins were absent from non-seed-plant genomes, while some family members had predicted galactose-binding domains-like and glycosylphosphatidylinositol anchor sites. Although evidence suggest that DUF642 may play a role in modification or strengthening of cell walls through the interaction with cell wall polysaccharides, the biological and molecular function remains elusive (Vázquez-Lobo et al. 2012).

1.3.3. Proteins of obscure features (POFs)

Whereas DUFs have recognized domains, POFs refer to a large number of proteins that lack such domains or motifs (Gollery et al. 2006; Gollery et al. 2007). These proteins may be found in known or unknown proteins (Supplementary Figure S1.1). Gollery et al. (2007) showed that these proteins constituted approximately 19%, 27% and 33% of the predicted proteome of *Arabidopsis*, *Populus* and rice, respectively. POFs contain regions with high propensity of disorder which are highly correlated with protein-protein interaction as well as conformational flexibility and variability allowing regulatory functioning with binding partners (Gollery et al. 2006; Gsponer and Madan Babu 2009; Sugase et al. 2007; Tompa 2002). The large number of POFs in plant genomes suggests diverse biological roles. One of these may be a role in stress responses. Luhua et al. (2008) demonstrated the roles of several POFs in stress responses in *Arabidopsis* and found that overexpression conferred single stress type tolerance in some cases while in other cases was associated with susceptibility. Two *Arabidopsis* POFs conferred enhanced tolerance to osmotic stress (Luhua et al. 2008). POFs that have been implicated in cell wall formation through a number of cell wall related studies are listed in Appendix A, Supplementary Table S1.2.

Unlike DUF-containing proteins, where functional inference can be extended via the conserved domains to distantly related proteins, functional inferences for POFs only extend to closely related family members with significant sequence similarity across the entire protein. While some POFs may be inherently disordered and characterized by the absence of a folded structure (referred to as intrinsically unstructured proteins, Tompa (2002)); across-species comparative genomics may facilitate the generation of HMM profiles for shared POF homologs allowing for defined signatures for the family and concomitant removal of these proteins from the POF category and reclassification as DUFs (Gollery et al. 2006; Gollery et al. 2007).

1.4. In search of a function for CW-PUFs

A decrease in cost and an increase in throughput of *next generation* sequencing technologies have fueled genome-wide sequencing efforts, resulting in a steady increase in the number of publicly available

genomes. Currently, 41 sequenced and annotated plant genomes are accessible through Phytozome. The result has been an increase in the number of predicted gene models, many of which are PUFs. According the version 10 release of The Arabidopsis Information Resource, between 26 to 34% of the protein coding genes have unknown molecular function, biological process and cellular component according to Gene Ontology annotation (Lamesch et al. 2012). Given that the aforementioned percentages does not exclusively relate to cell wall biology, a general problem is identifying and prioritizing PUFs related to this biological process for further functional characterization. Several studies have generated cell wall-related candidate gene lists thereby implicating a number of genes including PUFs (Table 1.1) however; prioritization of PUFs for functional characterization is a challenge due to the lack of hypothesized biological or molecular functions to guide experimental design. Despite the availability of commercially available knockout lines in *Arabidopsis* (Alonso et al. 2003; Kuromori et al. 2004; Rosso et al. 2003; Sessions et al. 2002), the challenge is finding appropriate experimental conditions to allow expression of the phenotype. This is especially important given the complexity of plant cell walls, combined with the diversity of phenotypes that may be linked to cell wall-related genes (Lloyd and Meinke 2012).

Furthermore, many PUFs may be part of larger families suggesting functional redundancy. It is therefore understandable why only a handful of PUFs have elucidated biological and/or molecular functions e.g. DUF231 (Gille et al. 2011; Yuan et al. 2013) and DUF579 (Lee et al. 2012; Urbanowicz et al. 2012). Integrated analysis of plant omics data is widely used for functional annotation in *Arabidopsis* (Bradford et al. 2010; Clare et al. 2006; Heyndrickx and Vandepoele 2012; Kourmpetis et al. 2011; Lan et al. 2007; Lee et al. 2010; Warde-Farley et al. 2010). Such studies allow high-throughput functional prediction of PUFs and demonstrate that integrated analysis of different *omics* data types leads to a higher accuracy of prediction as confirmed by subsequent experimental validation (Bradford et al. 2010; Lan et al. 2007; Lee et al. 2010). There are many plant omic based approaches which can be can be used for functional inference for the different types of CW-PUFs by means of guilt-by-association (Figure 1.1). Intrinsic properties such as co-expression patterns, shared promoter *cis* elements, protein domains, etc., may provide valuable functional clues for guiding experimental design.

1.4.1. Prioritizing CW-PUFs with genomic analyses

Genes that are involved in the same biological process and under the common transcriptional regulation often share *cis*-regulatory elements which can be used to predict biological function for unknown proteins (Vandepoele et al. 2009). For a largely conserved developmental processes such as cell wall biosynthesis, these regulatory elements are also found to be conserved across species and genera provided the orthologous transcription factors and target genes are present (Creux et al. 2013; Freeling and Subramaniam 2009; Lockton and Gaut 2005). For example, Ding et al. (2012) found that approximately 4% of the total shared *cis*-regulatory sequences were probably regulating genes relating to bioenergy including cell wall biosynthetic genes.

Genetic mapping of quantitative trait loci (QTL) is a method to identify genomic regions associated with phenotypic variation. Such studies can also be used to identify candidate genes co-locating with QTLs of interest such as lignin content, syringyl:guaiacyl ratio (S:G), cellulose pulp yield and fibre length (Capron et al. 2013; Chavigneau et al. 2012; Courtial et al. 2012; Courtial et al. 2013; Kullan et al. 2012; Price 2006; Ranjan et al. 2010). However, the resolution of QTL mapping is very low which means that QTL intervals include hundreds of genes typically including CW-PUFs. Therefore to prioritize positional candidate CW-PUFs, other data types such as expression and metabolite profiles can be used in a genetical genomics approach which aims to associate polymorphism in transcriptome and metabolome with phenotypic QTL (Jansen and Nap 2001). This can be used to identify and narrow down candidate genes affecting the phenotypic variation of a given trait of interest (Breitling et al. 2008; Hansen et al. 2008).

1.4.2. Functional inference of CW-PUFs through transcriptomics

High-throughput transcriptome profiling technologies such as mRNA-Seq and microarray have facilitated genome-wide identification of genes expressed during cell wall formation (Table 1.1 and Andersson Gunnerås et al. 2006; Demura et al. 2002; Hertzberg et al. 2001; Mizrachi et al. 2010; Schrader et al. 2004). The approach has important applicability to the functional annotation of PUFs as it provides a framework for putatively linking these genes to biological processes, and in the case of cell wall biology,

expression data may delineate genes specific to major cell wall-forming organs, tissues and cell types such as stems in herbaceous plants (Brown et al. 2005; Ehlting et al. 2005; Persson et al. 2005) or xylem tissue in woody perennials (Hertzberg et al. 2001; Mizrachi et al. 2010). Indeed, such studies have reported that many PUFs were differentially expressed in cell wall-forming tissue (Table 1.1 and Appendix A, Supplementary Table S1.1/2). Note that Table 1.1 contains mainly secondary cell wall associated genes due to the strong transcriptional co-regulation of secondary cell wall genes, and hence high expression correlation observed for these genes in tissue specific expression profiling. A comparison of stem xylem to leaf transcriptomic data from *Eucalyptus grandis* and *Populus trichocarpa* (Hefer, et al., in preparation) revealed that, of the 1055 and 1390 stem xylem up-regulated genes, 199 (~19%) and 234 (~17%) genes encoded PUFs in eucalyptus and poplar respectively (≥ 1.5 expression preference in xylem tissue compared to leaf tissue, Figure 1.2A). Furthermore, most of the stem xylem-preferential PUF genes in *E. grandis* and *P. trichocarpa* encoded POFs (126 and 146 respectively) while 73 and 88 contained DUFs respectively (Figure 1.2B). Several xylem-specific PUFs were identified in a recent study conducted in *Populus* in order to develop xylem-specific utility promoters (Ko et al. 2012). Two of the four PUFs, which had a 10-fold higher expression in developing xylem in the Ko et al. (2012) study, were also identified in Hefer, et al. (in preparation). In 2011, Ohtani et al. identified 63 differentially expressed poplar PUFs, which were up- or down-regulated by the *Arabidopsis* NAC domain protein VND7, a master regulator of xylem vessel formation.

Gene co-expression is valuable for functional prediction of genes based on the guilt-by-association principle (Walker et al. 1999). In gene co-expression, a query gene (bait) displaying similar expression patterning with functionally known genes will have an increased likelihood of sharing a regulatory pathway and thus, may be functionally related (Aoki et al. 2007; Saito et al. 2008). For example, Brown et al. (2005) and Persson et al. (2005) independently used the known secondary cell wall-specific cellulose synthase genes as bait genes to identify other highly co-expressed genes. The versatility of the gene co-expression approach is demonstrated at a genome-wide level where modules of highly correlated genes, within a co-expressed gene network, may be linked to particular biological processes due to overrepresentation of functional categories. In this way, biological function may be inferred for unknown

genes within a module (Aoki et al. 2007; Obayashi et al. 2011). Examples of publically available co-expression databases include ATTED-II (Obayashi et al. 2011), PlaNet (Mutwil et al. 2011) and AraNet (Lee et al. 2010). A genome-wide annotation of *Arabidopsis* PUFs was conducted by Horan et al. (2008) by clustering highly co-expressed known and unknown genes followed by annotation of resulting modules using GO enrichment. Thus, functional enrichment could be ascribed for a total of 1 541 PUFs where 16 PUFs were found in specific cell wall clusters (Horan et al. 2008).

Support for functional annotation via co-expression network analysis can also be obtained by identifying conserved network components across species (Bergmann et al. 2003; Ficklin and Feltus 2011; Mutwil et al. 2011). Oikawa et al. (2010) compared the co-expression gene network of xylan-related glycoside transferases from rice and *Arabidopsis* to identify novel genes involved in xylan biosynthesis (approximately 30% of high ranking co-expressed genes encoded PUFs), while Ruprecht et al. (2011) compared the co-expression networks of primary and secondary cellulose synthase genes across seven distantly related plant species. In the latter study, 13% of the 376 gene families identified were PUFs and a subset of these unknown proteins was then experimentally validated using reverse genetics approaches. This study identified several PUFs affecting cell wall polysaccharides, thereby supporting a role for these PUFs in cell wall biology (Ruprecht et al. 2011). Moreover, due to functional redundancy associated with large protein families, across-species expression profiling and expression correlation in combination with phylogeny are powerful tools to prioritize functionally equivalent orthologs across species (Mutwil et al. 2011; Patel et al. 2012). The basis of the approach is the assumption that the expression profile and gene co-expression pattern of orthologous genes with similar functions should be conserved even if they are not closest phylogenetic neighbors (Bergmann et al. 2003; Ruprecht et al. 2011).

1.4.3. Prioritizing CW-PUFs with proteomics

Cell wall proteins (CWP) refer to proteins localized to the plant cell wall (Jamet et al. 2006) which are highly likely to be involved in aspects of wall development. Therefore, a list of CWP may facilitate the simultaneous identification and prioritization of candidate CW-PUFs for functional studies. The generation

of a comprehensive list of CWP can be challenging, confounded by contamination from other cellular compartments as well as the inherent characteristics of proteins such as varying post-translational modifications, physiochemical properties and protein turnover rates which results in a biased capture of the CWP (Jamet et al. 2008; Rose and Lee 2010). Despite this, several attempts at the isolation of the CWP from species such as *Arabidopsis*, *Brassica oleracea, Oryza sativa* and *Nicotiana tabacum* have been made (Albenne et al. 2009; Borderies et al. 2003; Chen et al. 2009; Chivasa et al. 2002; Feiz et al. 2006; Ligat et al. 2011; Millar et al. 2009). *WallProtDB*, an online database (http://www.polebio.lrsv.ups-tlse.fr/WallProtDB/index.php), cumulative of several studies, lists approximately 500 and 800 CWP of *Arabidopsis* and rice respectively according to known function or conserved domains (Pont-Lezica et al. 2010). CW-PUFs accounts for approximately 12.2% of *Arabidopsis* wall proteins found in *WallProtDB* (Pont-Lezica et al. 2010). However, not all proteins involved in cell wall biogenesis are found within the apoplastic environment, for example, the matrix polysaccharide xylan is synthesized in the Golgi apparatus (Bolwell and Northcote 1983). Recently, Parsons et al. (2012) found that 13% of 371 proteins from the *Arabidopsis* Golgi proteome, isolated from protoplast cultures were PUFs, interestingly including an additional as yet un-described DUF579-containing protein (At1g27930).

Finally, annotation of predicted protein features forms an invaluable tool for functional inference and consequently, PUF prioritization. While the similarity approach might classify a protein as unknown if it does not display homology to a protein of known function, properties at the protein level such as conserved domain, tertiary structure and protein-protein interaction may provide functional clues. Prediction of subcellular localization of candidate proteins are an important step toward understanding their biological role (Chou and Cai 2003; Rost et al. 2003). Different subcellular localities have distinct proteomes defining the subcellular compartment (Heazlewood et al. 2007; Oikawa et al. 2010). The finding that xylan-related and lignin-related proteins are found in the Golgi apparatus (Bolwell and Northcote 1983) and endoplasmic reticulum (Boerjan et al. 2003; Chapple 1998; Ro et al. 2001) respectively corroborates this notion. There are many online prediction tools for subcellular localization available based on sorting signals, protein primary structure and experimental validation (Emanuelsson et al. 2007; Heazlewood et al. 2007). Oikawa et al. (2010) developed an algorithm, PFANTOM, based on

information from protein functional domains to predict subcellular localization of proteins found in endomembrane systems including the plasma membrane, the prediction of which is not reliable with current prediction tools.

Another important characteristic of proteins is the presence of domains, which fold and function independently of the constituting protein. While domain architecture may not lead to elucidation of biochemical activity, it may link biological process by means of domain co-occurrence and therefore an inference of conserved function (Bosgraaf and Van Haastert 2003; Leipe et al. 2002). Domain co-occurrence has particular applicability to DUFs as functionally known domains in combination with DUFs may guide hypothesis generation regarding function of the candidate domain-containing gene family. In Pfam 26.0, approximately 23% of total DUFs co-occurred with a functionally annotated domain (Punta et al. 2012). Currently, the annotation of conserved protein domains is incorporated into most, if not all, genome annotation pipelines including plant genome databases such as Phytozome.

Protein sequence motifs from *de novo* discoveries can be used to infer function on more distantly related proteins sharing the same motifs. An example of this is the TONNEAU1 Recruiting Motif (TRM) protein family which was delineated by *de novo* motif discovery (Drevensek et al. 2012). The protein family members could be grouped into eight subgroups with no significant similarity between groups apart from the conserved motifs. At the tertiary level, protein three-dimensional fold may also be useful in functional assignment since protein structure responsible for functionality is more conserved and thus more divergent relationships may be observed at this level (Bornberg-Bauer et al. 2005; Eddy 1998). Furthermore, the order of domains may also not necessarily be maintained at the sequence level and this may present a problem for conventional sequence searches (Amoutzias et al. 2004; Bornberg-Bauer et al. 2005; Kersting et al. 2012). The methodology has been useful in functional prediction of DUF266 and DUF246-containing proteins as putative GTs (Hansen et al. 2009; Hansen et al. 2012).

Finally, almost all biological processes are governed by a dynamic myriad of protein-protein interactions that culminate in a stable phenotype (Bork et al. 2004; Cusick et al. 2005; Vidal et al. 2011). As previously mentioned, two DUF642-containing proteins were shown to interact *in vitro* with a pectin methylesterase

protein, a leucine-rich repeat protein (FLOR1) and a vegetative storage protein (Zúñiga-Sánchez and Gamboa-de Buen 2012). It was also previously shown that the cellulose synthase-interactive protein1 (CSI1) interacts with cellulose synthase (Gu et al. 2010) thereby bridging the cellulose synthase complexes with cortical microtubules (Li et al. 2012). Therefore, a protein interactome is useful as it serves as an addendum to the gene co-expression guilt-by-association principle and can be used to ascribe putative function to PUFs for hypothesis testing (Schauer and Stingl 2010). A list of plant interactome resources can be found in Mochida and Shinozaki (2011). A number of studies have integrated protein interaction data with other *omics*-based datatypes in *Arabidopsis* to predict function of PUFs (Bradford et al. 2010; Heyndrickx and Vandepoele 2012; Kourmpetis et al. 2011; Lee et al. 2010), with success of each methodology dependent on the data source and integration algorithms.

1.5. Concluding remarks

The continuation of genome and transcriptome sequencing efforts will undoubtedly increase the number of predicted PUFs in plants, which will require functional characterization. Despite advances in understanding cell wall biology, there still remain many implicated proteins for which function has not been ascribed. Unraveling function for CW-PUFs is particularly challenging, as we do not know where to begin looking except for obvious morphological and cell wall chemistry phenotypes. Furthermore, large protein families may contribute to functional redundancy, as was seen with the REDUCED WALL ACETYLATION genes (Lee et al. 2011), which may complicate functional elucidation. The above necessitates the integration of complementary *omics*-based datasets to provide functional insight to guide hypothesis-driven experimentation for CW-PUFs. This, in combination with system biology approaches to studying metabolic pathways (Vanholme et al. 2012b) and system genetics approaches in biomass crops combining transcript, protein and metabolite data to explain trait variation at a population level (Mackay et al. 2009; Mizrachi et al. 2012) may consequently bring the research community closer to bridging gaps in our molecular understanding of cell wall development and structure.

1.6. Aim of the current study

While the long-term goal of this study is to contribute to the understanding of cell wall biology, this dissertation aimed at identifying candidate CW-PUFs in the most widely planted hardwood genus,

Eucalyptus, and functionally characterizing these genes in *Arabidopsis thaliana*. Genes preferentially expressed in *Eucalyptus* wood-depositing xylem tissue (Figure 1.2) were consolidated with *Populus* and *Arabidopsis* data (Figure 1.2 and Table 1.1, respectively), and further prioritized using the meta-analysis platform proposed in Figure 1.1. These efforts resulted in candidate genes from the DUF1218 family and a single POF for functional characterization. Based on the biological evidence for prioritizing these candidate genes, we postulate the importance of these genes in cell wall biology and hypothesize that gene perturbation may result in the modification of the cell wall composition. To test this hypothesis, I used reverse genetics in *Arabidopsis* to analyse growth and cell wall chemistry for several candidate genes. Furthermore, the study of each gene was complimented with promoter activity studies as well as confocal microscopy to determine the subcellular localization for protein function.

1.7. References

Albenne, C., Canut, H., Boudart, G., Zhang, Y., San Clemente, H., Pont-Lezica, R., et al. (2009) Plant cell wall proteomics: mass spectrometry data, a trove for research on protein structure/function relationships. *Molecular Plant* 2: 977-989.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301: 653.

Amoutzias, G.D., Robertson, D.L., Oliver, S.G. and Bornberg-Bauer, E. (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Reports* 5: 274-279.

Andersson Gunnerås, S., Mellerowicz, E.J., Love, J., Segerman, B., Ohmiya, Y., Coutinho, P.M., et al. (2006) Biosynthesis of cellulose enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *The Plant Journal* 45: 144-165.

Aoki, K., Ogata, Y. and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology* 48: 381-390.

Asif, M. and Muneer, T. (2007) Energy supply, its demand and security issues for developed and emerging economies. *Renewable and Sustainable Energy Reviews* 11: 1388-1413.

Atmodjo, M.A., Hao, Z. and Mohnen, D. (2013) Evolving views of pectin biosynthesis. *Annual Review of Plant Biology* 64: 747-779.

Bayer, E.M., Bottrill, A.R., Walshaw, J., Vigouroux, M., Naldrett, M.J., Thomas, C.L., et al. (2006) *Arabidopsis* cell wall proteome defined using multidimensional protein identification technology. *Proteomics* 6: 301-311.

Bergmann, S., Ihmels, J. and Barkai, N. (2003) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology* 2: 0085-0093.

Bischoff, V., Nita, S., Neumetzler, L., Schindelasch, D., Urbain, A., Eshed, R., et al. (2010a) *TRICHOME BIREFRINGENCE* and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. *Plant Physiology* 153: 590.

Bischoff, V., Selbig, J. and Scheible, W.R. (2010b) Involvement of TBL/DUF231 proteins into cell wall biology. *Plant Signaling & Behavior* 5: 1057.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31: 365-370.

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

Bolwell, G.P. and Northcote, D.H. (1983) Arabinan synthase and xylan synthase activities of Phaseolus vulgaris. Subcellular localization and possible mechanism of action. *Biochemical Journal* 210: 497.

Borderies, G., Jamet, E., Lafitte, C., Rossignol, M., Jauneau, A., Boudart, G., et al. (2003) Proteomics of loosely bound cell wall proteins of *Arabidopsis thaliana* cell suspension cultures: a critical analysis. *Electrophoresis* 34: 3421-3432.

Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (2004) Protein interaction networks from yeast to human. *Current Opinion in Structural Biology* 14: 292-299.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A. and Weiner, J. (2005) The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences* 62: 435-445.

Bosgraaf, L. and Van Haastert, P.J.M. (2003) Roc, a Ras/GTPase domain in complex proteins. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1643: 5-10.

Bradford, J.R., Needham, C.J., Tedder, P., Care, M.A., Bulpitt, A.J. and Westhead, D.R. (2010) GO-At: in silico prediction of gene function in *Arabidopsis thaliana* by combining heterogeneous data. *The Plant Journal* 61: 713-721.

Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., et al. (2008) Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics* 4: e1000232.

Brown, D., Wightman, R., Zhang, Z., Gomez, L.D., Atanassov, I., Bukowski, J.P., et al. (2011) *Arabidopsis* genes *IRREGULAR XYLEM* (*IRX15*) and *IRX15L* encode DUF579 containing proteins that are essential for normal xylan deposition in the secondary cell wall. *The Plant Journal* 66: 401-413.

Brown, D.M., Zeef, L.A.H., Ellis, J., Goodacre, R. and Turner, S.R. (2005) Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. *The Plant Cell Online* 17: 2281.

Capron, A., Chang, X.F., Hall, H., Ellis, B., Beatson, R.P. and Berleth, T. (2013) Identification of quantitative trait loci controlling fibre length and lignin content in *Arabidopsis thaliana* stems. *Journal of Experimental Botany* 64: 185-197.

Carpita, N., Tierney, M. and Campbell, M. (2001) Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. *Plant Molecular Biology* 47: 1-5.

Carroll, A. and Somerville, C. (2009) Cellulosic biofuels. *Annual Review of Plant Biology* 60: 165-182.

Cassman, C.K. and Liska, A.J. (2007) Food and fuel for all: realistic or foolish? *Biofuels, Bioproducts and Biorefining* 1: 18-23.

Chaffey, N., Cholewa, E., Regan, S. and Sundberg, B. (2002) Secondary xylem development in *Arabidopsis*: a model for wood formation. *Physiologia Plantarum* 114: 594-600.

Chang, V.S. and Holtzapple, M.T. (2000) Fundamental factors affecting biomass enzymatic reactivity. *Applied Biochemistry and Biotechnology* 84: 5-37.

Chapple, C. (1998) Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. *Annual Review of Plant Biology* 49: 311-343.

Chavigneau, H., Goué, N., Delaunay, S., Courtial, A., Jouanin, L., Reymond, M., et al. (2012) QTL for floral stem lignin content and degradability in three recombinant inbred line (RIL) progenies of *Arabidopsis thaliana* and search for candidate genes involved in cell wall biosynthesis and degradability. *Open Journal of Genetics* 2: 7-30.

Chen, X.Y., Kim, S.T., Cho, W.K., Rim, Y., Kim, S., Kim, S.W., et al. (2009) Proteomics of weakly bound cell wall proteins in rice calli. *Journal of Plant Physiology* 166: 675-685.

Chivasa, S., Ndimba, B.K., Simon, W.J., Robertson, D., Yu, X.L., Knox, J.P., et al. (2002) Proteomic analysis of the *Arabidopsis thaliana* cell wall. *Electrophoresis* 23: 1754-1765.

Chou, K.C. and Cai, Y.D. (2003) Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition. *Journal of Cellular Biochemistry* 90: 1250-1260.

Clare, A., Karwath, A., Ougham, H. and King, R.D. (2006) Functional bioinformatics for *Arabidopsis thaliana*. *Bioinformatics* 22: 1130-1136.

Cosgrove, D.J. (1997) Assembly and enlargement of the primary cell wall in plants. *Annual Review of Cell and Developmental Biology* 13: 171-201.

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

Courtial, A., Jourda, C., Arribat, S., Balzergue, S., Huguet, S., Reymond, M., et al. (2012) Comparative expression of cell wall related genes in four maize RILs and one parental line of variable lignin content and cell wall degradability. *Maydica* 57: 56-74.

Courtial, A., Thomas, J., Reymond, M., Méchin, V., Grima-Pettenati, J. and Barrière, Y. (2013) Targeted linkage map densification to improve cell wall related QTL detection and interpretation in maize. *Theoretical and Applied Genetics*: 1-15.

Creux, N.M., De Castro, M.H., Ranik, M., Maleka, M.F. and Myburg, A.A. (2013) Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in Eucalyptus. *Tree Genetics & Genomes*: 1-16.

Cusick, M.E., Klitgord, N., Vidal, M. and Hill, D.E. (2005) Interactome: gateway into systems biology. *Human Molecular Genetics* 14: R171-R181.

Demura, T., Tashiro, G., Horiguchi, G., Kishimoto, N., Kubo, M., Matsuoka, N., et al. (2002) Visualization by comprehensive microarray analysis of gene expression programs during transdifferentiation of mesophyll cells into xylem cells. *Proceedings of the National Academy of Sciences* 99: 15794-15799.

Ding, J., Hu, H. and Li, X. (2012) Thousands of cis-regulatory sequence combinations are shared by *Arabidopsis* and poplar. *Plant Physiology* 158: 145-155.

Doering, A., Lathe, R. and Persson, S. (2012) An update on xylan synthesis. *Molecular plant* 5: 769-771.

Drevensek, S., Goussot, M., Duroc, Y., Christodoulidou, A., Steyaert, S., Schaefer, E., et al. (2012) The *Arabidopsis* TRM1–TON1 interaction reveals a recruitment network common to plant cortical microtubule arrays and eukaryotic centrosomes. *The Plant Cell Online* 24: 178-191.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14: 755-763.

Ehlting, J., Mattheus, N., Aeschliman, D.S., Li, E., Hamberger, B., Cullis, I.F., et al. (2005) Global transcript profiling of primary stems from *Arabidopsis thaliana* identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. *The Plant Journal* 42: 618-640.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* 2: 953-971.

Feiz, L., Irshad, M., Pont-Lezica, R.F., Canut, H. and Jamet, E. (2006) Evaluation of cell wall preparations for proteomics: a new procedure for purifying cell walls from *Arabidopsis* hypocotyls. *Plant Methods* 2: 10.

Ficklin, S.P. and Feltus, F.A. (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiology* 156: 1244-1256.

Fischer, D. and Eisenberg, D. (1999) Finding families for genomic ORFans. *Bioinformatics* 15: 759-762.

Freeling, M. and Subramaniam, S. (2009) Conserved noncoding sequences (CNSs) in higher plants. *Current Opinion in Plant Biology* 12: 126-132.

Galperin, M.Y. and Koonin, E.V. (2004) 'Conserved hypothetical'proteins: prioritization of targets for experimental study. *Nucleic Acids Research* 32: 5452-5463.

Galperin, M.Y. and Koonin, E.V. (2010) From complete genome sequence to 'complete'understanding? *Trends in Biotechnology* 28: 398-406.

Gao, Y., Badejo, A.A., Sawa, Y. and Ishikawa, T. (2012) Analysis of Two L-Galactono-1, 4-Lactone-Responsive Genes with Complementary Expression During the Development of *Arabidopsis thaliana*. *Plant and Cell Physiology* 53: 592-601.

Gao, Y., Nishikawa, H., Badejo, A.A., Shibata, H., Sawa, Y., Nakagawa, T., et al. (2011) Expression of aspartyl protease and C3HC4-type RING zinc finger genes are responsive to ascorbic acid in *Arabidopsis thaliana*. *Journal of Experimental Botany* 62: 3647-3657.

Gilbert, H.J. (2010) The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiology* 153: 444-455.

Gille, S., de Souza, A., Xiong, G., Benz, M., Cheng, K., Schultink, A., et al. (2011) O-acetylation of *Arabidopsis* hemicellulose xyloglucan requires AXY4 or AXY4L, proteins with a TBL and DUF231 domain. *The Plant Cell Online* 23: 4041-4053.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.K., et al. (2006) What makes species unique? The contribution of proteins with obscure features. *Genome Biology* 7: R57.

Gollery, M., Harper, J., Cushman, J., Mittler, T. and Mittler, R. (2007) POFs: what we don't know can hurt us. *Trends in Plant Science* 12: 492-496.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40: D1178-D1186.

Gsponer, J. and Madan Babu, M. (2009) The rules of disorder or why disorder rules. *Progress in Biophysics and Molecular Biology* 99: 94-103.

Gu, Y., Kaplinsky, N., Bringmann, M., Cobb, A., Carroll, A., Sampathkumar, A., et al. (2010) Identification of a cellulose synthase-associated protein required for cellulose biosynthesis. *Proceedings of the National Academy of Sciences* 107: 12866-12871.

Hansen, B.G., Halkier, B.A. and Kliebenstein, D.J. (2008) Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends in Plant Science* 13: 72-77.

Hansen, S.F., Bettler, E., Wimmerova , M., Imberty, A., Lerouxel, O. and Breton, C. (2009) Combination of several bioinformatics approaches for the identification of new putative glycosyltransferases in *Arabidopsis*. *Journal of ProteomeResearch* 8: 743-753.

Hansen, S.F., Harholt, J., Oikawa, A. and Scheller, H.V. (2012) Plant glycosyltransferases beyond CAZy: a perspective on DUF families. *Frontiers in Plant Science* 3: 1-10.

Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I. and Millar, A.H. (2007) SUBA: the Arabidopsis subcellular database. *Nucleic Acids Research* 35: D213-D218.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., et al. (2001) A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* 98: 14732-14737.

Heyndrickx, K.S. and Vandepoele, K. (2012) Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiology* 159: 884-901.

Himmel, M.E., Ding, S.Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W., et al. (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315: 804-807.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., et al. (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiology* 147: 41-57.

Hossain, Z., Amyot, L., McGarvey, B., Gruber, M., Jung, J. and Hannoufa, A. (2012) The Translation Elongation Factor eEF-1Bβ1 Is Involved in Cell Wall Biosynthesis and Plant Development in *Arabidopsis thaliana*. *PloS One* 7: e30425.

Hu, H. and Brown, P.H. (1994) Localization of boron in cell walls of squash and tobacco and its association with pectin (evidence for a structural role of boron in the cell wall). *Plant Physiology* 105: 681-689.

Hu, H., Brown, P.H. and Labavitch, J.M. (1996) Species variability in boron requirement is correlated with cell wall pectin. *Journal of Experimental Botany* 47: 227-232.

Irshad, M., Canut, H., Borderies, G., Pont-Lezica, R. and Jamet, E. (2008) A new picture of cell wall protein dynamics in elongating cells of *Arabidopsis thaliana*: confirmed actors and newcomers. *BMC Plant Biology* 8: 94.

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

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

Jansen, R.C. and Nap, J.-P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* 17: 388-391.

Jensen, J.K., Kim, H., Cocuron, J.C., Orler, R., Ralph, J. and Wilkerson, C.G. (2011) The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in *Arabidopsis*. *The Plant Journal* 66: 387-400.

Karp, A. and Shield, I. (2008) Bioenergy from plants and the sustainable yield challenge. *New Phytologist* 179: 15-32.

Keegstra, K. (2010) Plant Cell Walls. *Plant Physiology* 154: 483.

Kersting, A.R., Bornberg-Bauer, E., Moore, A.D. and Grath, S. (2012) Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome biology and Evolution* 4: 316-329.

Ko, J.-H., Kim, H.-T., Hwang, I. and Han, K.-H. (2012) Tissue-type-specific transcriptome analysis identifies developing xylem-specific promoters in poplar. *Plant Biotechnology Journal* 10: 587-596.

Ko, J.H., Beers, E.P. and Han, K.H. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. *Molecular Genetics and Genomics* 276: 517-531.

Kourmpetis, Y.A.I., van Dijk, A.D.J., van Ham, R.C.H.J. and ter Braak, C.J.F. (2011) Genome-Wide Computational Function Prediction of *Arabidopsis* Proteins by Integration of Multiple Data Sources. *Plant Physiology* 155: 271-281.

Kullan, A.R.K., Van Dyk, M.M., Hefer, C.A., Jones, N., Kanzler, A. and Myburg, A.A. (2012) Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus grandis* and *E. urophylla*. *BMC Genetics* 13: 60.

Kuromori, T., Hirayama, T., Kiyosue, Y., Takabe, H., Mizukado, S., Sakurai, T., et al. (2004) A collection of 11 800 single-copy Ds transposon insertion lines in *Arabidopsis*. *The Plant Journal* 37: 897-905.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40: D1202-D1210.

Lan, H., Carson, R., Provart, N.J. and Bonner, A.J. (2007) Combining classifiers to predict gene function in *Arabidopsis thaliana* using large-scale gene expression measurements. *BMC Bioinformatics* 8: 358.

Lee, C., Teng, Q., Zhong, R. and Ye, Z.-H. (2011) The four *Arabidopsis* REDUCED WALL ACETYLATION genes are expressed in secondary wall-containing cells and required for the acetylation of xylan. *Plant and Cell Physiology* 52: 1289-1301.

Lee, C., Teng, Q., Zhong, R., Yuan, Y., Haghighat, M. and Ye, Z.-H. (2012) Three *Arabidopsis* DUF579 domain-containing GXM proteins are methyltransferases catalyzing 4-O-methylation of glucuronic acid on xylan. *Plant and Cell Physiology* 53: 1934-1949.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M. and Rhee, S.Y. (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature Biotechnology* 28: 149-156.

Lefebvre, V., Fortabat, M.-N., Ducamp, A., North, H.M., Maia-Grondard, A., Trouverie, J., et al. (2011) ESKIMO1 disruption in *Arabidopsis* alters vascular tissue and impairs water transport. *PloS one* 6: e16645.

Leipe, D.D., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) Classification and evolution of P-loop GTPases and related ATPases. *Journal of Molecular Biology* 317: 41-72.

Li, S., Lei, L., Somerville, C.R. and Gu, Y. (2012) Cellulose synthase interactive protein 1 (CSI1) links microtubules and cellulose synthase complexes. *Proceedings of the National Academy of Sciences* 109: 185-190.

Ligat, L., Lauber, E., Albenne, C., Clemente, H.S., Valot, B., Zivy, M., et al. (2011) Analysis of the xylem sap proteome of Brassica oleracea reveals a high content in secreted proteins. *Proteomics* 11: 1798-1813.

Lloyd, J. and Meinke, D. (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. *Plant Physiology* 158: 1115-1129.

Lockton, S. and Gaut, B.S. (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics* 21: 60-65.

Loomis, W. and Durst, R. (1992) Chemistry and biology of boron. *BioFactors (Oxford, England)* 3: 229-239.

Luhua, S., Ciftci-Yilmaz, S., Harper, J., Cushman, J. and Mittler, R. (2008) Enhanced tolerance to oxidative stress in transgenic *Arabidopsis* plants expressing proteins of unknown function. *Plant Physiology* 148: 280-292.

Mackay, T.F., Stone, E.A. and Ayroles, J.F. (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* 10: 565-577.

Mansfield, S.D. (2009) Solutions for dissolution—engineering cell walls for deconstruction. *Current Opinion in Biotechnology* 20: 286-294.

Mansfield, S.D., Kang, K.Y. and Chapple, C. (2012) Designed for deconstruction–poplar trees altered in cell wall lignification improve the efficacy of bioethanol production. *New Phytologist* 194: 91-101.

McCann, M.C. and Carpita, N.C. (2008) Designing the deconstruction of plant cell walls. *Current Opinion Plant Biology* 11: 314-320.

Mentzen, W.I. and Wurtele, E.S. (2008) Regulon organization of *Arabidopsis*. *BMC Plant Biology* 8: 99.

Millar, D.J., Whitelegge, J.P., Bindschedler, L.V., Rayon, C., Boudet, A.M., Rossignol, M., et al. (2009) The cell wall and secretory proteome of a tobacco cell line synthesising secondary wall. *Proteomics* 9: 2355-2372.

Minic, Z., Jamet, E., San-Clemente, H., Pelletier, S., Renou, J.P., Rihouey, C., et al. (2009) Transcriptomic analysis of *Arabidopsis* developing stems: a close-up on cell wall genes. *BMC Plant Biology* 9: 6.

Mizrachi, E., Hefer, C.A., Ranik, M., Joubert, F. and Myburg, A.A. (2010) De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 11: 681.

Mizrachi, E., Mansfield, S.D. and Myburg, A.A. (2012) Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* 194: 54-62.

Mochida, K. and Shinozaki, K. (2011) Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant and Cell Physiology* 52: 2017-2038.

Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., et al. (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell Online* 23: 895-910.

Mutwil, M., Ruprecht, C., Giorgi, F.M., Bringmann, M., Usadel, B. and Persson, S. (2009) Transcriptional wiring of cell wall-related genes in *Arabidopsis*. *Molecular Plant* 2: 1015.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O. and Persson, S. (2010) Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiology* 152: 29-43.

Nookaraju, A., Pandey, S.K., Bae, H.-J. and Joshi, C.P. (2013) Designing cell walls for improved bioenergy production. *Molecular Plant* 6: 8-10.

Obayashi, T., Nishida, K., Kasahara, K. and Kinoshita, K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant and Cell Physiology* 52: 213.

Ohashi-Ito, K., Oda, Y. and Fukuda, H. (2010) *Arabidopsis* VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *The Plant Cell Online* 22: 3461-3473.

Ohtani, M., Nishikubo, N., Xu, B., Yamaguchi, M., Mitsuda, N., Goué, N., et al. (2011) A NAC domain protein family contributing to the regulation of wood formation in poplar. *The Plant Journal* 67: 499-512.

Oikawa, A., Joshi, H.J., Rennie, E.A., Ebert, B., Manisseri, C., Heazlewood, J.L., et al. (2010) An Integrative Approach to the Identification of *Arabidopsis* and Rice Genes Involved in Xylan and Secondary Wall Development. *PloS one* 5: 263-289.

Parsons, H.T., Christiansen, K., Knierim, B., Carroll, A., Ito, J., Batth, T.S., 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.

Patel, R.V., Nahal, H.K., Breit, R. and Provart, N.J. (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *The Plant Journal* 71: 1038-1050.

Persson, S., Wei, H., Milne, J., Page, G.P. and Somerville, C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America* 102: 8633.

Pilate, G., Guiney, E., Holt, K., Petit-Conil, M., Lapierre, C., Leplé, J.-C., et al. (2002) Field and pulping performances of transgenic trees with altered lignification. *Nature Biotechnology* 20: 607-612.

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

Pont-Lezica, R., Minic, Z., Roujol, D., San Clemente, H. and Jamet, E. (2010) Plant cell wall functional genomics: Novelties from proteomics. In *Advances in Genetics*. Edited by Osborne, M. Nova Science Publishers, Hauppauge, NY.

Price, A.H. (2006) Believe it or not, QTLs are accurate! *Trends in Plant Science* 11: 213-216.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012) The Pfam protein families database. *Nucleic Acids Research* 40: D290-D301.

Quadrelli, R. and Peterson, S. (2007) The energy–climate challenge: Recent trends in CO< sub> 2</sub> emissions from fuel combustion. *Energy Policy* 35: 5938-5952.

Ranjan, P., Yin, T., Zhang, X., Kalluri, U.C., Yang, X., Jawdy, S., et al. (2010) Bioinformatics-based identification of candidate genes from QTLs associated with cell wall traits in *Populus*. *Bioenergy Research* 3: 172-182.

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

Ro, D.K., Mah, N., Ellis, B.E. and Douglas, C.J. (2001) Functional characterization and subcellular localization of poplar (*Populus trichocarpa*× *Populus deltoides*) cinnamate 4-hydroxylase. *Plant Physiology* 126: 317-329.

Rose, J.K.C. and Lee, S.J. (2010) Straying off the highway: trafficking of secreted plant proteins and complexity in the plant cell wall proteome. *Plant Physiology* 153: 433.

Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K. and Weisshaar, B. (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Molecular Biology* 53: 247-259.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O. and Ofran, Y. (2003) Automatic prediction of protein function. *Cellular and Molecular Life Sciences* 60: 2637-2650.

Rubin, E.M. (2008) Genomics of cellulosic biofuels. *Nature* 454: 841-845.

Ruprecht, C., Mutwil, M., Saxe, F., Eder, M., Nikoloski, Z. and Persson, S. (2011) Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Frontiers in plant Science* 2: 1-13.

Saito, K., Hirai, M.Y. and Yonekura-Sakakibara, K. (2008) Decoding genes with coexpression networks and metabolomics-majority report by precogs'. *Trends in Plant Science* 13: 36-43.

Schauer, K. and Stingl, K. (2010) 'Guilty by Association'–Protein-Protein Interactions (PPIs) in Bacterial Pathogens. In *Microbial Pathogenomics*. Edited by de Reuse, H. and Bereswill, S. pp. 48-61. Karger, Basel.

Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M., et al. (2004) A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *The Plant Cell Online* 16: 2278-2292.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., et al. (2002) A high-throughput *Arabidopsis* reverse genetics system. *The Plant Cell Online* 14: 2985-2994.

Showalter, A.M. (1993) Structure and function of plant cell wall proteins. *The Plant Cell* 5: 9.

Siew, N., Azaria, Y. and Fischer, D. (2004) The ORFanage: an ORFan database. *Nucleic Acids Research* 32: D281-D283.

Siew, N. and Fischer, D. (2003) Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Structure, Function, and Bioinformatics* 53: 241-251.

Solomon, S., Plattner, G.-K., Knutti, R. and Friedlingstein, P. (2009) Irreversible climate change due to carbon dioxide emissions. *Proceedings of the National Academy of Sciences* 106: 1704-1709.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., et al. (2004) Toward a systems approach to understanding plant cell walls. *Science Signalling* 306: 2206.

Sugase, K., Dyson, H.J. and Wright, P.E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* 447: 1021-1025.

Taylor, N.G. (2008) Cellulose biosynthesis and deposition in higher plants. *New Phytologist* 178: 239-252.

Tompa, P. (2002) Intrinsically unstructured proteins. *Trends in Biochemical Sciences* 27: 527-533.

Urbanowicz, B.R., Peña, M.J., Ratnaparkhe, S., Avci, U., Backe, J., Steet, H.F., et al. (2012) 4-O-methylation of glucuronic acid in *Arabidopsis* glucuronoxylan is catalyzed by a domain of unknown function family 579 protein. *Proceedings of the National Academy of Sciences* 109: 14253-14258.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. and Van de Peer, Y. (2009) Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. *Plant Physiology* 150: 535-546.

Vanholme, R., Morreel, K., Darrah, C., Oyarce, P., Grabber, J.H., Ralph, J., et al. (2012a) Metabolic engineering of novel lignin in biomass crops. *New Phytologist* 196: 978-1000.

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

Vanholme, R., Storme, V., Vanholme, B., Sundin, L., Christensen, J.H., Goeminne, G., et al. (2012b) A systems biology view of responses to lignin biosynthesis perturbations in *Arabidopsis*. *The Plant Cell Online* 24: 3506-3529.

Vázquez-Lobo, A., Roujol, D., Zuñiga-Sánchez, E., Albenne, C., Piñero, D., Buen, A., et al. (2012) The highly conserved spermatophyte cell wall DUF642 protein family: Phylogeny and first evidence of interaction with cell wall polysaccharides in vitro. *Molecular Phylogenetics and Evolution* 63: 510-520.

Vidal, M., Cusick, M.E. and Barabasi, A.L. (2011) Interactome networks and human disease. *Cell* 144: 986-998.

Vogel, J.P., Raab, T.K., Somerville, C.R. and Somerville, S.C. (2004) Mutations in PMR5 result in powdery mildew resistance and altered cell wall composition. *The Plant Journal* 40: 968-978.

Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D. and Klingler, T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Research* 9: 1198-1203.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* 38: W214-W220.

Wyman, C.E. (2007) What is (and is not) vital to advancing cellulosic ethanol. *Trends in Biotechnology* 25: 153-157.

Xin, Z., Mandaokar, A., Chen, J. and Last, R.L. (2007) *Arabidopsis* ESK1 encodes a novel regulator of freezing tolerance. *The Plant Journal* 49: 786-799.

Yamaguchi, M., Mitsuda, N., Ohtani, M., Ohme-Takagi, M., Kato, K. and Demura, T. (2011) VASCULAR-RELATED NAC-DOMAIN 7 directly regulates the expression of a broad range of genes for xylem vessel formation. *The Plant Journal* 66: 579-590.

Yang, X., Ye, C.Y., Bisaria, A., Tuskan, G.A. and Kalluri, U.C. (2011) Identification of candidate genes in *Arabidopsis* and *Populus* cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. *Plant Science* 181: 675-687.

Ye, C.Y., Li, T., Tuskan, G.A., Tschaplinski, T.J. and Yang, X. (2011) Comparative analysis of GT14/GT14-like gene family in Arabidopsis, Oryza, Populus, Sorghum and Vitis. *Plant Science* 181: 688–695.

Yong, W., Link, B., O'Malley, R., Tewari, J., Hunter, C.T., Lu, C.A., et al. (2005) Genomics of plant cell wall biogenesis. *Planta* 221: 747-751.

Yuan, Y., Teng, Q., Zhong, R. and Ye, Z.-H. (2013) The *Arabidopsis* DUF231 domain-containing protein ESK1 mediates 2-O-and 3-O-acetylation of xylosyl residues in xylan. *Plant and Cell Physiology* 54: 1186-1199.

Zhong, R., Lee, C. and Ye, Z.-H. (2010) Global analysis of direct targets of secondary wall NAC master switches in *Arabidopsis*. *Molecular Plant* 3: 1087-1103.

Zhou, Y., Li, S., Qian, Q., Zeng, D., Zhang, M., Guo, L., et al. (2009) BC10, a DUF266-containing and Golgi-located type II membrane protein, is required for cell-wall biosynthesis in rice (Oryza sativa L.). *The Plant Journal* 57: 446-462.

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

Zúñiga-Sánchez, E. and Gamboa-de Buen, A. (2012) The two DUF642 At5g11420 and At4g32460-encoded proteins interact *in vitro* with the AtPME3 catalytic domain. *Protein Interactions*: 119-142.

1.8. Figures and Tables

Table 1.1. *Arabidopsis* cell wall-related studies and the corresponding number of domains of unknown function (DUFs) and proteins of obscure features (POFs) identified.

	Reference	Type of study	Analysis type	Number of DUFS	Number of POFs
1	Bayer et al. (2006)	Proteomic	Multidimensional protein identification technology	4	0
2	Brown et al. (2005)	Transcriptome	Expression profiling & co-expression	4	0
3	Capron et al. (2013)	Genomic	QTL mapping	9	6
4	Ko et al. (2006)	Transcriptome	Expression profiling	8	4
5	Mentzen and Wurtele (2008)	Transcriptome	Co-expression	8	5
6	Minic et al. (2009)	Transcriptome/ Proteomic	Expression & proteomic profiling	20	31
7	Mutwil et al. (2009)	Transcriptome	Co-expression	15	1
8	Mutwil et al. (2010)	Transcriptome	Co-expression	7	4
9	Ohashi-Ito et al. (2010)	Transcriptome	Expression profiling	5	5
10	Oikawa et al. (2010)	Transcriptome	Co-expression	16	9
11	Persson et al. (2005)	Transcriptome	Co-expression	7	1
12	Ruprecht et al. (2011)	Transcriptome	Co-expression	20	1
13	Yamaguchi et al. (2011)	Transcriptome	Expression profiling	21	29
14	Yang et al. (2011)	Transcriptome	Co-expression	42	34
15	Zhong et al. (2010)	Transcriptome	Expression profiling	31	27

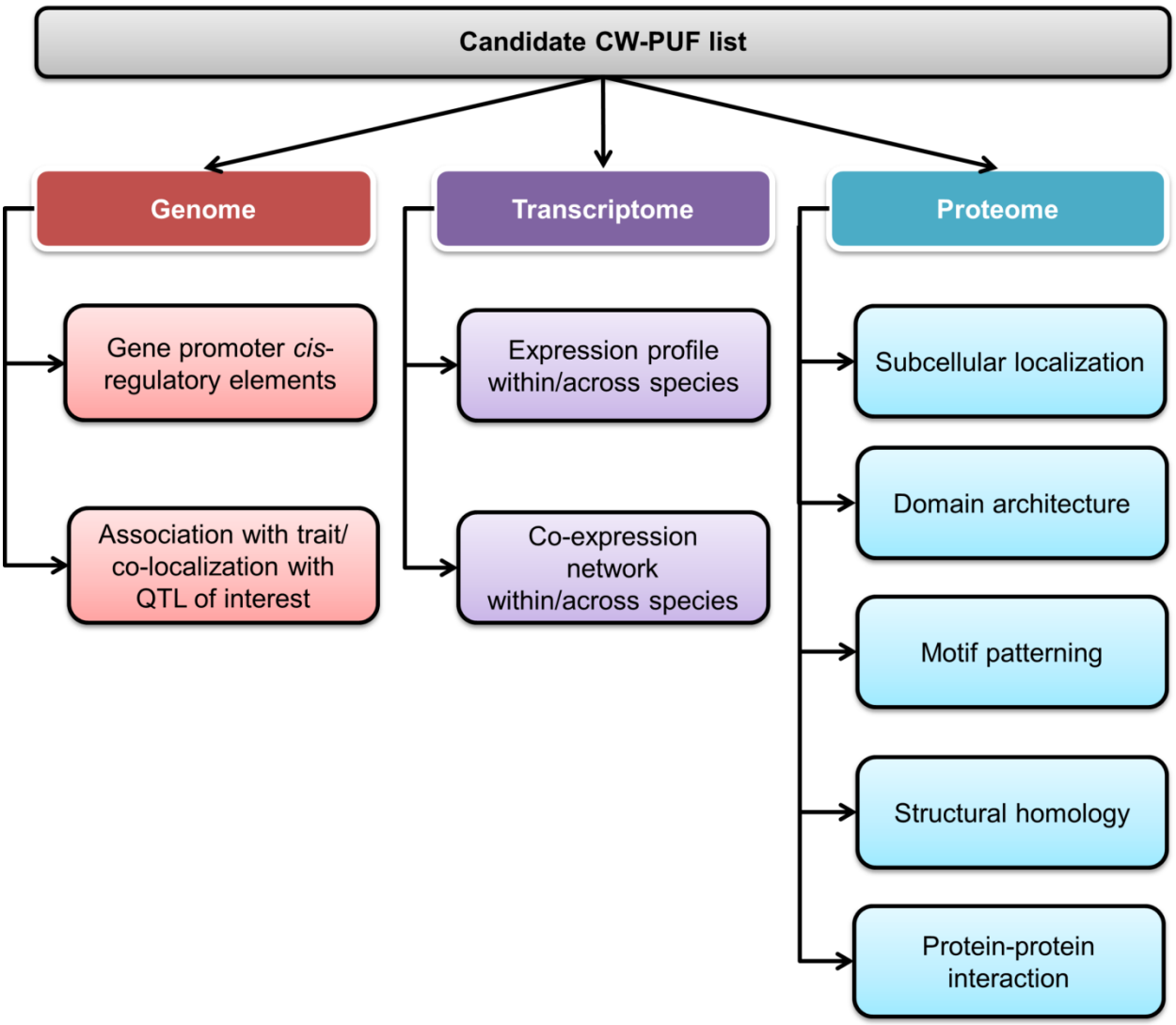


Figure 1.1. Complementary omics datatypes for functional prediction of CW-PUFs to guide experimental design.

Light colored boxes represent the types of data analysis, which can be conducted through the different *omics* levels (solid boxes) to prioritize candidate CW-PUFs.

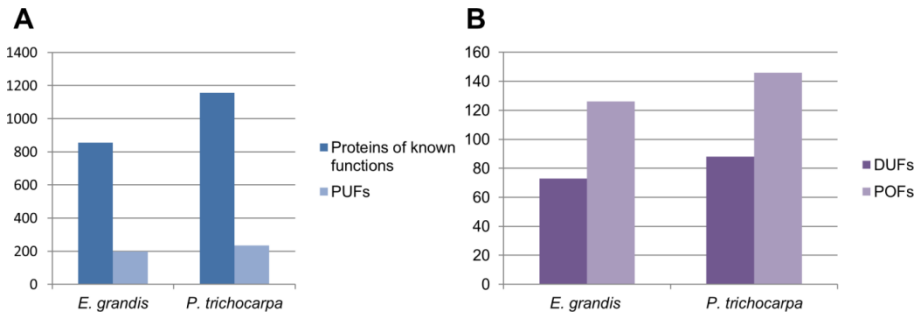
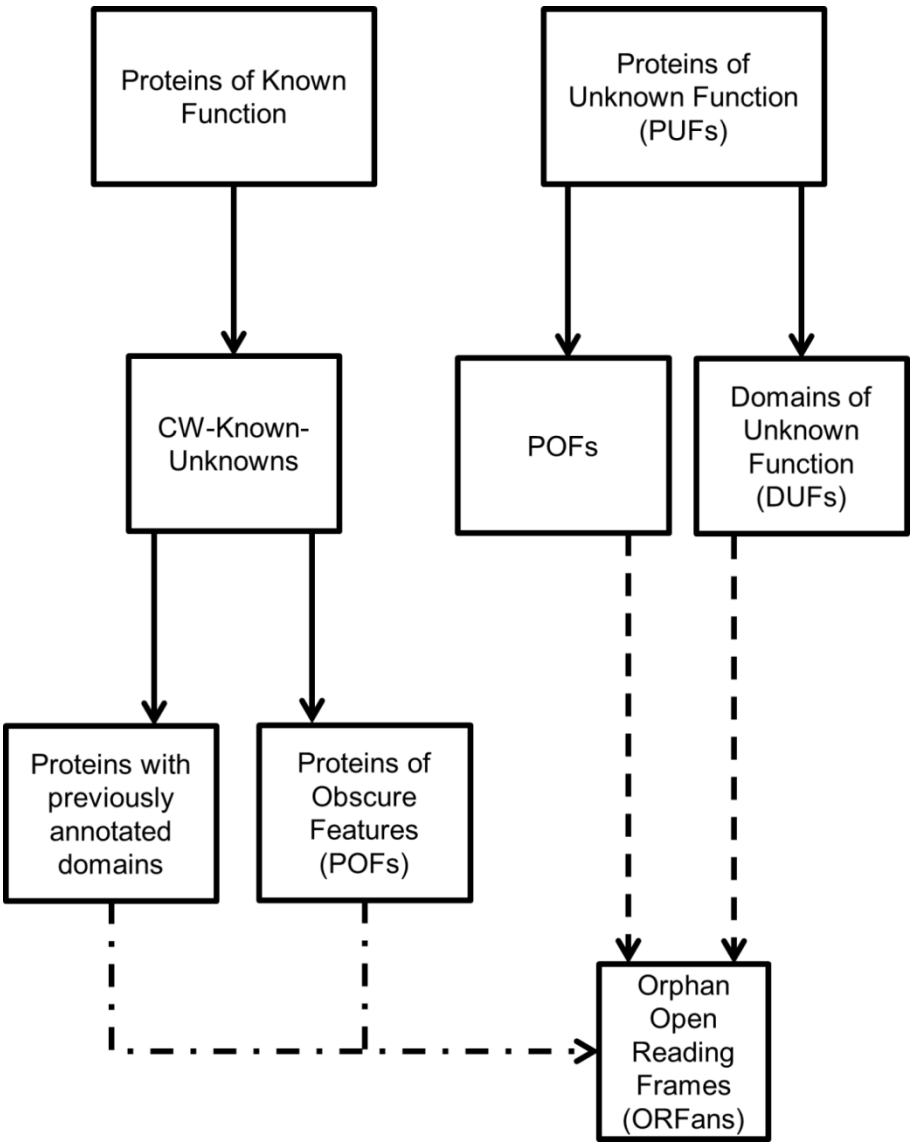


Figure 1.2. Identification of *Eucalyptus grandis* and *Populus trichocarpa* genes preferentially expressed in stem xylem tissue.

A. Xylem expressed genes were identified by ≥1.5 fold xylem-to-leaf expression (xylogenic:non-xylogenic tissue expression). A total of 1055 and 1390 significantly up-regulated genes were identified in *E. grandis* and *P. trichocarpa,* respectively (Hefer*, et al*., 2015). These genes were then classified as either proteins of known function or proteins of unknown function (PUFs). B. The PUF genes included Domains of Unknown Function (DUFs) and Proteins of Obscure Features (POFs). Most xylem preferential PUFs were POFs, while the remainder encoded DUFs.



Supplementary Figure S1.1. Categories of cell wall-related proteins of unknown function (CW-PUFs).

Different types of CW-PUF were identified from candidate cell wall-related gene list in literature (and in Table 1.1). Protein annotation via the similarity annotation approach resulted in either i) proteins of known function (significant homology to functionally annotated proteins present within a reference database), or ii) proteins of unknown function (PUFs) (functionally uncharacterized). Within proteins of known function are proteins referred to as CW-known unknown, for which there is a generalized prediction of biochemical or cellular function despite the lack of empirical evidence supporting cell wall function. CW-known unknowns may include proteins with previously annotated domains or proteins

lacking currently defined domains or motifs (referred to as proteins of obscure features (POFs). PUFs may be further subdivided into proteins which contain recognizable domains of unknown function (DUFs) or POFs. Finally, proteins from the above categories may have a limited phyletic distribution, specific to a species or group of species and is referred to orphan open reading frames (ORFans).