COMBINED PROFILING IN ASPEN

A SYSTEMS BIOLOGY APPROACH

OGONNA OBUDULU
Faculty of Forestry
Department of Forest Genetics & Plant Physiology
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2015

Acta Universitatis agriculturae Sueciae
2015:91

Cover: field-grown and transgenic aspen trees systems biology approach
(Photo: Obudulu O., Bourquin V. and Hvidsten TR)

ISSN 1652-6880
ISBN (print version) 978-91-576-8380-9
ISBN (electronic version) 978-91-576-8381-6

Print: Arkitektkopia, Umeå 2015

Combined profiling in aspen: a systems biology approach

Abstract

This thesis presents efficient systems biology modelling strategies for integrating complex multi-platform datasets in order to increase our understanding of stress biology, wood formation and secondary cell wall formation in *Populus* trees and connected responses to perturbations in organisms, particularly aspen. It is based on studies reported in four appended papers, designated Papers I-IV.

Paper I describes an enhanced framework for investigating and understanding multi-level oxidative stress responses and their influence on phenotypic variation in transgenic hipI-superoxide dismutase *Populus* plants. Paper II presents a modelling strategy based on a combination of Principal Component Analysis (PCA), Orthogonal Projections to Latent Structures (OPLS), and an enhanced stepwise linear modelling approach. This strategy revealed major transitions in the proteomes along the wood developmental series in *Populus tremula* (aspen) pinpointing, for example, the location of the cambial cell divisions leading to phloem and xylem cells, and the location of the secondary cell wall formation zone.

A pairwise multivariate OPLS approach was applied in the study described in Paper III to analyse proteome dynamics during tension wood formation of *Populus* trees. Pairwise comparisons of four corresponding phases/tissue types in normal and tension wood formation allowed identification of several processes that were strongly enhanced and/or unique during tension wood formation.

These multidisciplinary approaches together with a recently developed formulation of the OnPLS method was used in Paper IV to analyse a set of transgenic *Populus* trees carrying an RNAi construct for the *Populus* secretory carrier-associated membrane protein3 (PttSCAMP3) gene. Multilevel analysis of datasets from nine platforms (RNA sequencing, proteomic, metabolomic and wood chemical analyses) revealed a critical function for PttSCAMP proteins in wood formation and elucidated the underlying molecular mechanism. The systems biology approach presented in this thesis provides novel types of tools for elucidating stress responses and wood formation in *Populus* trees. Exploitation of very recent advances in multivariate methods such as OnPLS allowed the simultaneous combination of transcriptomic, proteomic and metabolomic data, as well as identification of differences and connections between the data sets, which would not have been possible using standard statistical methods.

Keywords: Systems biology, transcriptomics, proteomics, metabolomics, *Populus*, multivariate statistics, wood development, aspen

Author's address: Ogonna Obudulu, SLU, Department of Forest Genetics and Plant Physiology, P.O. Box 901 83 Umeå, Sweden. *E-mail:* ogonna.obudulu@slu.se

Dedication

To my strength and song, and all who have helped me in life to become who I am today.

Contents

List of Publications

This thesis is based on the work contained in the following papers, which are referred to by the corresponding Roman numerals in the text:

I Srivastava V*, **Obudulu O***, Bygdell J, Löfstedt T, Rydén P, Nilsson R, Ahnlund M, Johansson A, Jonsson P, Freyhult E, Qvarnström J, Karlsson J, Melzer M, Moritz T, Trygg J, Hvidsten TR, Wingsle G. (2013). OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI-superoxide dismutase *Populus* plants. *BMC Genomics*, 14:893.

II **Obudulu O**, Bygdell J, Sundberg B, Moritz T, Hvidsten TR, Trygg J, Wingsle G. Quantitative proteomics reveals protein profiles underlying major transitions in aspen wood development (submitted).

III Bygdell J, **Obudulu O**, Nilsson R, Srivastava MK, Srivastava V, Sundberg B, Trygg J, Wingsle G. Protein expression in tension wood formation in *Populus*, monitored at high tissue resolution (manuscript).

IV **Obudulu O**, Mähler N, Skotare T, Bygdell J, Abreu IN, Ahnlund M, Gandla ML, Petterle A, Gerber L, Moritz T, Hvidsten TR, Jönsson LJ, Wingsle G, Trygg J, Tuominen H. A systems approach reveals functions of secretory carrier-associated membrane proteins in wood formation of *Populus* trees (manuscript).

* Joint first authors and equal contributors

The contributions of Ogonna Obudulu to the papers included in this thesis were as follows:

I Data analysis and interpretation after OnPLS application, data mapping to KEGG and Mapman pathways and writing the manuscript.

II Preparation of the samples after sectioning, conducting the proteomic experiment, stepwise OPLS application and writing the manuscript.

III Pairwise OPLS application, data mapping to KEGG and Mapman pathways and writing the statistical parts of the manuscript.

IV Preparation of the samples, conducting the proteomics and metabolomics analysis, participation in data interpretation after OnPLS application, data mapping to KEGG and Mapman pathways and writing the manuscript.

Abbreviations

All abbreviations are explained as they first appear in the text**.**

1 Introduction

Sumner *et al.* (2003) describe systems biology as encompassing "…a holistic approach to the study of biology and the objective is to simultaneously monitor all biological processes operating as an integrated system". This view of systems biology is based on the central dogma of molecular biology, which assumes a sequential unidirectional transfer of information from DNA by transcription to RNA and then to proteins via translation (Crick, 1970). The omics cascade can be regarded as a more modern version of the central dogma, and has been described as a flow from the genome (the starting point, representing all processes that *can* happen) to the transcriptome (representing processes that *appear* to take place), the proteome (processes that influence events) and finally the metabolome (visible completed and ongoing effects) (Dettmer *et al.*, 2007). While analysis at any one of these levels can provide valuable information, multi-level analysis enables more reliable mapping of different molecular states by combining information obtained from multiple types of experiments and instruments to provide detailed insights into plants responses to perturbation (Weckwerth *et al.*, 2004; Kaever *et al.*, 2014). This thesis deals with such multi-level analyses.

Multi-level omics measurements can provide a broader view of biological processes than can be obtained by focusing on a single level to the exclusion of others, not least because information from different levels is often complementary and can reveal things that would otherwise be missed. For example, it is well known that the transcription level of a given gene does not necessarily correlate well with the levels of the corresponding protein (because of processes such as posttranslational modification, protein inactivation, and protein–protein interactions) or with the levels of any associated metabolites (Gygi *et al.*, 1999; Diz *et al.*, 2012). Figure 1 shows a schematic flowchart of the systems biology approach applied in the studies underlying this thesis, combining transcriptomic, proteomic and metabolomic profiling in order to investigate effects of oxidative stress in *Populus* (Srivastava *et al.*, 2013). The studies involved up to nine sources of data: transcriptomic, proteomic, three types of metabolomic (gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), pyrolysis–gas chromatography/mass spectrometry (Py-GC-MS), phenotypic-growth-height-density, saccharification, hemicellulose content and glucan analyses. All nine were used in the study described in Paper IV, and are briefly described in following sections.

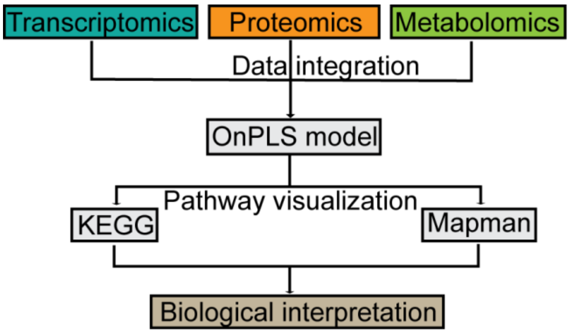

Figure 1. Schematic flowchart of the integrated profiling strategy employed in the studies underlying this thesis. In the first step, transcriptomic, proteomic and metabolomic data for the cambial region of *Populus* wild type and transgenic plants were collected individually. In the second step, these three omic datasets were integrated using the most recently developed formulation of the Orthogonal Projections to Latent Structures (OPLS) approach, which is known as OnPLS (an extension of O2PLS described in Chapter 3) (Lofstedt *et al*., 2013) to identify their joint variation - initially by applying OnPLS modelling to all variables and subsequently to targeted variables. Finally, the OnPLS model constructed in the second step was visualized with Mapman and KEGG to explore the pathways (genes-proteins-metabolites) affected in the transgenic plants and deepen interpretation of their oxidative stress responses. (From Figure 1: Srivastava *et al*., 2013).

When adequate funding is available and there are no time constraints, integrated multi-layered omics approaches are optimal for characterizing complex biological changes. Approaches of this kind have successfully revealed significant functional information and helped to generate new hypotheses concerning cellular-phenotypic variation that can provide robust foundations for future studies (Ideker *et al.*, 2001; Zhu *et al.*, 2012; Wu *et al.*, 2014).

Wood development is highly interesting for both researchers probing fundamental biological processes and applied scientists seeking to optimise and exploit cellulose and lignin production (Plomion *et al*., 2001). Previous efforts to elucidate the molecular processes involved in wood formation have been based on transcriptomic analyses and/or low-resolution sampling (Hertzberg *et al*., 2001; Andersson-Gunnerås *et al*., 2006). However, while transcriptomic data provide strong indications of protein abundance, quantitative proteomic information is needed to confirm the presence and relative abundance of different proteins and to account for the effects of post-translational modifications. Similarly, metabolomic analysis provides information about observed phenotypic variation (Fernie & Schauer, 2009; Sweetlove *et al*., 2014). My colleagues and I (hereafter we), have therefore studied dynamic changes in aspen under various conditions and used the resulting data to create statistical models of its transcriptomic, proteomic and metabolomic responses.

Paper I focuses on the complex oxidative stress tolerance mechanisms of plants. A multivariate modelling approach was used to statistically integrate information concerning global (transcriptomic, proteomic and metabolomic) responses to oxidative stress in the cambium region of a *Populus* model system. Samples were collected from the cambial region of wild-type controls and mutant poplar plants expressing antisense RNA of the gene encoding the antioxidative copper-zinc superoxide dismutase (SOD) enzyme hipI-SOD. The mutant and wild-type plants were studied using transcriptomic, proteomic, and metabolomic tools described in Chapter 3. The resulting datasets were then statistically integrated using OnPLS (Lofstedt *et al*., 2013). OnPLS identifies globally joint information in any number of data blocks while being fully symmetrical, i.e. giving no preference to any one data block (Lofstedt *et al.*, 2013; Srivastava *et al*., 2013).

In the study reported in Paper II, a proteomics technique (Ultra-Performance Liquid Chromatography-Quadrupole Time-Of-Flight Mass Spectrometry (UPLC-QTOF-MS), see section 3.3) was used to quantify protein expression in tangential 20-160 µm thick sections spanning all stages of wood development in *Populus tremula* from phloem through cambium, the expansion zone, xylem, and dead cells. The resulting high-resolution developmental data series represents 482 sections from four 47-year-old trees harvested in the forest. A combination of PCA, OPLS modelling and an enhanced novel stepwise linear modelling approach revealed major transitions in global protein expression, pinpointing (among other things) the location of the cambium division leading to phloem and xylem cells, and the location of secondary cell wall formation.

Paper III presents a high tissue resolution study of protein expression in tension wood formation. Global protein expression was analysed in cell types originating from phloem, cambium and xylem in normal and tension wood of poplar, representing a developmental gradient covering multiple distinct developmental stages. This enabled clarification of the cellular mechanisms involved in both strictly developmental processes (e.g. xylogenesis) and tension wood formation.

Paper IV describes the use of a systems approach to unravel functions of secretory carrier-associated membrane proteins (SCAMPs) in *Populus* wood formation. The function of the SCAMP3 gene in *Populus* trees was investigated using three transgenic p35S RNAi lines of Potri.019G104000 (SCAMP3). The living part of the xylem was scraped and collected from the stems of two-month-old trees (seven wild-type trees and four or five trees of each transgenic line) then analysed.

The results presented in Papers I and IV demonstrate that the integrated omics modelling strategy applied is an immensely powerful tool for elucidating multi-level responses to environmental changes in plants (exemplified by the oxidative stress response), analysing biological variability associated with mutation, and functions of specific genes (exemplified by the SCAMP3 gene). Papers II and III show that it is possible to integrate several multivariate statistical tools to obtain a clear understanding of transformative biological processes. The stepwise and pairwise modelling methods used in these works identified major transitions in global protein expression and pinpointed precise tissue-specific variation, suggesting the occurrence of novel and unexplored biological processes that will be the focus of future research on wood development.

1.1 Wood development in *Populus tremula* (aspen)

1.1.1 Why study trees?

Forest trees have great economic importance as sources of timber, pulp, biomass, and medicinally relevant compounds. They are also environmentally important as major elements of habitats, as sources of food, oxygen, and energy, and as providers of vital ecosystem services. Research aimed at understanding and improving their growth and quality is therefore highly important (Varshney *et al.*, 2014). Wood from forest trees is enormously interesting for both purely scientific and practical reasons due to its global abundance and diverse domestic and industrial applications, which are largely dependent on its contents of renewable biomass in the form of cellulose and lignin (Van Acker *et al.,* 2013).

All traits of trees (e.g. leaf characteristics, growth rate, phenology, dimensions, form, seed germination parameters, wood properties and stress resistance) are determined by interactions between genetic and environmental factors. Thus, traditional tree breeding approaches are currently used in combination with genetic analyses and genetic engineering techniques to study and modify these traits in order to produce faster-growing trees with desirable qualities (Plomion *et al*., 2001; Mellerowicz & Sundberg, 2008; Van Acker *et al.,* 2013). This thesis focuses on genetic modifications of aspen trees and the use of multivariate statistical methods to identify factors that could be tuned to improve wood properties.

Populus species, including poplar and aspen, are among the most important angiosperm model systems for studying wood formation, which occurs via a process known as xylogenesis (Plomion *et al*., 2001; Tuskan *et al*., 2006; Mellerowicz & Sundberg, 2008). Xylogenesis can be described as an ordered developmental process involving cell division, cell expansion, secondary wall deposition and lignification, and programmed cell death (Mellerowicz & Sundberg, 2008). Aspens grow rapidly, can be regenerated from their sprouts, and have a relatively close phylogenetic relationship to the extremely widely used plant model system *Arabidopsis thaliana*. All of these factors have helped to increase its popularity in forestry and plant research (Jansson & Douglas, 2007). Among woody plants, *Populus* can be considered as the ‘model’ tree for genomic research, partly because its genome is relatively small at 450–550 Mbp (Taylor, 2002; Tuskan *et al*., 2006; Goodstein *et al*., 2012; Nordberg *et al*., 2014). In addition, the *Populus trichocarpa* genome has been completely sequenced, providing a valuable bioinformatics dataset to support omics investigations (Tuskan *et al*., 2006; Goodstein *et al*., 2012; Nordberg *et al*., 2014).

Papers I and IV appended to this thesis focus on hybrid aspen (*Populus tremula x Populus tremuloides*) while Papers II and III focus on field-grown aspen (*Populus tremula*). The choice of model system in each case was dictated by the plant processes under investigation.

1.1.2 Biochemical composition and physiology of wood

Wood from trees is one of the most abundant renewable natural materials on earth, and wood development is highly interesting to both scientists exploring fundamental processes and commercial groups seeking to improve supplies of important products including timber, biofuel, pulp and paper (Mellerowicz & Sundberg, 2008; Niculaes *et al.,* 2014). Tree stems can be divided into two parts: the external protective layer known as the bark, and the inner woody layer. As trees grow the width of the woody layers increases resulting in increased stem size. Three major types of polymers (cellulose, hemicellulose and lignin) control the chemical and physical properties of the cells and stem during wood development, and wood yields can be manipulated by altering their production and interactions (Plomion *et al*., 2001; Mellerowicz &

Gorshkova, 2012). The wood development process involves five major phases (cell division, cell differentiation, cell expansion, cell wall formation and cell death), which progressively occur in poplar stems from the cambium inwards through the xylem expansion zone to mature xylem, and from the cambium outwards to mature phloem (Plomion *et al*., 2001; Mellerowicz & Sundberg, 2008; Van Acker *et al.*, 2013). The zones are illustrated in Figure 2 (adapted from Paper II) and described in detail by Hertzberg *et al*. (2001) and Mahboubi *et al.* (2013).

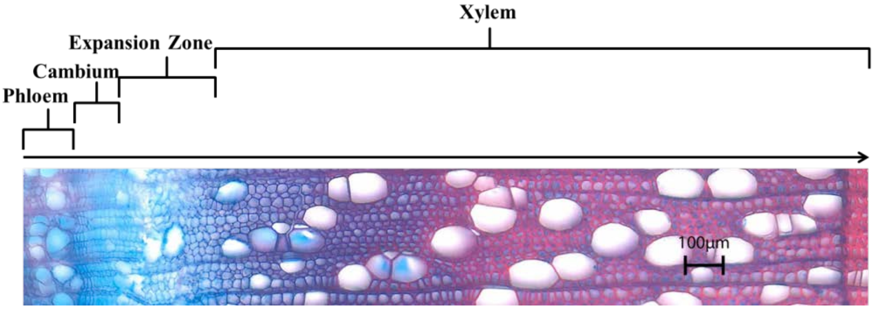

Figure 2. Overview of transverse sections prepared from a poplar tree specimen, showing the four wood developmental zones (phloem, cambium, expansion zone and xylem) in accordance with illustrations in Herzberg *et al.* (2001) and Mahboubi *et al.* (2013) (Paper II).

The phloem transports various substances, including phytohormones and photosynthetic products, from "source tissues" to "sink tissues", while the cambium is the site of active cell division. The xylem provides mechanical support, transports water, mineral ions and various other solutes from the roots to other parts of plants (Plomion *et al*., 2001; Mellerowicz & Gorshkova, 2012; Myburg *et al.*, 2013). It consists of different types of cells such as tracheids, vessels and fibres which deposit cell wall. Initially, the primary cell wall is composed of polysaccharides (mainly cellulose, hemicelluloses and pectins) and glycoproteins that provide rigidity, but also sufficient flexibility (mediated by various enzymes) to allow cellular expansion (Plomion *et al*., 2001; Mellerowicz & Gorshkova, 2012; Myburg *et al.*, 2013; Van Acker *et al.*, 2013). After cellular expansion ceases, a secondary cell wall (which includes polyphenols in addition to cellulose and hemicelluloses) is deposited. Secondary walls usually contain three recognized layers, designated S1, S2 and S3 (Figure 3) (Plomion *et al*., 2001; Mellerowicz & Gorshkova, 2012; Van Acker *et al.*, 2013). For further details of plant cell wall structures and formation processes, which are highly complex, see Plomion *et al.* (2001) and Mellerowicz & Gorshkova (2012).

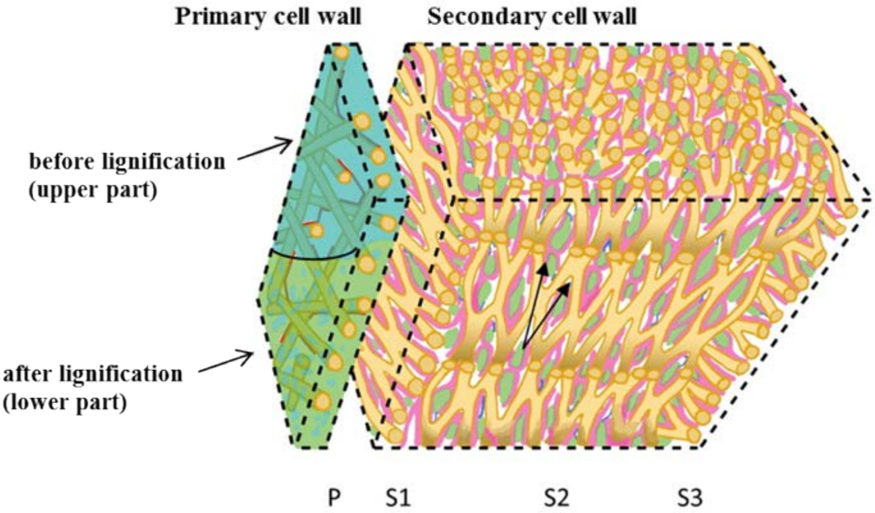
Primary cell wall
Secondary cell wall
before lignification (upper part)
after lignification (lower part)
P
S1
S2
S3

Figure 3. Illustration of plant cell wall organisation showing the primary (P) and secondary (S1, S2 and S3) layers (adapted from Mellerowicz & Gorshkova, 2012). The long arrows indicate angles relative to the cell axis microfibril angles (MFAs) in the S2 layer. Xyloglucan chains are shown in red, pectins in blue, xylan in pink, cellulose fibrils in beige, the cross-linked lignin network in green, and mannan chains in dark blue.

Due to the complexity of poplar cell walls costly engineering procedures, pre-treatment methods and genetic improvement strategies capable of modifying its lignin composition/amount, cellulose crystallinity, and hemicellulose amount/composition are needed to manipulate its saccharification efficiency and biomass production (Van *et al*., 2013; Porth *et al*., 2013). Some strategies of this sort are considered in Papers II, III and IV.

1.2 Omics applications in plant systems biology

1.2.1 Data evaluation strategies

According to Svante Wold “Statistics is the art of drawing conclusions from data, and making decisions, in the presence of variability” while “Chemometrics is the branch of chemistry concerned with the analysis of chemical data (extracting information from data) and ensuring that experimental data contain maximum information (the design of experiments)." Both are important in plant sciences, where variation must be interpreted efficiently (Wold, 1995). The selection of particular methods for data analysis, evaluation and statistical testing in order to draw biological inferences should be based on the specific experimental goals and desired outcomes (Wang *et al*., 2006). Several data evaluation strategies are currently used in plant sciences, ranging from methods for assessing the shape and distribution of data in order to obtain broad overviews of datasets' properties, to strategies centred on scaling and normalisation to examine variation within them.

Univariate methods are most commonly used with singular datasets or measurements and involve examining one variable at a time. Their output can be presented as a frequency distribution of sample categories in a tabular format, a chart, or some other graphical representation that emphasises measures of central tendency such as the sample mean, median and mode (Babbie, 2009). Univariate tests of statistical significance to detect differences within and between sampled populations are applied using published, recommended, or arbitrary cutoff values based on mean differences or statistics obtained using various methods that may include Student's T-test, F tests, fold-change analyses or analysis of variance (ANOVA).

Multivariate statistical analyses are often applied to larger datasets in order to detect trends in relationships, characteristics, effects and properties, to facilitate more descriptive and integrated, holistic biological interpretations, especially in cases involving multiple datasets. Several projection methods have been introduced for visualising results of such analyses. One of the most prominent and widely used is PCA, in which the first projection (Principal Component 1, denoted PC1) effectively describes the greatest variance within the data, while other PCs describe progressively smaller amounts of the remaining variation (Madala *et al*., 2014). In multivariate analyses, variables that differ significantly are often identified by considering loading values, variable importance (VIP) values, correlation coefficients, and so on.

Multivariate methods are useful for identifying correlations between individual measurements, highlighting global relationships, multifactor analysis and examining residuals (or non-interacting/undesired effects) (Rood *et al*., 2013).

The comparisons of transgenic lines versus their group controls presented in this thesis were mainly performed using multivariate models (PCA, OPLS and OnPLS) and in some cases checked by univariate statistics.

1.2.2 Visualisation tools

Bioinformatics and data visualization tools are invaluable for understanding and interpreting biological data as they greatly facilitate efforts to investigate large data blocks, summarise information they contain, and display systematic variations or perturbations in biological systems (Sumner *et al*., 2003). Visualisation tools can be broadly divided into a group for visualizing variation and another for interpreting responses. They are distributed and accessed in various ways, including via standalone downloads of free or commercial software packages, web interfaces (e.g. paintomics, or cytoskape), or databases hosted by research institutes from which one can download data files such as GMD@CSB.DB, the Golm metabolome database (Kopka *et al*., 2005). Transcriptomic microarray data have been visualised using GENEVESTIGATOR, and other analytical tools (Zimmermann *et al*., 2004).

Due to this plethora of pathway and visualisation tools users may have to conduct a tool-by-tool study in order to determine which is most suitable to meet their research objectives (Huang *et al*., 2009). To illustrate overall states associated with particular treatments or changes, and support meaningful interpretation of intrinsically complex omics datasets, networks derived from coexpression analysis are often used in connection with pathway analysis (Porth *et al*., 2013). Visualization methods based on such networks can display relationships between variables in clusters or graphs, and capture raw similarities between various omics datasets using hierarchical clustering, K-means clustering or correlation networks (Subramanian *et al*., 2005).

Co-expression network analysis is used to determine which genes participate in a biological process or have similar functions to chosen target genes. This approach examines topological features such as connections (nodes), closeness, or interactions (edges) between variables, and generates graphs based on “guilt by association” evaluated using some predefined scoring or enrichment function (Oliver, 2000). These graphs do not necessarily show the state of the focal system or its direction of change, but they are useful for revealing similarity or coregulation (Saito *et al.*, 2008), and provide very useful information when investigating molecular relationships and associations (Higashi & Saito, 2013). However, network analyses based on association or interaction alone may be inappropriate in some cases because their results can be unduly influenced by a small number of outlying interactions that cannot be reliably used to make generalizations or draw conclusions about the rest of the network. Consequently, this approach can generate false positives and negatives (Draghici *et al.*, 2007; Gillis & Pavlidis, 2012). It is therefore important to combine information from multiple platforms, using computational approaches such as OnPLS, which can efficiently handle outliers and thus enable robust data interpretation (Lofstedt *et al.*, 2013; Srivastava *et al*., 2013). We have used OnPLS to identify interesting variables and visualize genes' expression based on known ontological and functional data or established biological pathways.

2 Objectives

The main objectives of the studies this thesis is based upon were:

1. To enhance current understanding of wood biosynthesis by applying a systems biology approach to multilevel datasets concerning the process in field-grown and genetically modified aspen.
2. To characterize changes in expression of several genes, proteins and metabolites from different tissue types and developmental zones in aspen stems with significant impacts on, and implications for, wood formation.
3. To develop a systematic approach (based on transcriptomic, proteomic and metabolomic profiles of field-grown and transgenic aspen trees) for discovering functional connections.
4. To create reference collections of genetic data including detailed stage-specific expression profiles that can be used to explore and elucidate biological processes involved in wood formation.
5. To highlight a series of biological processes and pathways essential for wood development in hybrid aspen.
6. To provide resources and tools to facilitate further investigation of these biological processes and pathways.

In order to meet these objectives, we initially focused on evaluating and refining methods and tools for quantifying molecular events in wood samples, analysing the acquired data, and further characterization, pathway analysis and biological interpretation of detected patterns. Established multivariate methods were then used to study previously investigated and novel genes, using both wild-type plants and transgenics generated by SweTree Technologies, Umeå, Sweden. Finally we observed a series of biological processes essential for wood development in aspen.

3 Materials and methods

3.1 Plant materials and sampling.

The plant materials used in the work underlying the thesis were obtained from field-grown and transgenic aspen trees.

The samples described and discussed in Paper I were scraped from the cambial regions of stems of transgenic hybrid aspen (*Populus tremula x P. tremuloides*) plants expressing high-isoelectric-point superoxide dismutase (hipI-SOD) transcripts in antisense orientation, containing elevated superoxide levels, and wild-type controls. Three 12-week-old wild-type plants and three of each of two antisense lines (designated AS-SOD9 and AS-SOD24) were sampled.

The samples used in the study reported in Paper II were tangential sections with thicknesses of 20-160 µm spanning all stages of wood development from phloem, through cambium, cell expansion and xylem to cell death zones. In total, 482 sections were taken from the stems of four 47 year-old *Populus tremula* trees harvested in a Swedish forest.

In the study described in Paper III, field-growing aspen trees (*P. tremula*) were bound with string and held at a 45° angle relative to the ground in order to induce tension wood formation. Two biological replicates of control and induced trees were sampled after four weeks and tangential phloem, cambium and xylem tissues were obtained from the stem.

Paper IV presents an investigation of the SCAMP3 gene’s function in *P. tremula x P. tremuloides* trees using three transgenic p35S Potri.019G104000 (SCAMP3)-RNAi lines. The living part of the xylem was scraped and collected from two-month-old stems of seven wild-type trees and four to five trees representing each of the transgenic lines.

The samples used in the studies described in Papers I and IV were obtained by scraping with scalpels, after which they were immediately flash-frozen in liquid nitrogen as described by Celedon *et al.* (2007). All samples were then ground into a fine powder in a mortar cooled with liquid nitrogen and the resulting powder was analysed immediately or stored at -80°C until needed for analysis.

The sections used in studies reported in Papers II and III were cut by high-resolution tangential cryosectioning. Sections 20 µm thick (20 µm × 2 mm × 20 mm, ≈0.5 mg, fresh weight) were cut from each tree's stem, through the wood formation zones (extending from the phloem through the cambium to the mature xylem), following previously described procedures (Uggla *et al.*, 1996 and 1998). Some of the samples were pooled during analysis as described in Paper II with final sections ranging from 20-160 µm thick.

3.2 Transcriptomic Microarray Analysis and RNA sequencing

A commonly used tool for studying gene expression is microarray analysis, where the expression profile of various genes is estimated from the amount of hybridized mRNA bound to each site on an array. This is done by first extracting transcripts from the plant tissues under investigation, labelling the transcripts with fluorescent dyes such as cyanine dye Cy3/Cy5, hybridizing them to complementary DNA probes immobilised at predefined positions on an array, and washing. The fluorescent dyes in the stained transcripts are then stimulated with a laser, inducing emission of light, the intensity of which is measured and used as an indicator of the expression of corresponding genes. This enables the simultaneous global quantification of many mRNA transcripts, linking the information encoded in the transcripts to phenotypic variation (Malone & Oliver, 2011).

Transcriptome profiles can be acquired by microarray analysis or RNA deep sequencing. These two techniques sometimes produce similar or complementary results (Bloom *et al.*, 2009). The first methods for transcriptome analysis were based on microarrays, which were used to generate the transcriptomic data considered in Paper I. However, methods based on Next Generation Sequencing (NGS) of DNA (DNA-seq) or RNA (RNA-seq) are increasingly becoming the tools of choice because they offer broader gene coverage, higher resolution, superior sensitivity, better reproducibility from biological replicates, higher specificity (enabling allele-specific expression analysis), and abilities to discover novel transcripts and identify isoforms (Malone & Oliver, 2011). RNA-seq also avoids problems associated with probe redundancy and annotation, which can occur when using pre-designed microarray sequence detection probes. However, it should be noted that microarrays are still widely used because of user familiarity and the lower complexity of handling, storing, and analysing the data they provide (Zhao *et* *al*., 2014). Because of the many advantages of NGS methodology, the transcriptomic data considered in Paper IV were acquired by RNA-seq.

RNA-seq involves converting isolated transcripts into complementary DNA (cDNA), which is then directly sequenced using next-generation sequencing technologies (Wang *et al*., 2009; Kogenaru *et al*., 2012). A common RNA-seq protocol begins by generating a double-stranded cDNA library using random or oligo(dT) primers exhibiting bias towards the 5 and 3 ends of genes, which is used for mapping the ends of genes and identifying transcribed regions. The cDNAs, which are usually made from poly(A)+ RNA, are fragmented by DNase I and ligated to adapters. The resulting adapter-ligated cDNA fragments are then amplified and sequenced in a high-throughput manner to obtain very large numbers of short sequence reads (Nagalakshmi *et al.*, 2010). Whole genome transcriptome maps are constructed by aligning and mapping the reads obtained to a reference genome, or if no reference genome is available, by *de novo* assembly of short reads (Wang *et al*., 2009: Nookaew *et al.*, 2012). This mapping of the resulting reads onto a reference genome enables quantification of relative or absolute gene expression levels (Wang *et al*., 2009; Kogenaru *et al.*, 2012).

The transcriptomic data considered in Paper I were generated by microarray analysis and processed using standard in-house procedures developed at Umeå Plant Science Centre (UPSC) as described by Srivastava *et al*. (2013). The RNA-seq analysis presented in Paper IV was conducted using the Illumina platform. The analysed samples were sent to the Beijing Genome Initiative (BGI), China (Zhao *et al*., 2004), and raw reads received were mapped/assembled using standard in-house procedures.

3.3 Proteomic mass spectrometry

Identification and quantification of cells', tissues' or organisms' proteins provides a window into complex regulatory networks. Gel-based methods involving (for example) separating proteins in two dimensions, e.g. size and charge, were traditionally used to fractionate complex protein mixtures, but research groups are increasingly using mass spectrometry for this purpose due to its far greater ability to separate and characterize complex mixtures (Kelleher, 2004). In mass spectrometric analysis, proteins are initially purified and then cleaved into peptides using a sequence-specific protease (in our case trypsin). The peptides are then analysed by mass spectrometry (MS), during which they are broken up into ionized fragments whose molecular masses are measured by following their specific trajectories in a vacuum (Steen & Mann, 2004). Peptides are analysed because they are easier to handle than proteins. They have fewer solubility issues and are therefore less likely to create problems during MS analysis arising from the presence of detergents, they are less prone to modification during analysis, they are more readily ionised than the corresponding proteins, they are typically more amenable to MS detection (resulting in better MS sensitivity) and they provide more robust and comprehensive sequence information than can be obtained from analysis of whole proteins (Steen & Mann, 2004).

UPLC-QTOF-MS rapidly provides information on the nature of specific protein/peptide components in complex plant mixtures (Link *et al.*, 1999). Two commonly used ionization techniques for proteins are matrix-assisted laser desorption ionization (MALDI), which is suitable for high molecular weight proteins, and electrospray ionization (ESI), which is suitable for small quantities of material because it offers greater sensitivity, enabling detection of analytes at femtomole levels (Dutt & Lee, 2000). ESI was used in the UPLC-QTOF-MS analysis presented in this thesis.

Proteomic analysis steps such as digestion, purification, separation and data analysis affect the output from protein profiling (Wang *et al*., 2014). Thus, ways to improve all of these steps are being investigated. Major current objectives are to develop higher-resolution or more convenient separation techniques and more sensitive quantification protocols to facilitate analysis of post-translational modifications (Mann & Jensen, 2003). There is a particular need to develop methods for integrating information about levels of mRNA expression and the abundance of corresponding proteins in tissue samples in order to better understand the nature and functions of gene networks, and to explore their coordination (Dutt & Lee, 2000).

The MS analyses reported in Papers II, III and IV involved label-free quantitation using the phase-transfer surfactant sodium deoxycholate following protocols presented by Masuda *et al*. (2008), with modifications. Plant tissue samples were digested with trypsin and the resulting complex peptide mixtures were separated using a nanoACQUITY^{TM} UPLC system (Waters, Massachusetts), after which the separated peptides were analysed by QTOF-MS using a Waters Synapt^{TM} G2 HDMS mass spectrometer equipped with a nanoflow electrospray ionization (ESI) interface operating in positive ionization mode with a minimal resolution of 20,000. Data were processed using the Protein Lynx Global Server v.3.0 (Waters) and the resulting spectra were searched against *Populus trichocarpa* v3.0 sequences compiled in the JGI Comparative Plant Genomics Portal database (http://phytozome.jgi.doe.gov) (Tuskan *et al*., 2006; Goodstein *et al*., 2012; Nordberg *et al*., 2014)

3.4 Metabolomic mass spectrometry

Metabolomics is essentially the characterization and quantification of as many low molecular weight compounds in samples as possible, at least in relative terms (Dettmer *et al.*, 2007; Nicholson & Lindon, 2008). It is important because metabolic changes are the ultimate responses of biological systems to genetic modifications, representing the links between genotypes and phenotypes. However, reliable procedures for identifying and quantifying many plant metabolites have not yet been developed (Fiehn, 2002).

Gas chromatography mass-spectrometry (GC-MS) is a powerful analytical technique for analysing metabolites, but the key separation step is vaporization, so it can only be applied to analytes that are sufficiently volatile and resistant to thermal decomposition. Thus, it is often necessary to derivatize samples or extracts prior to GC-MS analysis, but even with derivatization, some metabolites are not amenable to analysis by this method (Dettmer *et al.*, 2007). In contrast, liquid chromatography (LC) using reversed-phase columns (RPLC) involves much milder differential partitioning of analyses between a liquid mobile phase and a stationary phase. Because analytes are not vaporised, it can accommodate a wider range of chemical species and no derivatization may be required. In the studies this thesis is based upon, both GC-MS and LC-MS were used in attempts to obtain information as many as possible of the primary and secondary metabolites present in the studied aspen samples, following standard procedures at the Swedish Metabolomic Centre (SMC), Umeå, Sweden. For reviews of mass spectrometry-based methods, see Dettmer *et al*. (2007) and Koek *et al*. (2011).

3.5 Plant carbohydrate and cell wall analysis

Plant carbohydrates and cell walls were analysed at the plant cell wall and carbohydrate analytical facility at UPSC/SLU. As described in Paper IV, the composition of polysaccharides was assessed by measuring monosaccharides released by methanolysis and acid hydrolysis (saccharification) following Gandla *et al*. (2015), in conjunction with Py-GC-MS techniques presented by Gerber *et al.* (2012; 2014).

3.6 Chemometrics and multivariate statistics

Biological systems analysis often generates massive datasets containing information on numerous responses, necessitating the use of chemometric techniques and multivariate statistics to untangle patterns in the data reflecting complex transcriptomic, proteomic and metabolomics changes in the studied material. We used PCA and both ‘traditional’ OPLS regression analyses and its most recent formulation (OnPLS), to analyze the data considered in Papers I-IV. PCA and OPLS were conducted using version 14.0 of the SIMCA software package (Umetrics, Umea, Sweden).

PCA is an unsupervised pattern recognition method that maps the main variation in multivariate datasets onto a low-dimensional subspace (Trygg & Wold, 2002 and 2003). It can be described as follows.

For a data matrix X with N rows (observations) in a K dimensional space (where K is the number of variables), PCA creates hyper planes in the K-dimensional space that approximate the data in the least squares sense and maximize the variance of the coordinates. Mathematically the data matrix X is modelled as

$$X = 1\bar{x} + TP' + E.$$

Where $\bar{x}$ is the column average vector, TP' is the matrix product that models the structure, T is the score matrix, P is the loading matrix, and E is the residual matrix.

The distribution pattern in a dataset can be examined using OPLS, which divides the systematic variation in the X data-block into two separate parts. One is predictive (denoted with the subscript p in the equation below) and describes the correlation between X and y. The other part is orthogonal (denoted with the subscript o) and describes the variation uncorrelated to y. An OPLS model may be unsupervised (if observations are classified into groups before modelling based on prior knowledge), supervised (if observations are not pre-classified), or discriminate (if a Y matrix of binary dummy variables is used) (Trygg & Wold, 2002 and 2003; Rantalainen *et al.*, 2008; Lofstedt *et al*., 2013).

For OPLS, the X-part of a single response variable or two-class discriminant model can be written as:

$$X = 1\bar{x} + t_p p_p' + T_o P_o' + E$$

and the y, OPLS model prediction, can be written as

$$y = \bar{y} + t_p q_p' + f,$$

where q_p' and f are the loading and residual vectors for y, respectively.

OPLS, as described above, was used in both stepwise and pairwise analyses. The stepwise, sequential, analyses involved first classifying sequential pairs of datasets representing biological transitions (notably, changes in expression patterns associated with transitions between word-formation zones or phases) and then creating a local Orthogonal Projections to Latent Structures-Discriminant Analysis (OPLS-DA) model for each transitional pair. Each acquired OPLS-DA model was evaluated and its significance was tested by cross-validation and ANOVA of the cross-validated residuals (CV-ANOVA) before proceeding to the next step. Next, the predictive component of each OPLS-DA model, representing the scale and change in variable expression was generated, referred to as the p_{dist} profile, which also represents the information about the direction and magnitude of change in the variable expression pattern between consecutive zones. Finally a PCA model of all the resulting p_{dist} profiles was created, providing an overview of the consecutive variation across the time/developmental series and used for biological interpretation.

The pairwise analytical procedure was similar, but OPLS models were constructed of differences within phases between controls and modified (transformed) material. Individual models were then connected to obtain an overall biological interpretation of the whole time series or process.

As already mentioned, the y term in the OPLS equation may be a separate response matrix or data-block, which can be analyzed by two-way OPLS (O2PLS) which models and predicts both X and y and separates the structured noise in X and y from their joint $X - y$ covariation (Trygg & Wold, 2002 and 2003). However when multiple datasets are involved, OnPLS (an extension of O2PLS) provides better generalization of the data (Trygg & Wold, 2002 and 2003; Rantalainen *et al.*, 2008; Lofstedt *et al*., 2013).

OnPLS modelling efficiently handles differences in sizes and connections of datasets obtained using different platforms (independently of platform order), while simultaneously linking intrinsic flows of information between different platforms or datasets. This enables robust interpretation of joint factors affecting variation and facilitates biological interpretation of data from individual, connected or correlated platforms. OnPLS analyses data acquired from all platforms or sources simultaneously and separates the corresponding variation into: a shared part (common to all blocks, referred to as globally joint variation); local parts (shared by some, but not all, blocks; referred to as locally joint variation); and unique parts (variation that is unique to a specific platform or dataset) (Lofstedt *et al.*, 2013; Srivastava *et al*., 2013). Figures 4a and 4b show numbers of possible levels of biological examination enabled by OnPLS in cases involving three (7 or more possible levels of analysis) and five (21 or more possible levels of analysis) different platforms/data sources, respectively. This ability to examine relationships between different datasets at multiple levels enables the comprehensive, correlated and detailed characterization of complex and diverse responses, which is essential for systems biology.

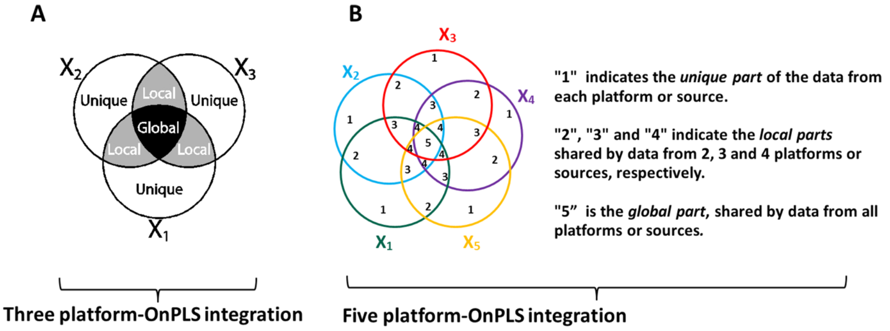
A
X2
X3
Unique
Local
Unique
Global
Local
Local
Unique
X1
Three platform-OnPLS integration
B
X3
X2
X4
X1
X5
"1" indicates the unique part of the data from each platform or source.
"2", "3" and "4" indicate the local parts shared by data from 2, 3 and 4 platforms or sources, respectively.
"5" is the global part, shared by data from all platforms or sources.
Five platform-OnPLS integration

Figure 4. An illustration of what OnPLS does in cases involving three data blocks (X_1, X_2 and X_3) and five blocks (X_1, $X_{2,}$ $X_{3,}$ X_4 and X_5) from separate platforms. Panel A is adapted from Srivastava *et al.* (2013). Panel B is adapted from illustrations in Paper IV.

The equation of the first matrix in an OnPLS model for three blocks can be written as:

$$\boldsymbol{X}_1 = \underbrace{(\boldsymbol{X}_1 \cap \boldsymbol{X}_2 \cap \boldsymbol{X}_3)}_{globally\ joint\ part} + \underbrace{((\boldsymbol{X}_1 \cap \boldsymbol{X}_2) \backslash \boldsymbol{X}_3) + ((\boldsymbol{X}_1 \cap \boldsymbol{X}_3) \backslash \boldsymbol{X}_2)}_{locally\ joint\ part} + \underbrace{(\boldsymbol{X}_1 \cap \overline{(\boldsymbol{X}_2 \cup \boldsymbol{X}_3)})}_{unique\ part}$$

where ∪ is the set union operator, ∩ is the set intersection operator, \ is the set difference operator and $\bar{X}$ is the set complement (adapted from Srivastava *et al.* (2013).

Papers I and III focus on global expression analysis, which provides a simultaneous, unifying biological theme (Subramanian *et al.*, 2005), and comprehensive overlap across all studied platforms. More detailed descriptions of PCA, OPLS, its discriminant analysis variant (OPLS-DA), and OnPLS can be found in Trygg & Wold (2002 and 2003), Rantalainen *et al.* (2008) and Lofstedt *et al.* (2013). In the study reported in Paper I, OnPLS was used to integrate data from the following analytical platforms: transcriptomic (microarrays), proteomic (UPLC-QTOF-MS) and metabolomic (Gas Chromatography-Time of Flight-Mass Spectrometry (GC-TOF-MS), Ultra Performance Liquid Chromatography-Mass Spectrometry (UPLC-MS), and Ultra High Performance Liquid Chromatography coupled to a Linear Quadrupole Ion Trap-Orbitrap Mass Spectrometer (UHPLC-LTQ/MS).

Paper II describes a stepwise modelling approach in which sequential OPLS models were created to represent changes in successive developmental zones within the studied stems.

OPLS and pairwise modelling were used in Paper III to compare tension and normal wood at distinct developmental stages. These approaches enabled a comprehensive investigation of poplar wood development and interpretation of the relevant biological pathways using several multivariate statistical tools.

Paper IV presents an analysis of data obtained from nine platforms. OnPLS was used to integrate data from transcriptomic microarray results and four sources of metabolomic data: UPLC-QTOF-MS, GC-TOF-MS, UHPLC-MS, UHPLC-LTQ-MS, and Py-GC-MS described in the Paper. Data from four other platforms were analyzed individually: monosaccharide analysis by methanolysis followed by silylation using hexamethyldisilazane and trimethylchlorosilane then GC-MS analysis, monosaccharide analysis by acid hydrolysis and high-performance and anion exchange chromatography (HPAEC), saccharification analysis, and phenotypic measurements (growth, height, density). The results of these individual analyses were subsequently correlated with the OnPLS results during biological interpretation.

3.7 Pathway analysis and visualisation

Pathway analysis and visualization is essential for comprehensive understanding of large omics datasets and thus is increasingly common in plant sciences (Porth *et al.,* 2013).

We examined pathways affected in the material we analysed using information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa & Goto, 2000; Kanehisa *et al*., 2011) and MAPMAN (a user-driven tool providing pathway and biological process information (Thimm *et al.,* 2004). The information obtained from the KEGG and MapMan databases was painted using Paintomics Version 2.0 (http://www.paintomics.org; García-Alcalde *et al.*, 2011), to map and visualize the associated transcripts, proteins and metabolites. This enabled efficient examination, connection and visualization of several pathways, providing an overall view of the corresponding omic changes, and revealing correlated transcript-, metabolite-, and protein-level changes as shown in Figure 5.

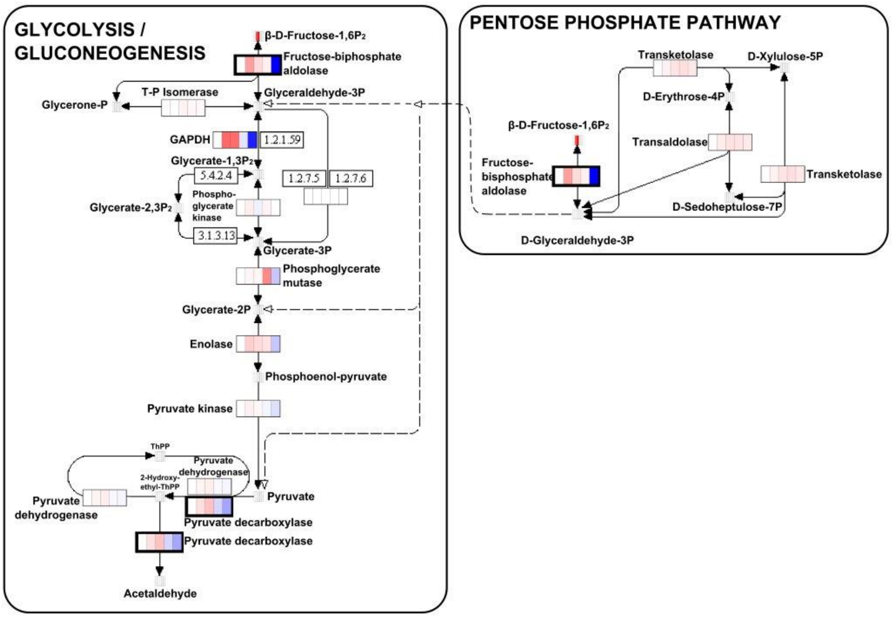

Figure 5. KEGG Glycolysis/Gluconeogenesis pathway and Pentose Phosphate Pathways 'painted' with transcriptomic, proteomic and metabolomic data from the targeted OnPLS model. Black-bordered entry boxes indicate significant differences between transgenic and WT plants at both transcript and protein levels. The first three sections of each gene box (left to right) indicate WT, AS-SOD9 and AS-SOD24 transcript levels, respectively, and the last two protein levels in AS-SOD9 and AS-SOD24 lines, respectively. The first sections in the metabolite entry boxes represent WT and the coloured boxes levels in the AS-SOD9 and AS-SOD24 lines. Reduced levels in the transgenics are coloured blue and increased levels red. (Figure 4 from Srivastava *et al.*, 2013)

Wood expression patterns across the series of wood development stages were visualized using PermutMatrix software v.1.9.3 (Caraux & Pinloche, 2005) in Papers II and III.

4 Results and discussion

4.1 Proper integration of multiple data measurements

The study reported in Paper I was our first application of systems biology to analyse oxidative stress responses in aspen. Oxidative stress tolerance mechanisms in plants are complex (Mittler, 2002; Srivastava *et al*., 2013), so a multivariate modelling approach was used to statistically integrate information on global (transcriptomic, proteomic and metabolomic) responses to oxidative stress in the cambium region of the *Populus* model system. Samples were collected from the cambial region of wild-type controls and aspen plants expressing hipI-SOD transcripts in antisense orientation. Data were generated using transcriptomic (microarray analysis), proteomic (UPLC-QTOF-MS), and metabolomic (GC-TOF-MS, UPLC-MS, and UHPLC-LTQ-MS) platforms. These data were then statistically integrated using the most recent formulation of the established OPLS approach, OnPLS.

Both abiotic and biotic stressors can disrupt the cellular redox state in plants, thereby causing rises in levels of Reactive Oxygen Species (ROS), with corresponding effects on various physiological and developmental processes (Apel & Hirt., 2004; Scandalios, 2005; Mittler *et al.*, 2011). Oxidative stress responses are complex (Mittler, 2002) and ROS have been shown to play key roles in plant stress signalling cascades affecting numerous biological processes including growth, development, and responses to diverse stimuli. They are also hypothetically involved in crosstalk with several signalling pathways that regulate various responses in plants (Morgan & Liu, 2011; Srivastava *et al.*, 2013; Baxter *et al.*, 2014). However, few attempts have been made to comprehensively and simultaneously characterize related transcriptomic, proteomic and metabolomics profiles, which is essential for fully elucidating stress responses (Higashi *et al.*, 2006). The systems-level approach described in Paper I is therefore essential for understanding the complex interplay between gene regulation, post-translational modifications and metabolic fluxes in aspen (Yuan *et al.*, 2008; Paper I; Srivastava *et al.*, 2013). The multi-omic profiling required for full analysis generates extremely large, complex datasets, and biologically meaningful interpretation of such datasets requires use of powerful systems biology techniques for integrating multidimensional information into networks.

Results presented in Paper I provided significant information about functional and pathway responses to oxidative stress from the OnPLS model.

We observed that as the plants were stressed protective antioxidant processes were induced, as manifested in the degradation of oxidized proteins. This appeared to be mediated by an induced, free 20S proteasome which plays a non-proteolytic role in transcriptional regulation with elevated proteasome activity, enhancing the responsiveness of plants' signal transduction pathways and increasing their stress resistance levels by accelerating removal of damaged proteins (Kurepa & Smalle, 2008). This was also evident from the upregulation at both protein and transcript levels of PBA1 (20S proteasome beta subunit A1, POPTR_0018s14290) in the transgenic plants. The *Arabidopsis* homologue of the 20S proteasome *At4g31300* has been previously been shown to be up-regulated in response to stress and involved in cell wall division/regeneration (Kurepa & Smalle, 2008; Polge *et al*., 2009).

Paper I also highlighted an important role of a less intensively investigated pathway (the pentose phosphate pathway, PPP) in *Populus* model systems, further supporting a systems biology approach. Notably, many transcripts and proteins involved in carbon metabolism pathways, such as the glycolysis/gluconeogenesis and PPP were strongly affected in the transgenics. However, while pyruvate decarboxylase (POPTR_0016s12760, PDC1.5), fructose-bisphosphate aldolase (POPTR_0006s17940) and glyceraldehyde-3-phosphate dehydrogenase (POPTR_0015s10330, GAPDH 1.2) were upregulated at transcript level, they were downregulated at protein level (Figure 5). A cytosolic fructokinase (POPTR_0007s01850), transaldolase (POPTR_0003s16030) and transketolase (POPTR_0002s14730) were upregulated at both transcript and protein levels. Transketolase and transaldolase both provide reversible links between the PPP and glycolysis (Matsushika *et al.*, 2012), and PPP and glycolysis have been suggested to contribute to ROS balance and scavenging (Casado-Vela *et al.*, 2005; García-Leiro *et al.*, 2010; Krüger *et al.*, 2011). In combination with the downregulation of sucrose and xylose levels, these observed changes in transcript and protein levels indicate a shift towards the breakdown of carbohydrates with a profound rearrangement of primary carbon metabolism in response to an imbalanced redox state. These observations are very important since sugar signaling is strongly linked to plant growth and development (Lastdrager *et al.*, 2014). They also reveal strong connections between glycolysis, PPP, carbon metabolism, oxidative stress and an "emergency strategy" that reroutes metabolic fluxes from glycolysis to the PPP as an immediate and protective response to oxidative stress (Ralser *et al.*, 2007).

Ribosomal proteins are involved in protein synthesis and plant growth, which are highly energy demanding processes (Roux & Topisirovic, 2012; Lastdrager *et al.*, 2014). Ribosomal proteins (r-proteins) were highly downregulated, indicating downregulation of protein synthesis, in the results in Paper I. Ribosome biogenesis and mRNA translation are also highly energy-demanding processes (Roux & Topisirovic, 2012), so this downregulation would provide major savings in energy consumption and the low energy levels would trigger cells to switch to an energy preservation mode. This supports the hypothesis that r-protein downregulation may contribute to a reprogramming of energy transformation processes that consequently affects maturation and cell death-associated signals in the transgenic plants (Paper I; Srivastava *et al.*, 2013).

These findings require further validation in plants since most previous observations supporting them were obtained from single-level studies of other systems, corroborating the need for more detailed multi-level (transcript-protein-metabolite) systems approaches.

4.2 Developmental regulators and molecular foundations of wood development.

Detailed understanding of wood formation is essential for numerous aspects of both pure and applied plant science, due to the associated cellulose and lignin biomass production (Plomion *et al*., 2001). For example, knowledge of molecular events involved in wood formation is essential for screening and forward genetics approaches to modify wood quality in desired ways (Mishima *et al*., 2014). However, transcriptomic data do not always correlate with levels of expressed proteins (Fournier *et al.*, 2010) due to effects of post-translational modifications and variations in turnover rates, so proteomic analysis is essential to characterize key transitions during wood development. Proteomics offers a high-throughput approach for determining key regulators during successive stages of wood developmental. However, because of difficulties related to sample preparation and low protein concentrations, most studies have largely focused on transcripts, selected protein families and/or relatively large wood sections (Vander *et al.*, 2000; Schrader *et al*., 2004; Bylesjö *et al.*, 2008; Dharmawardhana *et al.*, 2010; Zhong *et al*., 2011; Lin *et al*., 2014).

In the study reported in Paper II, proteomic (UPLC-QTOF-MS) analysis was used to quantify protein expression in tangential, 20-160 µm thick sections spanning all wood development zones in *Populus tremula* from phloem, through cambium and the expansion, xylem maturation and cell death zones. This high-resolution developmental series comprised 482 sections from four, 47-year-old trees harvested in the forest.

Paper III presents a proteomic analysis of global expression of phloem, cambium and xylem cell types in normal and tension wood of *Populus*. This revealed several cellular processes influencing tension wood formation through changes, for instance, in the expression of proteins involved in starch and sucrose metabolism, protein signalling, and possible reprogramming of the plant's energy transformation machinery, reflecting the high energy demands associated with the process.

The study presented in Paper II involved modelling by a combination of PCA, OPLS and an enhanced stepwise linear approach that identified major transitions in global protein expression profiles, pinpointing (*inter alia*) the location of the cambium division leading to phloem and xylem cells, and the secondary cell wall formation zone.

Potential regulators of phloem functions (Schrader *et al*., 2004), which include several sucrose transporters (Potri.015G029100, SUS6.1; Potri.004G081300, SUS6.2; Potri.017G139100, SUS5; and Potri.012G037200, SUS6.3), were significantly upregulated in phloem compared to cambium (Figure 6 in the thesis and Additional File 2: Tables S2.1-7 in Paper II), supporting the hypothesis that active sucrose transport from the phloem into the cambium region and beyond is important for wood formation (Mahboubi *et al*., 2013).

Wood formation in trees is initiated in the vascular cambium and involves undifferentiated cambial cells developing into phloem (outwards) and xylem cells (inwards) through the processes of division, expansion, secondary wall formation, lignification, and finally (in xylem) programmed cell death (Hertzberg *et al.*, 2001). The multivariate analysis described in Paper II revealed clear, progressive trends across the wood-forming zone with the phloem, cambium and expansion zone samples clustered in the top left quadrant of score plots, the X1, X2 and X3 xylem samples segregated mainly to the right, while xylem-X4 samples clustered in the bottom left quadrant (Figure 6; see the Methods section in Paper II for descriptions of the X1-4 xylem zones). There were also clear trends, not only between sequences of clusters in the PCA model plot including data from all samples collected from all sampled trees, but also within clusters, as illustrated by the almost linear transition within and through the Xylem-X3 zone to the X4 zone, not previously reported to our knowledge. Figure 6 shows a novel visualisation plot which can be used to facilitate time series analysis in future tree selection.

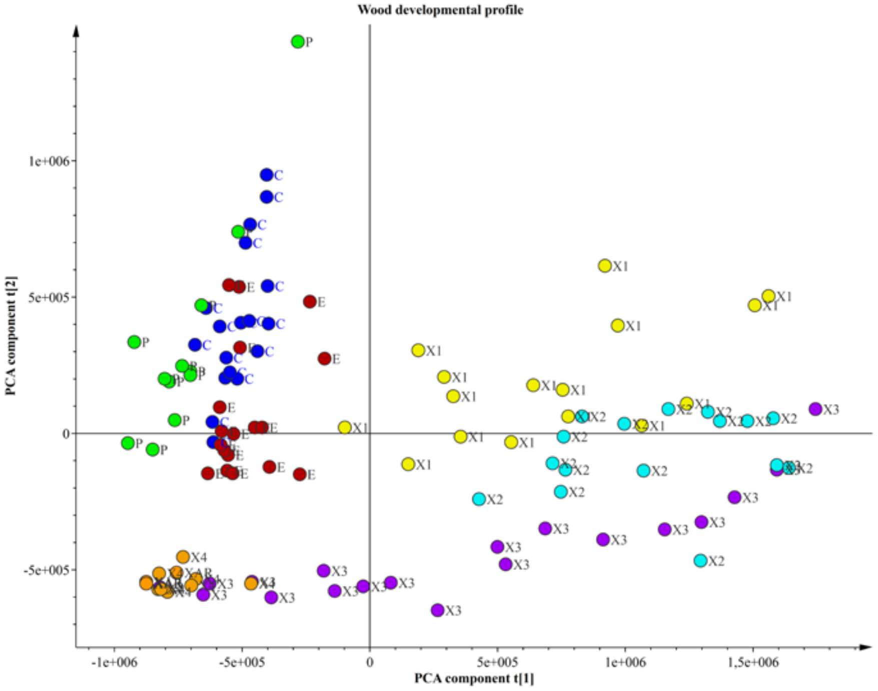

Figure 6. Joint genotype effect scores from the global PCA model. The PCA model includes data from all samples, from all four sampled trees. The score plots revealed clear progressive trends across the series of development zones in the sampled wood. P, C, E, X and XAR refer to phloem, cambium, expansion zone, xylem (divided into four sub-zones: X1, X2, X3 and X4), and xylem annual ring border, respectively (Paper II).

A major component of primary cell walls is pectin, and changes in pectin composition identified in poplar in relation to cambial cell differentiation indicate that pectin methylesterases are useful early markers of cambial differentiation into either phloem or xylem (Guglielmino *et al.*, 1997). Pectin methylesterases have been shown to act as negative regulators of symplastic and intrusive growth of developing wood cells in tissues of hybrid aspen, causing changes specifically in expanding wood cells (Siedlecka *et al.*, 2008). We observed that the pectin acetylesterase protein Potri.003G046200 was downregulated in the Cambium-Expansion Zone and Expansion Zone-Xylem

X1 transitions, while the pectin methylesterase protein Potri.001G162400 was upregulated in the phloem-cambium transition and downregulated in the cambium-expansion zone transition. These observations show that pectin is degraded and influences regulation of intrusive growth in woody tissues.

Roles of some proteins in specific zones have been at least partially elucidated in previous studies but analysis such as this provides opportunities for more comprehensive interpretation of stage-specific profiles and processes, which is essential for understanding wood formation. Moreover, this analysis identified other key proteins and associated pathways underlying these developmental landmarks, some of which are shown in Figures 7 and 8, and discussed in detail in Paper II.

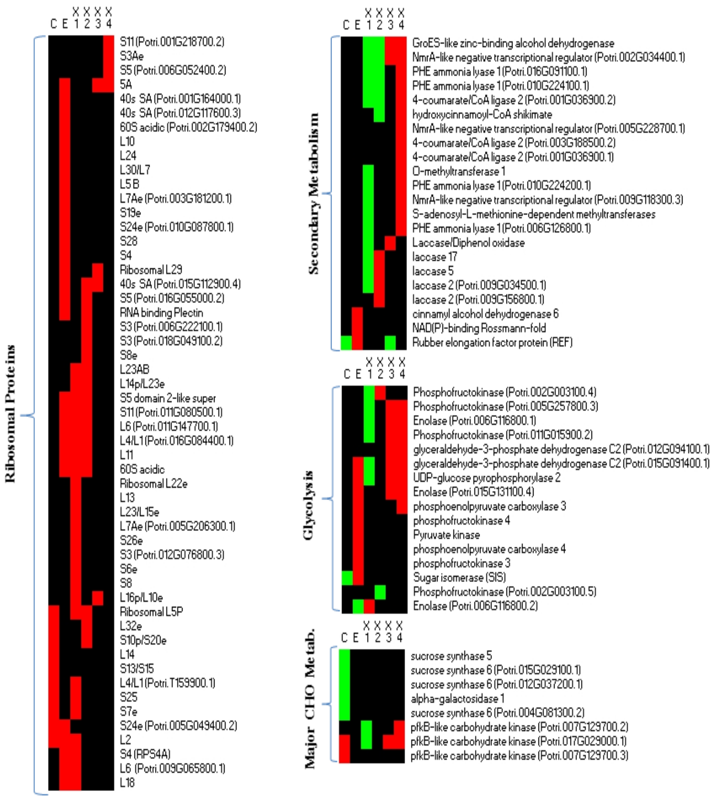

Figure 7. Visualization of patterns and trends of differentially expressed protein families across the series of development zones. Profiles of protein families encoding or involved in ribosomal biogenesis, secondary metabolism; glycolysis and major CHO metabolism. The red, green, and black color-codes indicate downregulation, upregulation and no change in expression,

respectively, in indicated transitions. Symbols: C, Phloem-Cambium transition: E, Expansion Zone; X, Xylem (divided into X1, X2, X3, X4 zones).

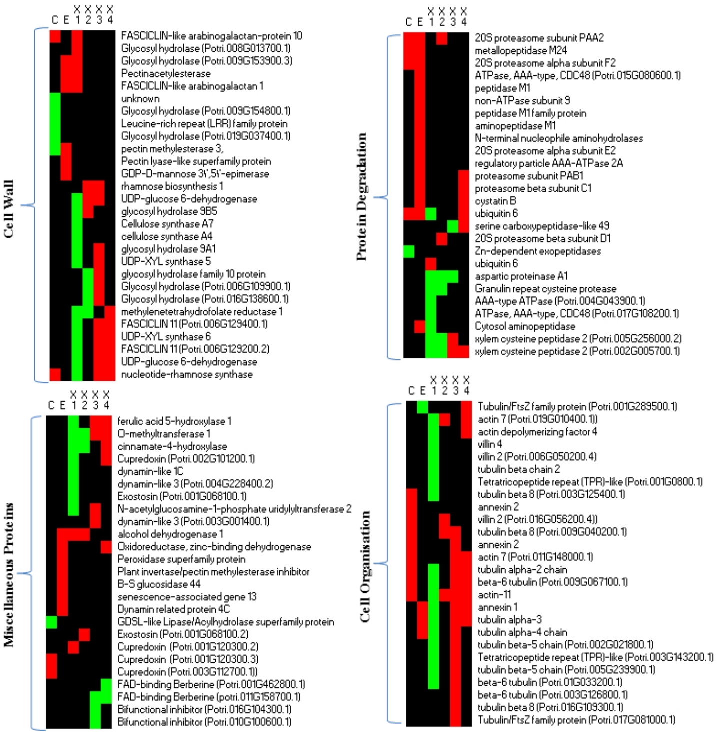

Figure 8. Visualization of patterns and trends of the differentially expressed protein families across the series of development zones. Profiles of protein families encoding or involved in cell wall, cell organization, protein degradation and miscellaneous proteins. The red, green, and black color-codes show downregulation, upregulation and no change in expression, respectively, in indicated transitions. Symbols: C, Phloem-Cambium transition: E, Expansion Zone; X, Xylem (divided into X1, X2, X3, X4 zones).

Tension wood (TW) is produced as a result of mechanical stress in plants and has distinctive structural, chemical composition and wood properties (Mellerowicz & Sundberg, 2008). The processes involved in its formation include stimulation of cambial growth at the upper stem side (TW), inhibition of growth at the lower (opposite wood) side, and a switch from vessel to fibre differentiation (Hellgren *et al*., 2004). Paper III reports a proteomic analysis of global expression of phloem, cambium and xylem cell types in normal and tension wood of *Populus*. Understanding of the mechanisms underlying TW formation are still limited, and most previous analyses of the process have focused on differential gene expression profiles (Andersson-Gunneras *et al*., 2006; Chen *et al.*, 2015). In contrast, we have provided the first (to our knowledge) extensive, multivariate and high-resolution stage-specific proteomic profiles. In our study, we focused on proteins based on the quantification of unique peptides in order to examine the expression pattern along the wood series.

Various aspects of several cellular processes influencing TW formation, such as the high energy demand manifested in variations in proteins involved in the starch and sucrose metabolism pathway, protein signalling, and possible reprogramming of the plant's energy transformation machinery, were unravelled and discussed in detail in Paper III. The key roles of many proteins, e.g. several members of the fasciclin-like arabinogalactan (FLA) protein family, S-adenosylmethionine synthetase family protein (SAM), sucrose synthase (SUS) and UDP-glucose pyrophosphatase (UGP) in Normal wood (NW) and TW were unravelled in Paper II.

In the imminent future sequences of most genes of several plant species will be available in public databases, but elucidating the biological functions of some proteins they encode may pose a major challenge (Vander *et al*., 2000). Proteomic research is likely to play valuable roles in efforts to address this challenge as rapid advances are massively enhancing both the quantity and quality of information it provides. Another important contribution of Papers II and III in this context is that they provide first steps towards construction of a protein stage-specific resource or database, and foundations for detailed characterization of key players throughout the entire wood development process in stems of plants generally and *Populus* particularly.

4.3 Systems analysis of variation and phenotypic perturbations.

According to Yang *et al*. (2014), "System biology is defined as the study of interactions among biological components to integrate genes, metabolites, proteins, regulatory elements, and other biological components". We adopted this approach most intensively in an investigation of the role of secretory carrier membrane proteins (SCAMPs) in wood formation (Paper IV). These are highly conserved 32–38 kDa proteins that are involved in membrane trafficking, but their trafficking routes in aspen are largely unknown (Law *et al*., 2012). In our study (the first comprehensive investigation of SCAMPs' roles in plants to our knowledge) we addressed effects of nine biological factors and functions of *Populus* SCAMP genes in transgenic *Populus* trees carrying an RNAi construct for *Populus tremula x tremuloides* SCAMP3

(PttSCAMP3; Potri.019G104000), which suppressed expression of two PttSCAMP genes: PttSCAMP3 and PttSCAMP6 (Figure 9).

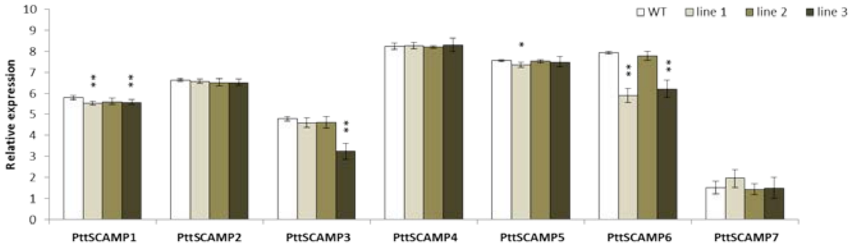

Figure 9. Expression of the PttSCAMP genes in the wild type and three PttSCAMP3 RNAi lines. Relative expression values ±SD were obtained from the RNA sequencing datasets.

Our analysis revealed that the secondary xylem of the transgenic trees exhibited increased deposition of all major secondary wall polymers, including carbohydrates, lignin and suberin, resulting in increases in both wood density and formation of suberized bark. These findings are summarised in Figure 10 (adapted from Paper IV).

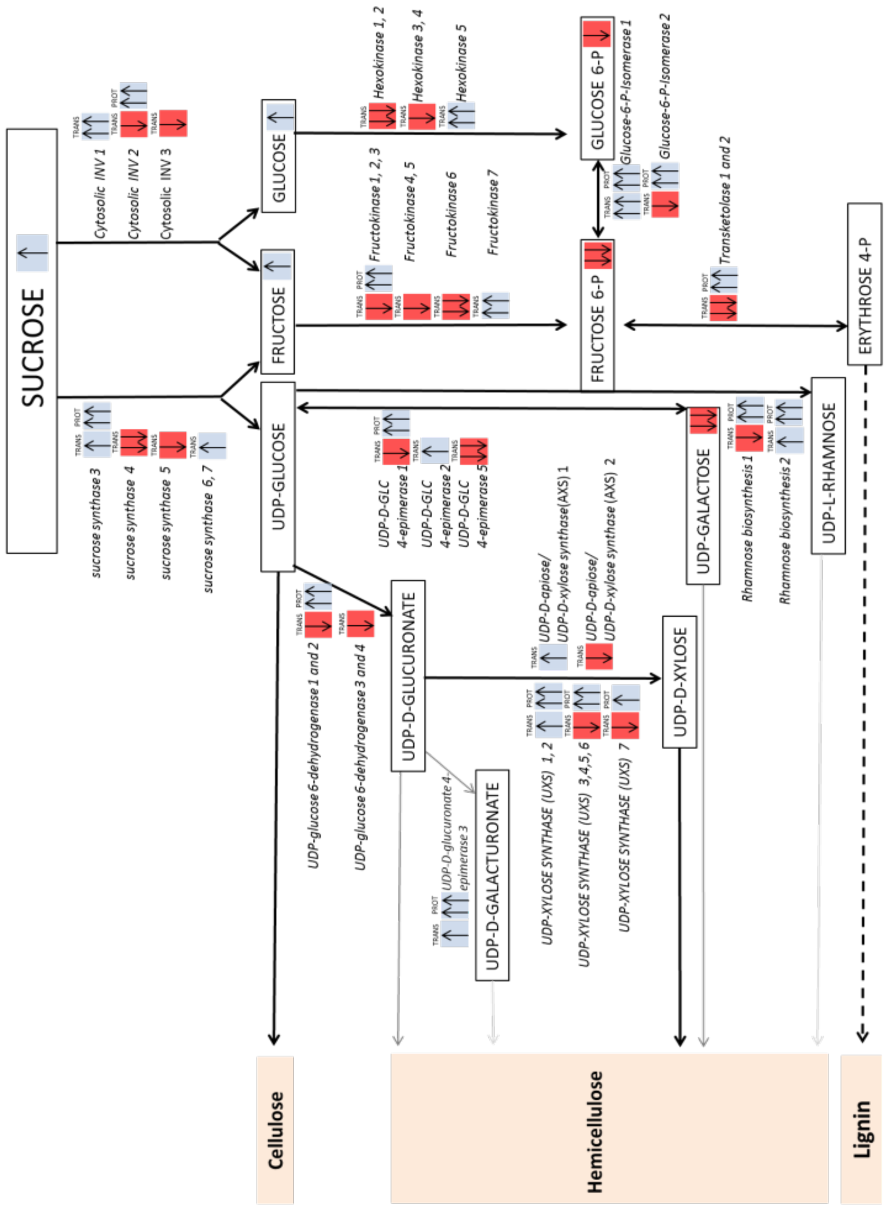

Figure 10. Biosynthetic pathways of cellulose, hemicellulose and lignin. The abundance of the secondary cell wall-related metabolites, transcripts (Trans) and proteins (Prot) in the PttSCAMP3 RNAi lines compared to wild type plants. ↑ in blue square, upregulated in PttSCAMP3 RNAi lines; ↑↑ in blue, significantly (|p(CORR)| >0.5) upregulated in PttSCAMP3 RNAi lines; ↓ in red square, downregulated in PttSCAMP3 RNAi lines; ↓↓ in red square, significantly (|p(CORR)| >0.5) downregulated in PttSCAMP3 RNAi lines. Metabolites are shown in bold boxes with the colour codes as indicated above.

The genes and proteins depicted in the figure correspond to the following *Populus trichocarpa* gene models according to JGI V3.0: Fructokinase 1 (Potri.017G126300); Fructokinase 2 (Potri.007G129700); Fructokinase 3 (Potri.017G029000); Fructokinase 4 (Potri.012G132700); Fructokinase 5 (Potri.004G089300); Fructokinase 6 (Potri.015G134900); Fructokinase 7 (Potri.019G063600); Hexokinase 1 (Potri.001G190400); Hexokinase 2 (Potri.005G238600); Hexokinase 3 (Potri.009G050000); Hexokinase 4 (Potri.018G088300); Hexokinase 5 (Potri.001G254800); Cytosolic INV 1 (Potri.014G188100); Cytosolic INV 2 (Potri.013G110800); Cytosolic INV3 (Potri.019G082000); Sucrose synthase 3 (Potri.002G202300); Sucrose synthase 4 (Potri.006G136700); Sucrose synthase 5 (Potri.018G063500); Sucrose synthase 6 (Potri.004G081300); Sucrose synthase 7 (Potri.012G037200); Glucose-6-phosphate isomerase 1 (Potri.008G118900); Glucose-6-phosphate isomerase 2 (Potri.002G104000); UDP-D-apiose/UDP-D-xylose synthase 1 (Potri.009G150600); UDP-D-apiose/UDP-D-xylose synthase 2 (Potri.004G189900); UDP-glucose 6-dehydrogenase 1 (Potri.017G092000); UDP-glucose 6-dehydrogenase 2 (Potri.004G118600); UDP-glucose 6-dehydrogenase 3 (Potri.010G159800); UDP-glucose 6-dehydrogenase 4 (Potri.008G094300); UDP-GLC 4-epimerase 1 (Potri.003G123700); UDP-GLC 4-epimerase 2 (Potri.003G140900); UDP-GLC 4-epimerase 5 (Potri.001G090700); UDP-D-xylose synthase (UXS) 1 (Potri.006G214000); UDP-D-xylose synthase (UXS) 2 (Potri.014G129200); UDP-D-xylose synthase (UXS) 3 (Potri.010G207200); UDP-D-xylose synthase (UXS) 4 (Potri.002G204400); UDP-D-xylose synthase (UXS) 5 (Potri.001G237200); UDP-D-xylose synthase (UXS) 6 (Potri.008G053100); UDP-D-xylose synthase (UXS) 7 (Potri.016G080500); Transketolase 1 (Potri.002G146300); Transketolase 2 (Potri.014G068200); UDP-D-glucuronate 4-epimerase (Potri.002G146500); Rhamnose biosynthesis 1 (Potri.006G272700); Rhamnose biosynthesis 2 (Potri.001G383500).

Systems application of multivariate OnPLS in the modelling of five large datasets (describing the transcriptome, proteome, GC-MS metabolome, LC-MS metabolome and pyrolysis-GC/MS metabolome) collected from the secondary xylem tissues of stems revealed systemic variation of the studied variables in the transgenic lines as well as changes that correlated with changes in abundance of the cell wall polymers. Moreover, the OnPLS model identified covariation of a large number of proteins such as those related to secretion, endocytosis, and cell wall biosynthesis, many of which were more abundant in the PttSCAMP transgenic lines than in the wild type.

Although we focussed on the globally joint component in Papers 1 and IV, our enhanced OnPLS and pathway approach provides possibilities for future studies to investigate the unique variation in each block or platform, locally joint variation and residuals, as illustrated in Figure 11.

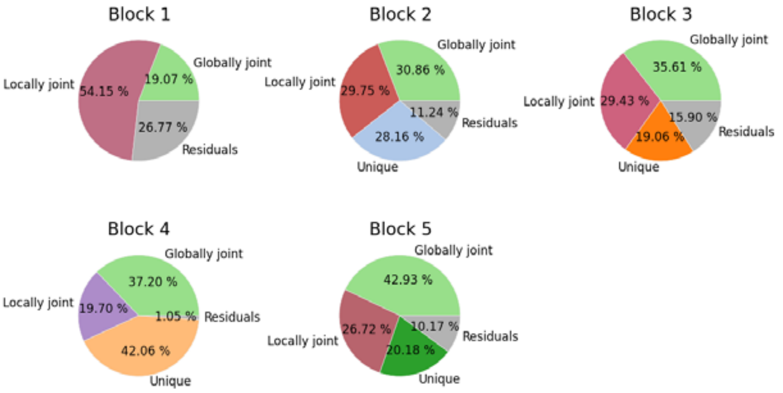

Figure 11. Illustration of variation explained and possibilities for applying the OnPLS approach to blocks/platforms: the globally joint component, locally joint component, unique variation and residuals.

4.4 Challenges in omics analysis

Systems biology investigations, such as those presented in this thesis, pose several challenges that are discussed in this section. Although transcriptomic, proteomic and metabolomic analyses are interrelated, and each of the experiments may probe patterns in a defined space- and time-frame, they may produce datasets of radically different types and sizes, which are challenging to combine in order to obtain a complete overview (Benkovic *et al.*, 2013). The differences in data size and type can be addressed by normalisation procedures or the integration capabilities of OnPLS to some extent, but many practical challenges remain in cross-experiment comparisons, and are likely to increase as the number of datasets or platforms is likely to rise in the imminent future (Joyce & Palsson, 2006). Equally challenging and crucial concerns arise when matching transcripts to protein families or comparing different species (e.g. *Populus* spp. and *Arabidopsis thaliana*). Searches for similar sequences in databases (e.g. BLAST searches of Phytozome) may be highly valuable for identifying orthologous genes or protein matches, but this is time consuming and may require advanced bioinformatics knowledge (Tuskan *et al*., 2006; Goodstein *et al*., 2012; Nordberg *et al*., 2014).

However, pathway analysis can greatly reduce the complexity of the patterns in considered datasets and massively increase explanatory power during the interpretation of observations, thereby enormously enhancing insights into the processes reflected by differentially expressed genes, proteins and metabolites (Khatri *et al*., 2012). Numerous studies have examined one or two pathways, but in the studies presented here attempts have been made to characterise changes in all possible pathways to examine all possible biological responses and processes captured in the datasets, and the inter-connections and information flows between pathways. A complicating factor is that integrating transcriptomic, proteomic, and metabolomics datasets generates vast amounts of information related to numerous genes, gene products, biological responses and/or pathways. This creates challenges for selecting the aspects that are most relevant to the focal biological questions, while ensuring that others (which may be less significant according to the multivariate analysis, but still important) are not overlooked. The challenges associated with biological complexity and integrating fragmented aspects of biological research are being intensively addressed (Efron & Tibshirani, 2007; Akula *et al*., 2009).

Localisation and functional annotation of identified transcripts and proteins (which are often critical for robust biological interpretation of acquired data) are also challenging. These crucial data processing steps often involve Gene Ontology (GO) classifications and/or inferences of localisation from published experimental or computational work, but experimental data are rarely available for all the genes or taxa covered, and many genes are expressed in multiple locations (Ashburner *et al*., 2000; Buza *et al*., 2008; Gene Ontology Consortium. 2013).

Evolutionary events such as whole-genome and segmental duplications have raised further challenges for orthologous gene identification across multiple species, and hence for interpreting and comparing results from poplar studies to published findings regarding other taxa (Patel *et al*., 2012). Equally challenging and crucial problems are posed by isoforms and paralogs in proteomic analysis. These are currently being addressed, to some extent, by focusing on unique peptides characteristic for each protein or protein isoform, using stable isotopes to monitor individual protein turnover rates or specific sets of peptides for targeted analysis. However, post-transcriptional events (splicing) and post-translational changes (including modifications, proteolytic processing and complex formation) still pose practical problems that hinder quantitative proteomics, and further advances in analytical techniques are required to solve them (Doherty *et al*., 2005; Kuster *et al*., 2005; Domon *et al*., 2006).

All platforms (transcriptomic, proteomic, metabolomic and wood chemistry) are rapidly developing and the challenges discussed above are being intensively addressed (Srivastava *et al*., 2013). Thus, further advances are likely to massively increase the scope of omic research. However, my colleagues and I believe that the approach presented in this thesis provides sound foundations for overcoming some of these challenges and improving understanding of plants’ responses and development.

5 Conclusions and future perspectives

The work described in this thesis has introduced new systems biology modelling strategies that enable the efficient integration of complex multi-platform datasets in order to increase our understanding of correlated multi-level and connected responses to perturbations, or developmental processes, in plants.

We began in the study presented in Paper I by proposing an enhanced framework for investigating and understanding multi-level oxidative stress responses and their influence on phenotypic variation in transgenic hipI-superoxide dismutase *Populus* plants. This analysis provided systems-level information on the biological pathways involved in the oxidative stress response. A previous attempt along the same lines using the O2PLS method had failed to adequately integrate the diverse data blocks because of its non-symmetrical nature and order-dependence. The new approach adopted in the study was based on the multivariate OnPLS method, which is fully symmetrical and data block-independent. The analysis revealed new multilevel biological information and highlighted results that could not have been readily obtained using traditional statistical methods.

We continued by proposing a new modelling strategy: a combination of PCA, OPLS modelling and an enhanced stepwise linear modelling approach linking profiles of successive zones (Paper II). This technique identified major transitions in global protein expression, pinpointing (among other things) the location of the cambium division leading to phloem and xylem cells, and the secondary cell wall formation zone in aspen. Our analysis provided the first clear, transparent, stepwise overview of the entire wood formation process in aspen, and illustrated the potency of combined application of several multivariate statistical tools.

The study presented in Paper II was based on data gathered from four 47-year-old trees. Paper III describes a similar study that compared the biological differences between normal and tension wood in different zones. We applied a multivariate approach related to the method described in Paper II, focussing on pairwise correlations between the tension and normal wood in each zone. The method was based on a combination of PCA and pairwise OPLS, with subsequent complementary evaluation of the results by multiple pairwise univariate analyses.

Finally, we sought to consolidate our systems understanding by using a combined multivariate approach to integrate data from nine platforms to investigate as much variation associated with selected system perturbations as possible. A multidisciplinary approach involving the most recent formulation of the OnPLS method was used to analyze a set of transgenic *Populus* trees that were identified as having a superior rate of biomass production. OnPLS was used to create a model integrating data from five separate platforms. The results obtained indicate that the PttSCAMP proteins control membrane trafficking in fine-tuning of the abundance of cell wall polymers. Consequently, targeted modification of the SCAMP proteins could be a powerful tool for forest biotechnology, e.g. for increasing deposition of secondary cell wall material in wood without impairing trees' overall growth.

The systems approach adopted in the research underlying this thesis provided new understanding of several biological processes, including oxidative stress responses, tension wood formation and secondary cell wall formation. It also enhanced our understanding of wood biosynthesis through the simultaneous characterisation of changes in transcript, protein and metabolite expression. A key outcome was the creation of a broad reference genetic data collection for both field-grown and transgenic lines, providing a valuable resource and tools for future investigations of biological processes and pathways. We anticipate that the strategies and results presented in this thesis will significantly facilitate the future design and generation of forest trees with improved properties.

In summary, our proposed systems approach provides a comprehensive understanding of various biological processes and pathways in wood formation and will facilitate the acquisition of new insights into molecular developmental processes in aspen. The multivariate OnPLS method applied (incorporating recent advances) allowed identification of co-variation within several datasets of different types, such as those derived from the transcriptomic, proteomic and metabolomic analyses, in a manner that would not be possible using standard statistical methods. Moreover it revealed a critical role of the SCAMP-dependent pathway in growth of secondary cell walls in stems, and provided a systems-level interpretation of biological responses and pathways controlled by SCAMP proteins in *Populus* woody tissues.

Future systems biology approaches may focus on specific processes or pathways controlling biomass production, or on selected classes of proteins such as the cellulose synthases, peroxidases and glycoproteins involved in aspen wood formation. We have provided a resource for biological interpretation from omics analysis, but there is still a need for further development of tools (notably user-friendly integration and visualisation tools) that can integrate and efficiently highlight integrated biological pathway responses from multilevel omics analysis.

6 References

Akula, S. P., Miriyala, R. N., Thota, H., Rao, A. A., & Gedela, S. (2009). Techniques for integrating -omics data. *Bioinformation*, 3(6), 284–286.

Andersson-Gunnerås, S., Mellerowicz, E. J., Love, J., Segerman, B., Ohmiya, Y., Coutinho, P. M., ... & Sundberg, B. (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(2), 144-165.

Apel, K., & Hirt, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu. Rev. Plant Biol.*, 55, 373-399.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25-29.

Babbie, E. R. (2009) *The Practice of Social Research*, 12th edition, Wadsworth Publishing, ISBN 0-495-59841-0, p. 426-433.

Baxter, A., Mittler, R., & Suzuki, N. (2014). ROS as key players in plant stress signalling. *Journal of Experimental Botany*, 65(5), 1229-1240.

Benkovic, S. J., Theriot, J., & Ringe, D. (2013). Open questions-in brief: Beyond-omics, missing motor proteins, and getting from molecules to organisms. *BMC biology*, 11(1), 8.

Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M., & Caudy, A. A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC genomics*, 10(1), 221.

Buza, T. J., McCarthy, F. M., Wang, N., Bridges, S. M., & Burgess, S. C. (2008). Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic acids research*, 36(2), e12-e12.

Bylesjö, M., Nilsson, R., Srivastava, V., Gronlund, A., Johansson, A. I., Jansson, S., ... & Trygg, J. (2008). Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *Journal of proteome research*, 8(1), 199-210.

Caraux, G., & Pinloche, S. (2005). PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*, 21(7), 1280-1281.

Casado-Vela, J., Sellés, S., & Bru Martínez, R. (2005). Proteomic approach to blossom-end rot in tomato fruits (Lycopersicon esculentum M.): Antioxidant enzymes and the pentose phosphate pathway. *Proteomics*, 5(10), 2488-2496.

Celedon, P. A. F., de Andrade, A., Meireles, K. G. X., da Cruz Gallo de Carvalho, M., Caldas, D. G. G., Moon, D. H., ... & Labate, C. A. (2007). Proteomic analysis of the cambial region in juvenile *Eucalyptus grandis* at three ages. *Proteomics*, 7(13), 2258-2274.

Chen, J., Chen, B., & Zhang, D. (2015). Transcript profiling of Populus tomentosa genes in normal, tension, and opposite wood by RNA-seq. *BMC genomics*, 16(1), 164.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1), 51-78.

Dharmawardhana, P., Brunner, A. M., & Strauss, S. H. (2010). Genome-wide transcriptome analysis of the transition from primary to secondary stem development in Populus trichocarpa. *BMC genomics*, 11(1), 150.

Diz, A. P., Martínez-Fernández, M., & Rolán-Alvarez, E. (2012). Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology*, 21(5), 1060-1080.

Doherty, M. K., Whitehead, C., McCormack, H., Gaskell, S. J., & Beynon, R. J. (2005). Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates. *Proteomics*, 5(2), 522-533.

Domon, B., & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science* 312(5771), 212-217.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., ... & Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome research*, 17(10), 1537-1545.

Dutt, M. J., & Lee, K. H. (2000). Proteomic analysis. *Current opinion in biotechnology*, 11(2), 176-179.

Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The annals of applied statistics*, 107-129.

Fernie, A. R., & Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement?. *Trends in Genetics*, 25(1), 39-48.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2), 155-171.

Fournier, M. L., Paulson, A., Pavelka, N., Mosley, A. L., Gaudenz, K., Bradford, W. D., ... & Washburn, M. P. (2010). Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for

Ggc1 in cellular sensitivity to rapamycin. *Molecular & Cellular Proteomics*, 9(2), 271-284.

Gandla, M. L., Derba-Maceluch, M., Liu, X., Gerber, L., Master, E. R., Mellerowicz, E. J., & Jönsson, L. J. (2015). Expression of a fungal glucuronoyl esterase in *Populus*: Effects on wood properties and saccharification efficiency. *Phytochemistry*, 112, 210-220.

García-Alcalde, F., García-López, F., Dopazo, J., & Conesa, A. (2011). Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 27(1), 137-139.

García-Leiro, A., Cerdán, M. E., & González-Siso, M. I. (2010). Proteomic analysis of the oxidative stress response in Kluyveromyces lactis and effect of glutathione reductase depletion. *Journal of proteome research*, 9(5), 2358-2376.

Gene Ontology Consortium. (2013). Gene Ontology annotations and resources. *Nucleic acids research*, 41(D1), D530-D535.

Gerber, L., Eliasson, M., Trygg, J., Moritz, T., & Sundberg, B. (2012). Multivariate curve resolution provides a high-throughput data processing pipeline for pyrolysis-gas chromatography/mass spectrometry. *Journal of Analytical and Applied Pyrolysis*, 95, 95-100.

Gerber, L., Zhang, B., Roach, M., Rende, U., Gorzsas, A., Kumar, M., ... & Sundberg, B. (2014). Deficient sucrose synthase activity in developing wood does not specifically affect cellulose biosynthesis, but causes an overall decrease in cell wall polymers. *New Phytologist*, 203(4), 1220-1230.

Gillis, J., & Pavlidis, P. (2012). 'Guilt by association' is the exception rather than the rule in gene networks. *PLoS Comput Biol*, 8(3), e1002444.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., ... & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(D1), D1178-D1186.

Guglielmino, N., Liberman, M., Catesson, A. M., Mareck, A., Prat, R., Mutaftschiev, S., & Goldberg, R. (1997). Pectin methylesterases from poplar cambium and inner bark: localization, properties and seasonal changes. *Planta*, 202(1), 70-75.

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, 19(3), 1720-1730.

Hellgren, J. M., Olofsson, K., & Sundberg, B. (2004). Patterns of auxin distribution during gravitational induction of reaction wood in poplar and pine. *Plant Physiology*, 135(1), 212-220.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., ... & Sandberg, G. (2001). A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences*, 98(25), 14732-14737.

Higashi, Y., & Saito, K. (2013). Network analysis for gene discovery in plant-specialized metabolism. *Plant, cell & environment*, 36(9), 1597-1606.

Higashi, Y., Hirai, M. Y., Fujiwara, T., Naito, S., Noji, M., & Saito, K. (2006). Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. *The Plant Journal*, 48(4), 557-571.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1), 1-13.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., ... & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518), 929-934.

Jansson, S., & Douglas, C. J. (2007). *Populus*: a model system for plant biology. *Annu. Rev. Plant Biol.*, 58, 435-45.8.

Joyce, A. R., & Palsson, B. Ø. (2006). The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3), 198-210.

Kaever, A., Landesfeind, M., Feussner, K., Morgenstern, B., Feussner, I., & Meinicke, P. (2014). Meta-analysis of pathway enrichment: combining independent and dependent omics data sets. *PloS one*, 9(2), e89.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, gkr988.

Kelleher, N. L. (2004). Peer reviewed: Top-down proteomics. *Analytical chemistry*, 76(11), 196A-203A.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2), e1002375.

Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C., & Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics*, 7(3), 307-328.

Kogenaru, S., Yan, Q., Guo, Y., & Wang, N. (2012). RNA-seq and microarray complement each other in transcriptome profiling. *BMC genomics*, 13(1), 629.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., ... & Steinhauser, D. (2005). GMD@ CSB. DB: the Golm metabolome database. *Bioinformatics*, 21(8), 1635-1638.

Krüger, A., Grüning, N. M., Wamelink, M. M., Kerick, M., Kirpy, A., Parkhomchuk, D., ... & Ralser, M. (2011). The pentose phosphate pathway is a metabolic redox sensor and regulates transcription during the antioxidant response. *Antioxidants & redox signaling*, 15(2), 311-324.

Kurepa, J., & Smalle, J. A. (2008). Structure, function and regulation of plant proteasomes. *Biochimie,* 90(2), 324-335.

Kuster, B., Schirle, M., Mallick, P., & Aebersold, R. (2005). Scoring proteomes with proteotypic peptide probes. *Nature Reviews Molecular Cell biology*, 6(7), 577-583.

Lastdrager, J., Hanson, J., & Smeekens, S. (2014). Sugar signals and the control of plant growth and development. *Journal of experimental botany*, 65(3), 799-807.

Law, A. H. Y., Chow, C. M., & Jiang, L. (2012). Secretory carrier membrane proteins. *Protoplasma*, 249(2), 269-283.

Lin, Y. C., Li, W., Chen, H., Li, Q., Sun, Y. H., Shi, R., ... & Chiang, V. L. (2014). A simple improved-throughput xylem protoplast system for studying wood formation. *Nature protocols*, 9(9), 2194-2205.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., ... & Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology*, 17(7), 676-682.

Lofstedt, T., Hoffman, D., & Trygg, J. (2013). Global, local and unique decompositions in OnPLS for multiblock data analysis. *Analytica Chimica Acta.*, 791,13-24.

Madala, N. E., Piater, L. A., Steenkamp, P. A., & Dubery, I. A. (2014). Multivariate statistical models of metabolomic data reveals different metabolite distribution patterns in isonitrosoacetophenone-elicited Nicotiana tabacum and Sorghum bicolor cells. *SpringerPlus*, 3(254), 10-1186.

Mahboubi A, Ratke C, Gorzsás A, Kumar M, Mellerowicz EJ, Niittylä T (2013) Aspen SUCROSE TRANSPORTER3 allocates carbon into wood fibers. *Plant Physiol* 163: 1729-40.

Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1), 34.

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3), 255-261.

Masuda, T., Tomita, M., & Ishihama, Y. (2008) Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J Proteome Res,* 7(2), 731-740.

Matsushika, A., Goshima, T., Fujii, T., Inoue, H., Sawayama, S., & Yano, S. (2012). Characterization of non-oxidative transaldolase and transketolase enzymes in the pentose phosphate pathway with regard to xylose utilization by recombinant Saccharomyces cerevisiae. *Enzyme and microbial technology*, 51(1), 16-25.

Mellerowicz, E. J., & Gorshkova, T. A. (2012). Tensional stress generation in gelatinous fibres: a review and possible mechanism based on cell-wall structure and composition. *Journal of experimental botany*, 63(2), 551-565.

Mellerowicz, E. J., & Sundberg, B. (2008). Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Current opinion in plant biology*, 11(3), 293-300.

Mishima, K., Fujiwara, T., Iki, T., Kuroda, K., Yamashita, K., Tamura, M., ... & Watanabe, A. (2014). Transcriptome sequencing and profiling of expressed genes in cambial zone and differentiating xylem of Japanese cedar (Cryptomeria japonica). *BMC genomics*, 15(1), 219.

Mittler, R. (2002). Oxidative stress, antioxidants and stress tolerance. *Trends in plant science*, 7(9), 405-410.

Mittler, R., Vanderauwera, S., Suzuki, N., Miller, G., Tognetti, V. B., Vandepoele, K., ... & Van Breusegem, F. (2011). ROS signaling: the new wave?. *Trends in plant science*, 16(6), 300-309.

Morgan, M. J., & Liu, Z. G. (2011). Crosstalk of reactive oxygen species and NF-κB signaling. *Cell research*, 21(1), 103-115.

Myburg, A. A., Lev-Yadun, S., & Sederoff, R. R. (2013). Xylem structure and function. *eLS.* John Wiley & Sons, Ltd, Chichester, UK.

Nagalakshmi, U., Waern, K., & Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology*, 4-11.

Nicholson, J. K., & Lindon, J. C. (2008). Systems biology: metabonomics. *Nature*, 455(7216), 1054-1056.

Niculaes, C., Morreel, K., Kim, H., Lu, F., McKee, L. S., Ivens, B., ... & Boerjan, W. (2014). Phenylcoumaran benzylic ether reductase prevents accumulation of compounds formed under oxidative conditions in poplar xylem. *The Plant Cell*, 26(9), 3775-3791.

Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic acids research*, gks804.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., ... & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic acids research*, 42(D1), D26-D31.

Oliver, S. (2000). Proteomics: guilt-by-association goes global. *Nature*, 403(6770), 601-603.

Patel, R. V., Nahal, H. K., Breit, R., & Provart, N. J. (2012). BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *The Plant Journal*, 71(6), 1038-1050.

Plomion, C., Leprovost, G., & Stokes, A. (2001). Wood formation in trees. *Plant physiology*, 127(4), 1513-1523.

Polge, C., Jaquinod, M., Holzer, F., Bourguignon, J., Walling, L., & Brouquisse, R. (2009). Evidence for the existence in Arabidopsis thaliana of the proteasome proteolytic pathway activation in response to cadmium. *Journal of Biological Chemistry*, 284(51), 35412-35424.

Porth, I., Klápště, J., Skyba, O., Friedmann, M. C., Hannemann, J., Ehlting, J., ... & Douglas, C. J. (2013). Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions. *New Phytologist*, 200(3), 727-742.

Ralser, M., Wamelink, M. M., Kowald, A., Gerisch, B., Heeren, G., Struys, E. A., ... & Krobitsch, S. (2007). Dynamic rerouting of the carbohydrate flux is key to counteracting oxidative stress. *J Biol*, 6(10), 301-312.

Rantalainen, M., Cloarec, O., Ebbels, T.M., Lundstedt, T., Nicholson, J.K., Holmes, E., & Trygg, J. (2008). Piecewise multivariate modelling of sequential metabolic profiling data. *BMC Bioinformatics*, 9(1),105.

Rood, S. B., Ball, D. J., Gill, K. M., Kaluthota, S., Letts, M. G., & Pearce, D. W. (2013). Hydrologic linkages between a climate oscillation, river flows, growth, and wood Δ13C of male and female cottonwood trees. *Plant, cell & environment*, 36(5), 984-993.

Roux, P. P., & Topisirovic, I. (2012). Regulation of mRNA translation by signaling pathways. *Cold Spring Harbor perspectives in biology*, 4(11), a012252.

Saito, K., Hirai, M. Y., & Yonekura-Sakakibara, K. (2008). Decoding genes with coexpression networks and metabolomics–'majority report by precogs'. *Trends in plant science*, 13(1), 36-43.

Scandalios, J. G. (2005). Oxidative stress: molecular perception and transduction of signals triggering antioxidant gene defenses. *Brazilian Journal of Medical and Biological Research*, 38(7), 995-1014.

Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M., & Sandberg, G. (2004). A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *The Plant Cell*, 16(9), 2278-2292.

Siedlecka, A., Wiklund, S., Péronne, M. A., Micheli, F., Leśniewska, J., Sethson, I., ... & Mellerowicz, E. J. (2008). Pectin methyl esterase inhibits intrusive and symplastic cell growth in developing wood cells of Populus. *Plant Physiology*, 146(2), 554-565.

Srivastava, V., Obudulu, O., Bygdell, J., Löfstedt, T., Rydén, P., Nilsson, R., ... & Wingsle, G. (2013). OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI-superoxide dismutase Populus plants. *BMC genomics*, 14(1), 893.

Steen, H., & Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology*, 5(9), 699-711.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545-15550.

Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6), 817-836.

Sweetlove, L. J., Obata, T., & Fernie, A. R. (2014). Systems analysis of metabolic phenotypes: what have we learnt?. *Trends in plant science*, 19(4), 222-230.

Taylor, G. (2002). *Populus*: Arabidopsis for Forestry. Do We Need a Model Tree? *Annals of Botany*, 90(6), 681–689.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., ... & Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6), 914-939.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics*, 16(3), 119-128.

Trygg, J., & Wold, S. (2003). O2-PLS, a two-block (X±Y) latent variable regression (LVR) method with an integral OSC® lter². *J. chemometrics*, 17, 53-64.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., ... & Henrissat, B. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793), 1596-1604.

Uggla, C., Mellerowicz, E. J., & Sundberg, B. (1998). Indole-3-acetic acid controls cambial growth in Scots pine by positional signaling. *Plant Physiology*, 117(1), 113-121.

Uggla, C., Moritz, T., Sandberg, G., & Sundberg, B. (1996). Auxin as a positional signal in pattern formation in plants. *Proceedings of the National Academy of Sciences*, 93(17), 9282-9286.

Van Acker, R., Vanholme, R., Storme, V., Mortimer, J. C., Dupree, P., & Boerjan, W. (2013). Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in Arabidopsis thaliana. *Biotechnol Biofuels*, 6(1), 46.

Vander Mijnsbrugge, K., Meyermans, H., Van Montagu, M., Bauw, G., & Boerjan, W. (2000). Wood formation in poplar: identification, characterization, and seasonal variation of xylem proteins. *Planta*, 210(4), 589-598.

Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLOS Biology*, 12(6).

Wang, G., Wu, W. W., Zeng, W., Chou, C. L., & Shen, R. F. (2006). Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. *Journal of proteome research*, 5(5), 1214-1223.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63.

Wang, Z., Wang, D., Zheng, S., Wu, L., Huang, L., & Chen, S. (2014). Ultra-performance liquid chromatography-quadrupole\ time-of-flight mass spectrometry with multivariate statistical analysis for exploring potential chemical markers to distinguish between raw and processed Rheum palmatum. *BMC complementary and alternative medicine*, 14(1), 302.

Weckwerth, W., Wenzel, K., & Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, 4(1), 78-83.

Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it?. *Chemometrics and Intelligent Laboratory Systems*, 30(1), 109-115.

Wu, Y., Williams, E. G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S. M., ... & Aebersold, R. (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell*, 158(6), 1415-1430.

Yang, D., Du, X., Yang, Z., Liang, Z., Guo, Z., & Liu, Y. (2014). Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. *Engineering in Life Sciences*, 14(5), 456-466.

Yuan, J. S., Galbraith, D. W., Dai, S. Y., Griffin, P., & Stewart, C. N. (2008). Plant systems biology comes of age. *Trends in plant science*, 13(4), 165-171.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, 9(1).

Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., ... & Wang, J. (2004). BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research*, 32(suppl 1), D377-D382.

Zhong, R., McCarthy, R. L., Lee, C., & Ye, Z. H. (2011). Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology*, 157(3), 1452-1468.

Zhu, J., Sova, P., Xu, Q., Dombek, K. M., Xu, E. Y., Vu, H., ... & Schadt, E. E. (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol*, 10(4), e1001301.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., & Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant physiology*, 136(1), 2621-2632.

7 Acknowledgement

I wish to express my heartfelt gratitude to special people who have touched me in many ways.

I am very grateful to all my supervisors. I deeply appreciate and feel highly privileged to have been tutored by such an experienced team. First I thank my main supervisor, Prof. Gunnar Wingsle for accepting me for as a PhD student and providing outstanding education, support and guidance throughout my PhD studies, thank you also for the boat ride during the summer, I have enjoyed working with you. Next, I thank my assistant supervisors: Prof. Thomas Moritz for training me in the field of metabolomics, and providing wise counsel and advice during preparation of the manuscripts and thesis; and Prof. Torgeir R. Hvidstein for expert advice, help with the transcriptomic analysis, proof-reading, practical advice and training. I am also highly grateful to my Chemometrics supervisor, Prof. Johan Trygg, for his excellent training and tips on modelling: going to his office to learn was a pleasant and educative experience. I am also grateful to Prof. Hannele Tuominen, an amazingly brilliant person: thanks for the financial support of my PhD program, wise counsel, inspirational advice, help with drafting manuscripts and educative meetings. I also wish to thank the various funding bodies listed in the papers and manuscripts for their financial support.

I wish to thank senior colleagues and Professors whose counsels have contributed to my success and taught me salient, practical lessons in the school of Science: Karin Ljung, Björn Sundberg, Solomon Tesfalidet, Göran Samuelsson, Stephanie Yvette Robert, Olivier Keech, Henrik Antti, Pär Jonsson, Ulrika Ganeteg, Uwe Sauer, Gerhard Gröbner, Knut Irgum, Mikael Elofsson, Leszek Kleczkowski, Urs Fischer, Ewa Mellerowicz, Rishikesh Bhalerao, András Gorzsás, Maria Rosario Garcia Gil, Åsa Strand, Anders Nordström, Paul Geladi, Totte Niittylä, Anders Fries, Edouard Pesquet, Anita Sellstedt, Harry Xiaming Wu, Christiane Funk and Ove Nilsson.

I wish to thank Maria Israelsson Nordström for being a remarkable, efficient and winsome personality, quick at helping PhD students. I wish to thank fellow BIOIMPROVE PhD graduate students (Bioimprove-Improved biomass and bioprocessing properties of wood- Swedish Research Council, Formas): Sacha Escamez, Melis Kucukoglu, Prashant Mohan Pawar, Henrik Serk and Amir and SweTree Technologies.

I wish to thank my office mates for being incredible companions: Thi Hai Hong Nguyen (especially for your Vietnamese food), Thomas Vain (especially for help with figures, inscape and office camaraderie), Maximiliano Kawahata Pagliarini, Siamsa Melina Doyle and Delphine Gendre.

Past and present members of the aspen wood development project Group, Umea Plant Science Centre; Next Gen seq. & Computational life Science cluster of Umea University / SLU (UPSC); Proteomics platform KBC, Umea University / Swedish University of Agricultural Sciences (SLU); and the Swedish Metabolomic Centre (SMC) are equally appreciated for all the discussions, tips and help they provided, including: Izabella Surowiec, Rui Pinto, Tomas skotare, Nabil Souihi, Sergiu Netotea, Rickard Sjögren, Tommy Löfstedt, David Sundell, Nicolas Delhomme, and Chanaka Mannapperuma. I wish to say thank you to Joakim Bygdell, Thomas Kieselbach, Robert Nilsson and Vaibhav Srivastava for help with the proteomic analysis and training: and to Shaochun Zhu, Irene Granlund, Miles Trupp, Malin Nording, Carolina Arias Gomez, Jehad Shaikhali, Enni Väisänen and Xiaoling Zhao who have also worked with or at the proteomics platform.

I wish to thank the following members of the Swedish Metabolomics Centre, Umea, for help with the metabolomic analysis: Ilka Nacif Abreu, Jonas Gullberg, Jenny Hällqvist, Annika Johansson, Maria Ahnlund, Jonas Gullberg, Inga-Britt Carlsson, Hans Stenlund, Roger Granbom, Krister Lundgren and Gun Lövdahl (thanks for the picture).

I wish to thank many at Umea University and SLU with whom I have enjoyed pleasant conversations during lunch breaks/ conferences etc., and some who have become my close friends: Kate Bernett, Pernilla linden, Jenny Hällqvist, Edward Kenneth Businge, Liudmila Angelcheva, Yogesh Mishra, Junko Takahashi Schmidt, Ioanna Antoniadi, Sonali Ranade, Kristina Frech Irgum, Natalia Bukhanko, Madhavi Latha Gandla, Judith Felten, Carlos Martín, Delphine Menard, Marta Derba-Maceluch, Thomas Dobrenel, Jerker Fick, Per-Anders Enquist, Hardy Craig Hall, Henrik Lars Svennerstam, Anne Honsel, Michal Karady, Daniela Elisa Liebsch, Pal Miskolczi, Emiko Murozuka, John Baison, Gloria Sancho, Helder Miranda, David Öhman, Ahmed Elhaj, Lorenz Gerber, Anna Petterle, Pieter Nibbering, Rubén Casanova-Sáez, Raphaël Julien Decou, Federica Brunoni, Jose Alfredo Zambrano R, Iftikhar Ahmad, Domenique André, Zhi-Qiang Chen, Ainhoa Calleja-Rodriguez, Daria Anna Chrobok, Huizu Vicki Guo, Haleh Hayatgheibi, Christoffer Johnsson, Mateusz Stanisław Majda, Anastasia Matrosova, Ingrid Helena Michelson, Jean Claude Nzayisenga, Stefano Papazian, Carlos Martín Medina, Umut Rende, Shashank Sane, Henrik Serk, Barbara Katarzyna Terebieniec, Bernard Andries Wessels, Christian Kiefer, Sara V. Petersson, Jakob Prestele, Kathryn Robinson (thanks for the household items at the start of my PhD), and Peter Immerzeel.

I wish to thank all members of the technical staff who have been immensely helpful: Veronica Bourquin, Gun Lövdahl, Roger Granbom, Thomas Hiltonen, Bengt Olsson, Jan Karlsson, Rosìe Forsgren and Ingela Sandström.

I wish to thank members of the administrative staffs for their splendid and outstanding work: Gun-Britt Långnabba, Ann Sehlstedt, Ulrika Qvist Wallmark, Inga-Britt Carlsson and Inga-Lis Johansson for always willing to go an extra mile to help. You all demonstrate strong affection, amiability, care and concern towards others.

I wish to thank the information technology unit for help with my computer and IT needs: Stefan Löfmark, Kjell Engström and Simon Birve.

I wish to thank other friends and their families: Samuel Ohiomokhare, Calvin Kouokam, Ikechukwu Okafo, Oluwatosin Ayodele, Paul Akowe, Anthony Ojo, Obinna Nkemjika, Tord Lennart Jakobsson, Lis-Britt Elaine Jakobsson and Carl Clarence Jakobsson, John Karl Filip Dahlstrom, Osei Ampomah, Samuel Awe, Stephen Famurewa, Opeoluwa S. Idowu, James Akinola, Ikenna R. Obi, Gbotemi Adediran, Douglas Joseph, Olusegun Ayodeji Oyewole, Olusola O. Alabi, Bright B. Kumordzi, Evans Korang Adjei, Shodeinde Peter Oladimeji, Joseph Mumba Zulu, and members of the DCLM charity organisation.

I wish to thank my family members and in-laws who support me in many ways for which I am ever grateful.

I wish to also specially thank others whose names I have not mentioned in this acknowledgement because of space limitations: I am equally grateful.

I am especially grateful to God for innumerable blessings.