METABOLOMIC ANALYSES OF WOOD ATTRIBUTES IN TREE SPECIES

by

ANDREW RAYMOND ROBINSON

B.Sc. (Hon.), Massey University, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Forestry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2009

Abstract

Metabolomics is an emerging field in functional plant biology that attempts to relate patterns in the molecular intermediates and products of metabolic pathways with genetic, gene expression, environmental and phenotypic traits - at the whole-tissue and/or whole-organism level. There is enormous potential for metabolomics tools to be applied in the study of tree species, and the demand for widespread application is promoting an ongoing evolution and refinement of newly-developed techniques. This body of research addresses the application of broad-scale, non-targeted metabolomics to questions of wood formation and quality in tree systems. Overall, it was shown that variation in metabolite profiles from developing xylem tissue was indeed correlated with the strength of specific phenotypic traits. Frequently, the strength of these relationships was such that phenotypic severity could be predicted accurately on the basis of metabolite profile data alone. The specific correlative patterns and metabolite/trait pairings observed in each study provided insight into the biological mechanisms by which these traits arise. Studies of secondary xylem development were conducted on breeding populations of Douglas-fir and radiata pine, as well as genetically modified hybrid poplar. In the Douglas-fir families studied, environment-induced variation in growth rate, fibre morphology and wood chemistry were correlated with metabolite profiles from developing xylem; metabolites involved in carbohydrate and lignin biosynthesis were primarily implicated in these relationships. Similarly, in juvenile trees from a series of radiata pine families, correlations were observed between metabolite profiles of developing xylem and the internal checking wood defect, a known heritable trait. In a different approach, two poplar hybrids, each modified separately with two exogenous gene constructs related to lignin biosynthesis, provided controlled model systems in which to investigate the interaction between genotype, metabolite profiles of developing xylem, and physico-chemical wood traits. Wood traits and metabolite profiles alike were altered by the genetic modifications, and it was found that the metabolic impact of the transgenes was not confined to pathways that were directly coupled to lignin biosynthesis. In fact, the scarcity of lignin-related metabolites in profiles from either the wild-type or modified genotypes suggested that metabolite channelling phenomena operate in the lignin biosynthetic pathway. Moreover, the

analyses demonstrated that transgene-induced gradients in phenotypic traits could be associated with similar gradients within broad-scale metabolite profiles, and also that the wood-forming metabolisms of different poplar hybrids can respond similarly to the influences of genetic manipulation, at a global level. To conclude, the demonstrated associations between genotype, the metabolism of wood formation, and wood phenotype, as revealed by metabolite profiles, confirm the value of non-targeted metabolomics as a systems biology approach to understanding and modeling growth and secondary cell wall biosynthesis in trees.

Table of Contents

List of Tables

List of Figures

Acknowledgements

This research project ended up being both larger and longer than everyone involved had hoped, and its final form is the product of many peoples' contributions over many years. My first thanks must go to my supervisor, Dr. Shawn Mansfield, who provided the vision and funding for this work, as well as his considerable time, dedication and expertise. In the same vein I'd also like to thank the other members of my committee, Drs. Brian Ellis, David Ellis, and John Kadla for their guidance, as well as Drs. Robert Kozak and Rick White for assistance in statistical analysis and Dr. Lacey Samuels for her microscopy work. David Kaplan and the staff of the UBC Horticulture greenhouse tended to my indoor forest season after season, and Drs. Peter Beets, Ken Wong, Johan Bosch, and Mike McConchie assisted in many ways during sample collection in Rotorua, New Zealand. Also many thanks to the hoards of Mansfield Lab members who helped with my research and particularly to my close collaborator Dr. Rebecca Dauwe, whose software coding abilities helped to make the analysis of my data so much better than would otherwise have been possible. A very special thank-you must be given to my good friends Tom Canam and Ian Cullis for mad humour, crazy missions and support both in and out of the lab. And lastly, thanks to my family, and most importantly my wife Kristan Sutherland, for always being there in support.

I would like to recognise the contribution of the musicians and vintage audio equipment manufacturers who provided the soundtrack to the recent and intensive assembly of this document. In particular, I would like to give credit to Sansui (amplifiers), Infinity (loudspeakers) and Technics (turntables), and make special mention of Curve's albums "Doppelgänger" and "Cuckoo", David Bowie's album "The rise and fall of Ziggy Stardust and the Spiders From Mars", and The KLF's timeless, classic, and multi-remixed singles "What time is love?" and "Last Train to Trancentral" – all of which were on painfully high rotation during this challenging time.

This work is dedicated to all designers and manufacturers of well-engineered, highly reliable equipment.

Co-authorship Statement

Chapter 2: Andrew Robinson was involved with performing research, data analysis and manuscript preparation. Shawn Mansfield was involved with identification and design of the research program and manuscript preparation. Nicholas Ukrainetz and Kyu-Young Kang were involved with performing the research and data analysis.

Chapter 3: Andrew Robinson was involved with the design of the research, performing research, data analysis and manuscript preparation. Shawn Mansfield, was involved with identification and design of the research program and manuscript preparation. A. Lacey Samuels was involved with performing the research. Nicholas Ukrainetz was involved with data analysis.

Chapter 4: Andrew Robinson was involved with the design of the research, performing research, data analysis and manuscript preparation. Shawn Mansfield was involved with identification and design of the research program and manuscript preparation. David Ellis was involved with the design of the research. Rana Gheneim was involved with performing the research. Robert Kozak was involved with data analysis.

Chapter 5: Andrew Robinson was involved with the identification and design of the research, performing research, data analysis and manuscript preparation. Shawn Mansfield was involved with identification and design of the research program and manuscript preparation. Rebecca Dauwe was involved with performing the research and data analysis.

CHAPTER 1

Introduction

1.1 Scenario

Human activity has long been associated with tree-derived materials, and wood in particular has utility in a plethora of roles. The specific chemical and physical properties of wood influence its suitability for particular end uses (*e.g.* pulp *vs.* lumber *vs.* biofuels feedstock). Ongoing demand to extend the limits of wood's applicability, both in terms of productivity and economics, has promoted the breeding of "elite" families and/or genotypes globally in a variety of species (both gymnosperms and angiosperms). Over the past century breeding and plantation-based forestry have brought about gains in forest productivity, although the long generation times of many tree species have made improvement through classical breeding techniques a relatively slow process, especially in conifers. Furthermore, selection criteria have typically focused on easily determined macroscopic properties, such as trunk form and growth rate (wood volume), with the effects of breeding on the physico-chemical properties of the wood, and trait stability under environmental variation receiving less attention. New technologies that enable the early growth stage assessment of physico-chemical wood traits in mature trees could have a profound effect on classical breeding and substantially improve the forest industry in its pursuit of both higher productivity and improved quality in trees.

The technologies of molecular biology have a significant role to play in the continued evolution of tree breeding for specific tree and wood properties. In theory, genomic sequences, gene expression, protein biosynthesis and metabolite fingerprints are all capable of providing markers for specific phenotypic traits that could then be incorporated into low-to-high throughput screening platforms. The development of genetic and gene expression-based markers is technically straightforward. The technology for the broad-scale analysis of DNA and mRNA is therefore fairly mature, when compared to similar platforms focussing on protein- and metabolism-based markers. Unfortunately, the complex 'regulatory space' between a physical phenotype and its molecular marker increases the likelihood of marker inaccuracy, and along the continuum from genes to phenotype, genetic and gene expression-based makers are the most distant from the final phenotype. Conversely, the regulatory space between phenotypic traits and markers based on metabolic features is much smaller, since metabolic changes reflect the cellular activity immediately preceding the emergence of a physical phenotype and presumably would integrate inputs from the upstream

genotype, gene and protein expression patterns, temporal and spatial subcellular localisation of gene products, and the influences of environmental and developmental factors. Markers for wood traits based on metabolite accumulation features therefore have the potential for greater accuracy than their upstream counterparts. Despite this potential, full exploitation of such markers has yet to be realised in plant biology, in contrast to medical biology, where urinalysis and blood analysis for disease diagnosis are commonplace.

The development and implementation of wood-associated metabolic marker systems for tree breeding has been slow because broadscale analyses of plant metabolism for the purpose of differentiating wood traits poses considerable technical challenges. Present technology does not permit routine measurement of flux through metabolic pathways *in vivo*, so the only viable option for creating a metabolite 'fingerprint' is to measure the abundance of the intermediates in biochemical pathways (metabolites) that exhibit pooling behaviour, and treat this profile as a surrogate indicator of metabolic state. Confounding this is the enormous range in abundance and variety of physical properties exhibited by different metabolites involved in the primary and secondary metabolism of wood formation, which ensure that, for a specific tissue, comprehensive profiles can only be generated by employing a combination of analytical techniques. In addition, custom designed software, extensive statistical analysis and substantial computing power are required to process metabolite profile data.

The greater vision of the research described in this thesis was to assemble a platform for generating and analysing metabolite profiles from tree species, and to demonstrate correlative associations between physico-chemical wood traits and features in metabolite profiles from developing xylem. Confirmation that such relationships can be routinely detected and characterised constitutes an initial phase in the development of trait screens and diagnostics for wood-related tree breeding based on metabolic markers. Secondly, in order to be viable in the setting of breeding for the forest industry, trait screens based on metabolic markers will not only have to be accurate, but also efficient and cost effective. Marker complexity will therefore have to be kept to a minimum so that the screening of many individuals can be accomplished in a cost and time-effective manner. As such, an important aspect of this work was to test statistical approaches for their ability to identify those metabolite signals that contribute

most strongly to correlative relationships with other phenotypic traits of interest (*e.g.* wood chemistry). The ability to generate refined profiles consisting of the fewest elements required to achieve accurate screening will be an essential part of any viable screening platform. In the course of this work, attempts were made to identify the metabolites detected in the profiles and to relate the identities of metabolites implicated in correlative relationships to their roles in the biological system, although this goal was considered secondary to the goal of demonstrating the potential utility of metabolic markers for early assessment of wood traits.

1.2 Metabolomics in plant species

1.2.1 The "metabolome" and "metabolomics"

In systems biology, the term "metabolome" refers to the complete set of small molecules (*i.e.* metabolites) that participate in, or are products of, metabolic reactions within an organism or tissue. Metabolomics is therefore primarily concerned with the identification and quantification of such molecules to better understand their biochemical fate in a given pathway or biological response, as well as aiding in the development of novel biomarkers. As an eventual product of gene expression under the influence of environment, cellular metabolism is the immediate progenitor of phenotype and, as such, the relationships between phenotypic and metabolomic traits are potentially less complicated than for the genomic, transcriptomic and proteomic counterparts. However, in comparison to the other "omics", for which rapid technological advances have been seen, the development of methodologies for the comprehensive analysis of the metabolome has been slow. Whereas the genome, transcriptome and proteome are each comprised of a single class of polymeric molecule, the metabolome exhibits an enormous degree of physico-chemical molecular variety such that no single current instrument platform is capable of analysing all metabolites. The consequent need to employ a series of preparative and analytical techniques to (imperfectly) span the metabolome, and the technical difficulty of merging disparate and/or overlapping data generated by these diverse means have restricted the potential of metabolomics to date. However, the significant technological and methodological advances of the past decade's research are now finally being reflected in the abundance of metabolomics studies reported in the plant biology literature.

Metabolomics tools now available finally have enough practicality and biological resolution to encourage widespread implementation of this technology.

Metabolomics analyses have been broadly classified as either "targeted" or "non-targeted". Targeted analysis, otherwise known as 'metabolite profiling', typically focuses on quantifying a defined group of metabolites that are related by either a metabolic pathway or a molecule class. These studies tend towards a higher degree of *a priori* knowledge as far as compound identity and interrelationship are concerned, and in their most refined form become 'target analysis' - the measurement of one or very few metabolites to serve as, for example, phenotypic biomarkers. Conversely, non-targeted analysis aims to measure as broad a range of metabolites as possible, with the intention of creating a global metabolic fingerprint. In the first instance, global fingerprinting is not as concerned with the metabolites' identity and absolute abundance, as it is with their relative abundance and interrelationships, and aims primarily to classify samples based on metabolic 'features'. Ultimately, though, the reductive approaches commonly employed in these analyses usually lead to the identification of subsets of discriminating metabolites whose abundances correlate with specific treatments or phenotypic traits of interest. Subsequently, attempts can be made to identify those compounds so that their biological significance may be rationalised. Whereas broad-scale metabolomics is a recent development of the last ten or so years, targeted analysis of metabolism has a much longer history. Although, due to the narrow focus, it is arguable that targeted analyses are not metabolomics in the strict sense, they do comprise the origin from which non-targeted, global metabolomics approaches have been derived with the assistance of advancing technology. As such, there is obvious interdependency between targeted metabolite profiling and non-targeted metabolic fingerprinting, which has arisen through their shared ultimate objectives: improved biological understanding and diagnostic capabilities. Because this conceptual bridge exists, recent metabolomics research in plants has frequently fallen into a middle ground, in terms of the degree of prior knowledge of the identity and role of the metabolites being analysed, the breadth of metabolites being analysed and the basis for their inclusion. Clearly, the scale and rationality of analyses do not allow a practical distinction between modern metabolomics and historical metabolic analyses to be made. In reality, metabolomics is

defined by a new working environment – one in which powerful new analytical tools, abundant computing power and powerful data-handling software have made it conceivable to tackle metabolic issues at the whole organism or tissue level, with an emphasis on deconvoluting biological complexity.

1.2.2 The analytical process

Practical metabolomics is concerned with measuring and analysing metabolite pools in an attempt to understand metabolic networks and develop biological markers. An expanding range of analytical and software tools are available to assist in this endeavour. In many cases, the rate of flux through metabolic pathways would be a more robust and informative measure of metabolic activity. However, the current limitations of metabolite flux analysis make its broad-scale implementation largely impractical, and dictate the use of the more easily measured metabolite pooling phenomenon as a somewhat ambiguous indicator of metabolic activity.

1.2.2.1 Sample preparation.

The source material for samples should be relevant to the research objectives (e.g. developing xylem tissue is a good substrate for analysis of xylem biosynthetic metabolism), and may be comprised of whole plants (practical only for small species such as *Arabidopsis*), plant fluids (e.g. xylem or phloem sap), compounds released in gas exchange (e.g. volatile terpenoids), individual plant organs (e.g. root, leaf, stem or inflorescence), and now even laser capture microdissected cell groups (Schad *et al.*, 2005). While some of the analytical tools employed in metabolomics permit the determination of metabolite composition with minimal sample preparation (e.g. nuclear magnetic resonance spectroscopy; NMR), others require active extraction of metabolites from the specific tissue, prior to analysis. This process typically involves -80°C freezing or freeze-drying of samples and tissue disruption (Fiehn *et al.*, 2000a; Shepherd *et al.*, 2007), followed by some form of liquid solvent extraction, and, when required, further solvent partitioning of the crude extracts.

Although metabolite extractions based on single solvents (e.g. methanol or chloroform) are applicable, the composition of the extract obtained will exhibit bias toward metabolites that are highly soluble in the chosen solvent, which may be either desirable or undesirable in particular analyses. Because of this, sample preparation for completely non-targeted chromatography-based metabolomics has frequently employed

multi-solvent extractions, typically including water (very polar) and at least one other less-polar solvent. Variations of the extraction and derivatisation protocols for *Arabidopsis* published by Fiehn *et al.* (2000a; 2000b) are often employed. The extraction is based on a dual-phase water/methanol/chloroform extraction that yields polar metabolites in the water/methanol phase and less-polar metabolites in the methanol/chloroform phase. A method for single-phase extraction with these same solvents, combined in ratios that do not lead to phase separation, has also been established (Gullberg *et al.*, 2004). In situations where specific metabolite classes are being targeted (e.g. phenolics), selective extraction and subsequent metabolite partitioning and enrichment can be used to refine samples to achieve better resolution and signal-to-noise ratios for the target metabolites. Such a method involving methanol extraction, followed by lyophilisation and subsequent partitioning of the metabolites between water and cyclohexane, was employed for the concentration of phenolic metabolites in an aqueous phase (Damiani *et al.*, 2005). Another example of an extraction protocol specifically tailored to a subject metabolite class is that employed for the specific extraction of membrane phospholipids developed for *Arabidopsis* (Welti *et al.*, 2002; Yang *et al.*, 2007). Isopropanol with butylated hydroxytoluene (BHT) is used as the primary solvent, with various mixtures of chloroform, water and methanol with BHT used for subsequent, exhaustive tissue extraction. Combined extracts are washed with KCl solution then purified with water, and finally dried down prior to re-suspension in chloroform or a chloroform/methanol mixture.

When samples are to be analysed by gas chromatography (GC), it is common to derivatise the metabolites post-extraction as a way of increasing volatility and therefore the high mass cut-off of the analysis. In the classic approach to metabolite sample preparation (Fiehn *et al.*, 2000a; Fiehn *et al.*, 2000b), this involves the protection of carbonyl moieties by reaction with an alkoxy-oxyamine hydrochloride, followed by the elimination of acidic protons by reaction with a trimethylsilylating agent (e.g. N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA)). Where appropriate, it is also recommended that a methanol/chloroform-based trans-methylation of hydrocarbon chains be carried out prior to the other derivatisation reactions.

The documented optimisation of conditions for metabolite extraction and derivatisation in *Arabidopsis* leaves, stems and cell cultures (Gullberg *et al.*, 2004;

t'Kindt *et al.*, 2008), developing xylem of loblolly pine (Morris *et al.*, 2004) and potato tubers (Shepherd *et al.*, 2007) clearly illustrate the importance of ensuring that the process has enough stringency to achieve good metabolite extraction, but is not harsh enough to cause degradation of labile compounds. The susceptibility of the metabolite profile to variations in sample handling and analytical conditions is a known limitation of metabolomics, which demands consistent processing in order for comparable datasets to be generated from individual samples or sample batches.

1.2.2.2 Tools for measuring metabolites

A variety of analytical tools are available for the generation of metabolite profiles or fingerprints, with specific tools being more appropriate for the determination of metabolites having particular physico/chemical properties. In this regard, the analysis of the plant metabolome requires no special consideration over that of other organisms, with chromatography, mass spectrometry and NMR spectroscopy being the analytical mainstays across the field.

Gas chromatography (GC) is the chromatographic technique of choice for the analysis of smaller (MW < ~1000) molecules, owing to its applicability to a broad range of molecular classes and high resolution. Furthermore, the recent emergence of ultra-fast gas chromatography offers a significant increase in sample processing efficiency that promises to assist the development of very high-throughput metabolomics. The usual approach to sample introduction for GC is the evaporation of liquid phase extracts in the injector, although other techniques such as headspace extraction can be effective in specific scenarios and may avoid the need for lengthy sample preparation (Kjalstrand *et al.*, 1998; Wang *et al.*, 2006). Alternatively, high pressure liquid chromatography (HPLC) is useful for the separation of molecules too large or too labile for GC. Furthermore, the advent of ultra-high pressure liquid chromatography (U-HPLC) has facilitated the much-needed increases in the resolution of liquid chromatography for metabolomics (Grata *et al.*, 2008). Although the range of metabolites that may be analysed by liquid chromatography is frequently limited to a specific polarity range in the "middle ground" (Roepenack-Lahaye *et al.*, 2004), variant approaches including capillary-based and hydrophilic interaction chromatography (HILC) can be used to broaden this specificity (Roepenack-Lahaye *et al.*, 2004; Tolstikov and Fiehn, 2002; Tolstikov *et al.*, 2003). Capillary electrophoresis is another emerging liquid-based

separation technology with potential applications in plant metabolomics (Soga *et al.*, 2003).

Chromatographic separation systems require an attached quantitative detector, and mass spectrometers have achieved widespread popularity in metabolomics research. Quadrupoles, ion traps, Fourier transform (FT) and time-of-flight (TOF) mass analysers, combined with various sample introduction methods appropriate for the respective spectrometer and the preceding gas or liquid chromatography technique, have all been applied in various settings. The mass spectral data generated can be used to deconvolute signals from co-eluting metabolites, effectively increasing the resolution of the chromatographic analysis (discussed below), and provide extensive molecular structural information. Also popular are photodiode array (PDA) systems, which may be implemented as detectors for liquid chromatography – either alone or in combination with a mass spectrometer. PDAs measure light absorption across the ultraviolet and visible wavelengths, generating characteristic spectra for responsive analytes such as aromatics.

NMR spectroscopy is a popular alternative to chromatography/mass spectrometry for resolving compounds in metabolomics analyses (Charlton *et al.*, 2004; Ott *et al.*, 2003; Ratcliffe and Shachar-Hill, 2001; Sanchez *et al.*, 2008; Terskikh *et al.*, 2005). A major benefit of NMR spectroscopy is that it is non-destructive, meaning that samples (even living) may be analysed repeatedly over the course of an experiment, or studied by alternative approaches once NMR analysis is complete. Biological NMR spectroscopy usually exploits the magnetic properties of ^{1}H or ^{13}C nuclei, but may also target ^{31}P or ^{15}N (Bligny and Douce, 2001). The different nuclei in a molecule resonate at slightly different frequencies as a result of differences in local chemical environment, so particular compounds have characteristic nuclear resonance patterns for specific nuclei. Thus, an NMR spectrum provides information on the number and type of atomic nuclei, for example, ^{1}H nuclei in a mixture of metabolites, and it is possible to resolve the contribution of individual molecules to the spectra generated by complex metabolite mixtures. While this technique has been applied directly to intact plant tissues, or crude metabolite preparations, it has also seen some application as a detection tool in HPLC-based analyses (Wolfender *et al.*, 2003).

The provision of structural information by the detection system is crucial to the success of metabolomics analyses. Without compound identification, all that can be provided by the analysis of metabolite extracts is a metabolic fingerprint, and while potentially useful for distinguishing between distinct metabolic systems, a fingerprint alone is not at all informative about underlying biological relationships. This fact has fuelled the popularity of photodiode array-, NMR-, and mass spectrometry-based detection in metabolomics, as the molecule- or molecular class-specific spectral patterns generated by these techniques can facilitate the identification of metabolites, via matches with the signatures of standard compounds. Spectral matches, combined with matches for retention times or indices (in analyses that include chromatographic separation), can provide a high level of certainty for positive identifications. Furthermore, with appropriate spectrometers, elemental composition calculations, soft chemical ionisation techniques or MS_n analysis can also contribute to the identification of compounds in the absence of verified standards (Fiehn *et al.*, 2000b; Tolstikov and Fiehn, 2002).

Obviously, libraries of the spectral and retention index data for biological molecules are of enormous use when attempting to identify metabolite compounds, and extensive libraries are available for NMR and GC/(EI)MS. Although it is possible to assemble such libraries for LC/MS, the high degree of instrument- and eluent-dependent variation in analyte fragmentation patterns has meant that these libraries are not universally compatible. For GC/MS, however, inert carrier gases and the standard use of a 70eV potential for electron ionisation (EI) and molecule fragmentation have meant that not only can libraries be constructed, but they can also be shared between instruments and research groups. This has led to the publishing of extensive commercial and freely distributed libraries of EI mass spectra. While commercial libraries represent an extremely broad range of molecules (e.g. the 2008 NIST library contains more than 190,000 compounds including various states of derivatisation), smaller, free libraries such as those provided by the Gölm Metabolite Database (GMD) (Kopka *et al.*, 2005) are tailored specifically to the needs of plant metabolomics. These libraries are less redundant and frequently have more utility. Despite the growing number of available libraries, many compounds resolved from metabolite continue to elude identification. The ongoing expansion of mass spectral library resources is of

paramount importance, because the process of compound identification constitutes a major limiting factor in the plant metabolomics field.

1.2.2.3 Data processing and analysis.

In chromatograms of complex biological samples the partial or complete co-elution of metabolites is a frequent occurrence that, if not addressed, can limit biological resolution and introduce error into downstream data analyses. Additionally, the collation of metabolite profile data from multiple samples is required prior to statistical analysis, but manual collation becomes impractical in chromatography-based analyses involving large sets of samples and/or metabolites. This is because unavoidable fluctuations in temperature ramps, eluent gradients, column pressure or flow rates lead to inter-sample variation in the metabolite separation domain (which is time-based in most high-resolution chromatography), ensuring that the retention time of any given metabolite is seldom, if ever, a single exact value across all sample runs. Fortunately, however, the pressing need to resolve these issues has led to the development of algorithms that are able to deconvolute the signals from co-eluting metabolites. Both commercial and free software tools that semi-automate these tasks have emerged, with notable non-commercial offerings including NIST AMDIS (for deconvolution only), MSFACTs (Duran *et al.*, 2003), metAlign (Tikunov *et al.*, 2005; Tolstikov *et al.*, 2003), correlation optimised warping (COW) (Christin *et al.*, 2008; Nielsen *et al.*, 1998), and the highly capable XCMS (Smith *et al.*, 2006).

Data analysis in metabolomics has advanced at a considerable rate, with ongoing introduction of statistical analyses and other calculative tools to the field. Most statistical tools have been applied from exploratory or reductive perspectives. Classic, univariate tests between means, such as Student's t-test, the F-test and more robust incarnations like Tukey's "Honestly Significant Difference" (HSD) test have been used to individually identify metabolites exhibiting genotype- or treatment-related differences in abundance (Fiehn *et al.*, 2000a; Yeh *et al.*, 2006). Although useful, these tests deal with each metabolite as an isolated entity, and are unable to take the interdependence of the components of metabolite profiles into account (*i.e.* the "network" paradigm). Multivariate analyses are better suited to this task. The multivariate tools initially adopted in metabolomics were principal components analysis (PCA) (Chen *et al.*, 2003;

Fiehn *et al.*, 2000a) and hierarchical cluster analysis (HCA) (Roessner *et al.*, 2001a; Roessner *et al.*, 2001b), and the scope of many classic and contemporary analyses is limited to these two techniques. Both are useful for comparing complete profiles from multiple samples, and generate diagrammatic outputs that are visually appealing and easily interpreted. Although PCA does provide some information regarding the particular metabolites responsible for any distinction between sample classes, neither PCA nor HCA are very diagnostic, as they are unable to provide calculated measures of the relationships between metabolite profiles and, for example, phenotypic traits. Canonical correlation analysis (CCA) is one method that can better assist in defining the relationships between two sets of variables, such as metabolites and quantitative phenotypic traits (Meyer *et al.*, 2007). Essentially, CCA identifies groups of variables in one set that are correlated to groups of variables in the other, and indicates the relative contributions of individual variables to the relationship. However, in cases where diagnostics are an objective, techniques that generate models for the prediction of specific traits on the basis of metabolite profiles typically have more utility. To this end, multiple discriminant analysis (MDA) is useful for distinguishing samples by class (e.g. genotype, species), while partial least squares regression (PLSR) (Dijksterhuis *et al.*, 2005; Meyer *et al.*, 2007) and less conventional "stepwise" variable selection procedures (Klukas *et al.*, 2006; Li and Nyholt, 2001; Yamashita *et al.*, 2007) are powerful techniques for modeling quantitative phenotypic traits (e.g. the total lignin content of wood, or any other measurable property).

The graphical presentation of biochemical pathways and molecular interactions, as supported by metabolomic data, is an important part of metabolomics, and can contribute considerably to data interpretation and the derived understanding of biological relationships at the molecular level. Neural networking, as conducted by the "Pajek" software (Batagelj and Mrvar, 2002), is a graphical correlative statistical approach capable of effectively summarising interactive networks, which uses marker size and proximity to visualise the interactions within sets of variables. In metabolomics, the networks generated by this process provide valuable insight into the interdependency between specific metabolites, which can reveal hubs or metabolic control points within the systems being analysed (Batagelj and Mrvar, 2002; Fiehn, 2003; Giuliani *et al.*, 2004; Steuer *et al.*, 2003). Another option is to record the

behaviour of various metabolites on conceptual metabolic pathway scaffolds established using previous research. Obviously, this could be done manually, but scaffolding and annotation software such as MapMan (Thimm *et al.*, 2004) can expedite the process. These scaffolds are annotated with the contributions of interesting metabolites to a given relationship with, for example, a phenotypic trait. These may be represented by numerical scores (Hirai *et al.*, 2004) or colour-coding and heatmap output markers (Nikiforova *et al.*, 2005b).

1.2.3 The effective incorporation of metabolomics into systems biology

From the turn of the century, at the time when the first reports of broad-scale metabolite profiling were made, it has been suggested that metabolomics would evolve into a powerful, integrated branch of plant systems biology (Fernie *et al.*, 2004; Fiehn *et al.*, 2001; Weckwerth, 2003). Unfortunately, the technical demands of metabolomics have dictated that this concept could not be realised in the very short term; however, it has become clear that metabolomic data has the greatest utility, and provides the deepest insight, only when interpreted in conjunction with its genomic, transcriptomic and proteomic counterparts. This realisation promotes increasingly complex experimental scenarios and the accompanying logistical challenges, as it demands concurrent analysis of sample sets by multiple "omics" platforms. It is reassuring then, to see that despite the inherent difficulties, "multi-omic" analyses, and the data processing tools required to conduct them (Bylesjo *et al.*, 2007; Daub *et al.*, 2003; Klukas *et al.*, 2006; van Riel, 2006; Wurtele *et al.*, 2003), are becoming more commonplace, and that data from one "omics" can assist the interpretation of that from others. To date, the staple diet has been the combination of metabolomics and transcriptomics analyses (Colebatch *et al.*, 2004; Hirai *et al.*, 2004; Nakamura *et al.*, 2007; Osuna *et al.*, 2007; Urbanczyk-Wochniak *et al.*, 2005), with some excellent examples of combined data being presented in either correlation network (Nikiforova *et al.*, 2005a) or pathway scaffold (Tohge *et al.*, 2005) formats. Combined metabolomics/genetics studies are emerging slowly (Lisec *et al.*, 2008; Morreel *et al.*, 2006).

1.3 Application of metabolomics technology in the study of plant species

Concurrent with the rise of metabolomics technology over the last decade, metabolomics analyses have been carried out on numerous plant species, involving a

broad range of analytical and data processing techniques. Frequently, though, analyses have been conducted simply to demonstrate the application of new or improved technologies in plant systems, without addressing defined biological issues. This notwithstanding, applied metabolomics is rapidly becoming an important, highly utilised research approach, and many examples of biology-based metabolomics have been published. Here, several areas pertinent to this research project will be discussed.

1.3.1 Development

One approach to understanding a developmental process is to track the behaviour of metabolites through its course, either as it occurs naturally or when perturbed by environmental conditions or specific genetic modifications. Metabolomics techniques allow broad observation of metabolism and any unexpected relationships therein, as associated with developmental processes. For example, tomato fruit development, as both a naturally occurring and genetically modified process, has been a target for metabolomics analyses. Such studies have helped in developing a comprehensive picture of the molecular biology (*i.e.* the interactions between gene expression, post-translational mechanisms and metabolic patterns) of tomato ripening (Carrari *et al.*, 2006), and to define temporal aspects of the role of hexokinase phosphorylation in that process (Roessner-Tunali *et al.*, 2003). Similarly, the extent to which a chromatin remodelling factor, PICKLE (PKL), was responsible for the metabolic transitions observed upon *Arabidopsis* seed germination and root formation was defined with the assistance of gene knockout-mutants and a metabolomic analysis (Rider *et al.*, 2004). In a "lipidomics" analysis following the progression of cellular development, apoptosis, and taxol biosynthesis in cell cultures of two *Taxus* species, Yang *et al*. (2007) found that phospholipid composition in apoptotic cells was markedly different than in living cells. This observation prompted the suggestion that the alternation of these membrane phospholipids plays a role in regulating the processes of apoptosis and taxol production in at least some *Taxus* species. The metabolic sink-to-source transition of developing quaking aspen (*Populus tremuloides*) leaves was followed in the work of Jeong *et al.* (2004), who observed clear distinctions between young, expanded and mature leaves. Through ontogeny, multi-fold changes in two-thirds of the identified metabolites were observed, with major trends seen in carbohydrate and amino acid metabolism that

conformed to the photosynthetic and respiratory shifts associated with a transition from carbon heterotrophy to carbon autotrophy, and from rapid synthesis to maturation of cell structure.

An alternative application of metabolomics in developmental biology involves its use in identifying markers for polygenic quantitative traits. Research indicates that it is possible to describe complex traits as a function of metabolic composition, and that genome-wide metabolic genomics analysis can aid in the search for polygenic traits with potential for improvement through breeding. For example, a combination of morphological analysis and metabolic and genetic quantitative trait loci (QTL) analyses were used to demonstrate the utility of a multi-omic approach in a tomato fruit breeding scenario (Schauer and Fernie, 2006), and in recombinant inbred lines of *Arabidopsis*, a strong, generally negative correlation between biomass and a specific set of (mostly) primary metabolites was defined (Meyer *et al.*, 2007).

1.3.2 Response to growth conditions

The ability to respond to environmental factors that challenge homeostatic equilibrium has far reaching consequences for plant health and survival, and productivity in the case of cultivated crops. Such abiotic pressure demands co-ordinated, system-wide adjustment in order for equilibrium to be maintained, and metabolomics technologies have become popular tools for investigating the biochemical mechanisms of this process. In fact, environmentally pressured systems continue to be key subjects in the development and application of tools for multi-omics analysis in plants.

1.3.2.1 Nutritional stress

In the study of nutrient deficiency stress by metabolomics and multi-omic analyses, sulphur was the first nutrient to attract attention, with this area seeing one of the first attempts to combine broad-scale metabolomics and transcriptomics (Hirai *et al.*, 2004). By demonstrating broad genomic and metabolic coherency, and very tight coherency for a branch of glucosinolate metabolism, this work set a precedent for multi-omic analyses in sulphur-related and other research. In an elegant implementation of dynamic networking, Nikiforova *et al*. (2005a) integrated complex transcript and metabolite data from *Arabidopsis* plants perturbed by sulphur depletion. The influence of this nutrient on the biological system was shown to act predominantly through modulation of particular genes' expression, which in turn affected metabolic

reorganisation. Specific gene expression and metabolic 'hubs' were identified, which appeared to control homeostasis with respect to sulphur nutrition, including apparent hormone-related regulatory networks. Parallel work (Nikiforova *et al.*, 2005b), found that the co-ordinated adaptive response of *Arabidopsis* to reduced sulphur availability, which involved decreases in sulphurous amino acid pools, total RNA, chlorophylls, proteins and plant biomass, was associated with a globally coordinated metabolic response involving shifts in elements relating to efficient sulphur assimilation, re-establishing nitrogen balance, and increases in lipid breakdown, purine metabolism and photorespiration.

Carbon- and phosphorus-based nutritional stress in *Arabidopsis* have been recent targets of metabolomic analyses, in conjunction with genomic scale transcript and protein analysis (Morcuende *et al.*, 2007; Osuna *et al.*, 2007). These analyses have provided broad insight into the nature of plant nutrition response mechanisms. As with sulphur, deprivation and resupply of these essential nutrients prompted coordinated, system-wide reprogramming of gene expression and subsequently central metabolism, with the response to resupply comprised of rapid and gradual components in both cases. The resupply of carbon led to rapid changes in the expression of transcription factors, and rapid re-accumulation of sucrose, reducing sugars and starch. More gradual recovery was apparent in transcripts, enzyme activities and metabolites involved in glycolysis and nitrate assimilation, the shikimate pathway and myo-inositol, proline and fatty acid metabolism. In the case of phosphorus, deprivation led to extensive shifts in gene expression and the accumulation of carbohydrates, organic acids and amino acids; resupply prompted a rapid recovery of gene expression related directly to phosphate processing/allocation and a fairly rapid reduction of amino acid pools, but much slower readjustment of the other metabolite classes.

1.3.2.2 Environmental pressure

Metabolomics studies of the response of plants to salt stress collectively suggest that the acclimation process involves multiple metabolic pathways, and is associated with changes in inorganic acid, amino acid and sugar metabolism. In grape vines exposure to salinity stress resulted in a reduction of sucrose and organic acid pools, but an increase in fructose, malic acid, and osmoprotectant amino acids (proline and asparagine) (Cramer *et al.*, 2007). Furthermore, mining of metabolite data from

Arabidopsis cell cultures suggests that the methylation cycle for the supply of methyl groups, the phenylpropanoid pathway, and glycine betaine biosynthesis were collectively induced in the short term response to salt stress, whereas the long term response (>24h) was characterised by an induction of glycolysis and sucrose metabolism, and a reduction of the methylation cycle (Kim *et al.*, 2007). In a comparison of the response of *Arabidopsis* and the related halophyte *Thellungiella halophila* to short term salt stress, a more extensive metabolic response in the halophyte was observed, with greater accumulation of myo-inositol, galactinol and raffinose, and greater reductions in pools of fumaric, malic, phosphoric and aspartic acids, compared to its glycophytic counterpart (Gong *et al.*, 2005). More interesting, however, was that prior to salt exposure, the steady state pools of many stress-responsive transcripts and metabolites were notably more abundant in the halophyte, which suggests the evolution of constitutive adaptation mechanisms in such species.

Research in grape vines indicates that the gene expression and metabolic responses of plants to high salt and drought may be based on similar foundations, but are specialised in order to meet the specific demands of each (Cramer *et al.*, 2007). In this, water deficit appears to be the more demanding state, with an analysis of the metabolite composition of maize xylem sap and *Arabidopsis* leaves under extended drought revealing both simple, and temporally more complex changes in separate sets of signalling and adaptive metabolites (Alvarez *et al.*, 2008; Rizhsky *et al.*, 2004). Some of the concentration changes in osmoprotectant metabolites were very substantial. In particular, a thirty-fold increase was seen in proline concentration in *Arabidopsis* leaves (Rizhsky *et al.*, 2004). Additionally, the very positive response of malic and abscisic acids in maize lent support to their putative role as root-to-shoot signals for systemic response to drought, while it was postulated that the observed pooling of monolignol precursors may relate to a reduction in lignin biosynthesis and stiffening of xylem cell walls as structural protection against tension induced buckling of vessels, and stem collapse (Alvarez *et al.*, 2008).

The response of plants to cold-temperature stress is a long-standing, highly active field of research, and several recent studies in *Arabidopsis* have defined a systems biology approach that includes metabolomic analyses. In this, metabolomics has shown that as with adaptation to high salt and drought, adaptation to temperature

stress involves extensive and complex reconfiguration of the metabolome. Non-acclimated cold or freezing tolerance appears to be under the positive control of the *CBF3* cold-responsive C-repeat/dehydration responsive element binding factor, and is more pronounced in cold-tolerant ecotypes that exhibit a higher level of constitutive activation and responsiveness in the CBF pathway, and tailoring of metabolome architecture (Cook *et al.*, 2004; Hannah *et al.*, 2006). However, the comparative metabolic stability of leaves developed at low temperatures compared to those shifted to low temperatures, and the commonly extensive, yet distinct metabolic characters generated by these two scenarios suggest that whereas some cold-related metabolic networks are modulated by the environment, development under low-temperature conditions invokes other constitutive network adjustments (Gray and Heath, 2005). Interestingly, the lack of correlation between related transcripts and metabolites in the course of cold acclimation suggest that regulatory factors other than transcript abundance play important roles in coordinating this process (Kaplan *et al.*, 2007).

With regard to heat-induced stress, non-targeted analyses have found that the metabolic response of *Arabidopsis* to heat shock shares the majority of its elements with the response to cold stress, but is much less intense (Kaplan *et al.*, 2007; Rizhsky *et al.*, 2004). Furthermore, the response to heat appears to be much less temporally complex than to cold, with most metabolic shifts in response to heat shock occurring quickly while cold response appears to pass through several phases (Kaplan *et al.*, 2007).

It has been found that a combination of stress types can prompt a metabolic response that is distinct from a combination of those prompted by each type alone. Amongst other effects, a combination of drought and heat stress can apparently prompt the replacement of one major osmoprotectant with another, presumably as a mechanism for avoiding metabolite cytotoxicity at high temperatures (Rizhsky *et al.*, 2004). Such mechanisms highlight the ability of the plant system to respond to complex environmental conditions that occur in nature.

1.3.3 Intra-species, and transgenic or non-transgenic line differentiation

Metabolomics technology is useful in the characterisation and distinction of different plant systems, while generating data of sufficient breadth as to allow consideration of the biological basis of such distinctions. Attempts to conduct metabolite profile-based

chemotaxonomy have yielded informative results for species in several plant genera, including *Eucalyptus* species of Australia (Merchant *et al.*, 2006), naturally occurring, environmentally marginal populations of *Arabidopsis lyrata* spp., and domesticated cultivars of Sesame (Laurentin *et al.*, 2008). Additionally, some of the seminal research of plant metabolomics has been concerned with defining the effects of genetic modification at the metabolic level, focussing on the effects of several transgenes related to sucrose metabolism in potato (Roessner *et al.*, 2001a; Roessner *et al.*, 2000; Roessner *et al.*, 2001b). As a continuation, differential network analysis of silent phenotype potato lines (Weckwerth *et al.*, 2004) highlighted the potential application of chemometric analysis in assigning function to genes for isozymes or members of gene families appearing to exhibit functional redundancy. The success of early research made it apparent that metabolomic analyses offered an opportunity to assess the effect(s) of genetic modification beyond overt phenotypic traits, and would be applicable to scenarios such as the determination of "substantial equivalence" and the extent of so-called "unintended effects" between transgenic and parental lines in food crop species. Although the potential of this application has long been discussed (Kuiper *et al.*, 2003), reports of food safety-related metabolomics research in plants are rare. Notable examples include the analysis of mutant and transgenic lines of tomato in which dietary antioxidants have been increased (Le Gall *et al.*, 2003; Long *et al.*, 2006), of wheat lines containing additional copies of endogenous genes encoding high-molecular-weight protein subunits of glutein (Baker *et al.*, 2006), and of transgenic maize lines harbouring the Cry1Ab gene for biosynthesis of Bt toxin (Levandi *et al.*, 2008). These studies indicate that genetic modification can result in significant, sometimes extensive changes in metabolism beyond the intended target pathway, although these changes may fall within the extent of variation seen for the parental line under environmental extremes. Metabolomics has also been applied with success to the analysis of gene misregulation related to phenylpropanoid metabolism, as discussed in the following section.

1.3.4 Secondary xylem biosynthesis

1.3.4.1 Lignin-related gene misregulation

Extensive studies focused on the effects of misregulating various genes associated with the phenylpropanoid pathway and involved in monolignol biosynthesis have been

conducted (Dauwe *et al.*, 2007; Leple *et al.*, 2007; Rohde *et al.*, 2004). In these analyses, metabolomics, and in particular, the co-application of metabolomics and transcriptomics, has yielded comprehensive maps of the effects of each misregulation that detail gene expression and metabolic responses to altered monolignol biosynthesis, and provide insight into the function and breadth of influence of each gene, and the plasticity of plant systems. This work will be discussed in detail. TDNA-insertion mutation-based inactivation of two isozymes of the first enzyme of the phenylpropanoid pathway, phenylalanine lyase (PAL1 and PAL2), resulted in extensive shifts in gene expression and metabolism in stems of *Arabidopsis* (Rohde *et al.*, 2004). The inactivation of either PAL1 or PAL2 caused increases in phenylalanine, tryptophan and glutamine-related metabolites involved in the recycling of ammonium via the GS-GOGAT cycle. The effects on gene expression were more extensive, with transcriptomic evidence suggesting a greater role for PAL1 in phenylpropanoid metabolism. However, the double mutation of these two isozymes was required for the emergence of a (minor) physical phenotype, and brought extended effects on the metabolome and wood composition. The elimination of both PAL1 and PAL2 greatly reduced flux through the phenylpropanoid pathway, evidenced by increased phenylalanine over-accumulation, shifts in several additional amino acids, reduced accumulation of flavonol glucosides, glycosylated vanillic acid, scopolin, two coniferyl alcohol-coupled feruloyl malates, and a reduction in total cell wall lignin content, with increased syringyl:guaiacyl monomer ratio.

Cinnamoyl-CoA reductase (CCR) catalyses the conversion of feruloyl-CoA to coniferaldehyde, in what is considered to be the first committed reaction step in the monolignol-specific branch of the phenylpropanoid pathway. In an analysis of CCR down-regulated poplar, the dramatic decrease in lignin content, and observed increase in the incorporation of ferulic acid into lignin with an approximate doubling of the ratio between ferulic acid or sinapic acid, and coniferaldehyde or sinapaldehyde, suggested that the down-regulation caused a shift in flux from monolignol biosynthesis toward ferulic acid (Leple *et al.*, 2007). LC/MS analysis revealed an increase in the production of the glucosylated phenolics, glucopyranosyl sinapic acid and glucopyranosyl vanillic acid, while GC/MS analysis identified twenty known metabolites that accumulated differentially due to CCR down-regulation, with strong representation from participants

in respiration, ascorbic acid, sugar (e.g. glucose, mannose and *myo*-inositol) and hemicellulose and pectin metabolism. Thus, it was confirmed that the misregulation had affected not only phenylpropanoid metabolism, but also various other pathways associated with primary metabolism and secondary cell wall biosynthesis. The transcriptomic and metabolomic data from this study, as well as another involving CCR-down-regulated tobacco (Dauwe *et al.*, 2007), indicated that a down-regulation of general carbohydrate metabolism and reduction and remodelling of hemicellulose and pectin glycans that cross-link lignin monomers took place in response to signals arising from the lignin-related changes in chemical and structural properties of the developing secondary wall. While some of these changes in carbohydrate metabolism could have been part of a stress response in the modified lines, the tobacco studies in particular indicated an emergence of a stressed state, with metabolite and transcript shifts suggesting increases in photo-oxidative stress and photorespiration (Dauwe *et al.*, 2007). Furthermore, the accumulation of glycosylated and quinylated derivatives of feruloyl-CoA, the usual substrate of CCR, suggests the existence of detoxification mechanisms that work to limit the accumulation of this metabolite, and may be the sink for carbon made available from the degradation of starch in a situation of reduced cell wall biosynthesis.

Further downstream in the monolignol biosynthetic pathway, cinnamyl alcohol dehydrogenase (CAD) catalyses the reduction of coniferaldehyde or sinapaldehyde into coniferyl and sinapyl alcohol, respectively. An “omics” analysis of the stems of tobacco down-regulated in the CAD2 enzyme defined a response with similarities to that seen with CCR down-regulation, but notably less extensive as far as carbohydrate metabolism was concerned (Dauwe *et al.*, 2007). Proximal to the activity of CAD2, an accumulation of its usual substrates, coniferaldehyde and sinapaldehyde, was observed. Although a respective decrease was not seen in the immediate enzymatic products, coniferyl alcohol and sinapyl alcohol, and lignin content remained stable, decreases were observed in the pools of 16 oligolignols (all consistent with those decreasing as a product of CCR down-regulation). The somewhat puzzling stability of lignin content despite CAD2 down-regulation may be explained by redundancy in this step of the pathway due to the existence of an isozyme, CAD1, which acts on coniferaldehyde and contributes significantly to the biosynthesis of coniferyl alcohol in

tobacco (Damiani *et al.*, 2005). Nevertheless, CAD2 down-regulation had a considerable positive effect on the pooling of quinic acid and conjugated phenolics, such as 1-caffeoyl quinic acid, vanillic acid glucoside, syringic acid glucoside, and sinapic acid glucoside, which are all putative by-products of upstream-of-CAD metabolite detoxification mechanisms.

1.3.4.2 Physico-chemical variation

In the study of xylem/wood formation, metabolomics has contributed to an improved understanding of the systemic rearrangements in cellular metabolism giving rise to wood with different physico-chemical properties, either within individuals or species. Morris *et al*. (2004) conducted a GC/MS-based metabolomic analysis of the developing xylem of loblolly pine trees, representing two families that produce wood with ~45% and ~50% alpha cellulose content. A set of the most abundant metabolites found in the GC/FID chromatogram were analysed by PCA, which loosely clustered and partially separated the samples of the two families. Both primary and secondary metabolites associated with wood formation were implicated in this distinction, including citric acid, shikimic acid, glucose and fructose. Although limited in terms of sample count and metabolic scope, this experiment set a precedent for subsequent, more comprehensive research. To support mounting chemical and structural evidence, and their hypothesis that juvenile and compression woods of conifers were not as similar as had previously been suggested, Yeh *et al*. (2006) attempted to distinguish between the metabolism involved in biosynthesis of variant wood forms in juvenile loblolly pine by profiling polar metabolites extracted from developing xylem. Tight clustering and clean separation of sample treatment groups in PCA and HCA analyses of a set of 25 highly and moderately abundant metabolites showed that the formation of normal, wind-exposed, compression, and opposite wood were each accompanied by distinct metabolite profiles. The profiles of juvenile and compression wood were clearly distinguished by PCA component 1, thus validating their claim. The separation of reaction wood (wind-exposed and compression) from non-reaction wood (normal and opposite) in PCA component 3 was due to increases in lignin precursors, such as shikimic acid, *p*-glucocoumaryl alcohol and coniferin, free sugars and sugar alcohols such as glucose, fructose, maltose, inositol and pinitol, and TCA cycle intermediates and amino acid-related metabolites including malic acid, gluconic acid and glycine. This profile was

consistent with the increase in lignin content and altered lignin composition typically seen in the compression wood of gymnosperms. The nature of the metabolism giving rise to reaction wood was further investigated in an analysis of the metabolic and gene expression profiles in developing tension wood of poplar (Andersson-Gunneras *et al.*, 2006). Although this work was dominated by transcript analysis, the multivariate analyses in the metabolomic component did reveal 26 metabolites that differed significantly between normal secondary cell wall and G-layer biosynthesis. Linoleic and oleic fatty acids were increased. Xylose and xylitol increased, whereas other sugars and sugar alcohols such as sucrose, arabinose and inositol decreased. Notably, the monolignol precursor shikimate was also decreased, as were other organic and amino acids including phosphate, citric acid, pentonic acid, aspartic acid, and galactaric acid. When viewed in conjunction with the extensive gene expression data, these metabolic shifts suggested the reprogramming of mechanisms for cellulose, lignin and cell wall matrix carbohydrate biosynthesis, amongst others. In particular, the apparent decrease in the activity of the pentose phosphate and shikimate pathways, and the concurrent increase in UDP-D-glucose biosynthesis were certainly in keeping with the decreased lignification and cellulose enrichment typically observed in the G-layer.

The examples provided demonstrate the effective use of metabolomics to rapidly identify the distinguishing components in different metabolic systems related to wood formation. However, it is apparent that a very promising aspect of metabolomics has yet to be exploited extensively. Not only can these types of analyses help to improve our understanding of the molecular mechanisms of wood biosynthesis, but there is certainly great potential to develop accurate metabolic markers for physico-chemical wood traits, and to apply those markers in trait monitoring and prediction scenarios.

1.4 The biology of secondary xylem biosynthesis

Plant cell walls are complex biological products comprised of a diverse array of compounds, which arise from a myriad of primary and secondary metabolic processes. However, for the sake of brevity this discussion of secondary xylem biosynthesis will be limited to the major carbohydrate and phenolic structural components.

1.4.1 Temporal and spatial aspects of secondary xylem formation

Wood, *i.e.* secondary xylem, arises from the vascular cambium (meristem) as part of secondary growth, which is a process whereby lateral meristematic activity allows stems to continue increasing in diameter in regions that are no longer elongating. Secondary xylem consists largely of cells that are no longer alive - specifically, the mature tracheids in softwoods (gymnosperms), and fibres, vessel elements and tracheids in hardwoods (angiosperms). However, in order to achieve their final, functional morphology, xylem cells must go through several developmental stages including origin, enlargement, secondary wall thickening and lignification.

New secondary xylem cells are produced through inward periclinal divisions of axially (vertically) orientated fusiform initial cells and their immediate derivatives (mother cells) in the vascular cambial zone. Following origin, an axially orientated cell enters a phase of elongation, or 'apical intrusive growth'. At this time, the cell has only a thin primary cell wall, consisting mainly of radially orientated cellulose microfibrils and cross-linking hemicellulose glycan. This wall expands vertically under the pressure of protoplast turgor, involving the vertically inclined, yet somewhat chaotic reorientation of microfibrils. At the same time, additional layers of microfibrils, called 'strata' are laid onto the inside of the primary wall, maintaining its thickness and preventing rupture. As regions of a cell stop growing, the primary wall is cross-linked into its ultimate shape.

Deposition of secondary cell walls begins once the cell shape is established. Since there is no clean seasonal separation between cell elongation and cell wall thickening, wall thickening works outwards from the middle of the cell to allow ongoing elongation at the ends. Outer (S1), middle (S2) and inner (S3) sub-walls are constructed from layers (lamellae) of microfibrils deposited on the inside of the existing primary wall in specific, ordered orientations. During this process hemicellulose and lignin are also deposited into the secondary cell wall matrix. Hemicellulose binds to cellulose, pectin and lignin to form a network of cross-linked fibres in the cell wall, establishing lateral rigidity in the process (Helm, 2000; Lawoko *et al.*, 2006; Popper and Fry, 2008; Uraki *et al.*, 2007). In axially orientated xylem, this process continues until all reserves in the vacuole and protoplast are consumed and metabolism ceases. Depending on ultimate function, different cell types undergo different degrees of thickening. For example, fibres provide strength and are almost a solid mass of walls, whereas vessel elements conduct fluid and retain a much larger hollow central core.

1.4.2 Cellulose

Cellulose is a biopolymer of unbranched β-1,4-linked glucan chains in which successive glucose residues are inverted 180° to achieve a flat ribbon-like structure, with the β-1,4-linked glucose dimer, cellobiose, as the repeating biosynthetic subunit (Koyama *et al.*, 1997). In higher plants these linear chains achieve lengths of up to 7000 – 15000 glucose residues (Brett, 2000; Brown, 2004). When arranged in parallel, these chains are able to form extensive hydrogen bond networks with one another. It is believed that ~36 chains are combined in a cylindrical array to form a cellulose microfibril – a highly crystalline structure that is a fundamental constituent of plant cell walls (Delmer and Haigler, 2002). The assembly of microfibrils from monomeric glucose residues is apparently conducted by cellulose synthase complex (CSC) "rosette" structures, which move across the plasma membrane as they extrude microfibrils into the cell wall (Herth, 1983). The CSC rosettes themselves are comprised of specific collections of cellulose synthase (CesA) subunit proteins, which are derived from multi-gene families and share a conserved structure (Arioli *et al.*, 1998; Holland *et al.*, 2000; Joshi *et al.*, 2004).

1.4.2.1 Biosynthesis of UDP-glucose

UDP-glucose is the proposed substrate of the CSC in plants (Delmer and Haigler, 2002). As such, a co-ordinated mechanism for the creation and regulation of UDP-glucose supply to the CSC should exist. Several enzymes have been implicated in this process primarily due to the positive correlation of their activities with the onset and progression of secondary cell wall biosynthesis, and also the cell wall-related effects of their misregulation. These enzymes include sucrose synthase (SuSy; sucrose + UDP ↔ UDP-glucose + fructose) (Robinson, 1996), sucrose phosphate synthase (SPS; UDP-glucose + fructose 6-phosphate ↔ UDP + sucrose 6-phosphate) (Haigler *et al.*, 2001; Park *et al.*, 2008), UDP-pyrophosphorylase (UGPase; UTP + glucose 1-phosphate ↔ UDP-glucose + PPi) (Carpita and Delmer, 1981; Coleman *et al.*, 2007; Wafler and Meier, 1994), sucrose phosphate phosphatase (SPP; sucrose 6-phosphate → sucrose + Pi) (Delmer and Haigler, 2002), and invertase (sucrose → glucose + fructose) (Canam *et al.*, 2008; Wafler and Meier, 1994). Although the exact nature of their interrelationships remains unclear, a putative model of these interactions has been proposed (Delmer and Haigler, 2002); the primary elements of this model will be outlined here.

A source of carbon is required to feed cellulose biosynthesis. Photosynthetic cells have the luxury of locally-generated pools of carbohydrates, while non-photosynthetic cells that form secondary walls (such as those in developing xylem) must derive their carbohydrate supply from transport sugars, such as sucrose. If the transport of sucrose across the plasma membrane is direct, then conversion by SuSy (via the reverse reaction) would be the most straightforward mechanism for generating UDP-glucose. Accordingly, there is some evidence that two forms of SuSy exist and that one form may associate directly with the CSC at the plasma membrane (Amor *et al.*, 1995; Robinson, 1996). If, however, the translocated molecule is something other than sucrose (e.g. raffinose), or if the mechanism by which the translocated dimer enters the cytosol is via cleavage into its monomeric constituents by an apoplastic invertase, then additional elements must be included in the biosynthetic model. In any case, even though cellulose biosynthesis is a strong sink, mechanisms that partition the translocated carbon between that process and other cellular processes must exist in order for the primary metabolic core to function, and elements of secondary metabolism to be maintained. Consequently, cytosolic invertase and/or soluble cytosolic SuSy must be involved in cleaving sucrose to create hexose pools even in the event that sucrose enters the cytosol directly. Utilisation of these free sugars would first require their phosphorylation into a pool of hexose phosphates, which is achieved by the action of hexokinases with the assistance of isomerases. From this hexose pool, glucose 1-phosphate may be converted to UDP-glucose by UGPase (Wafler and Meier, 1994). As well as a potential feedstock for the CSC (provided there is some mechanism by which the two can be associated), this UDP-glucose product can be converted back into sucrose either by the forward activity of SuSy, or via an alternative path involving the concerted activity of SPS and SPP (Delmer and Haigler, 2002; Haigler *et al.*, 2001). Indeed, it has been suggested that the primary mechanism by which cytosolic hexoses are made available to the CSC is via conversion into sucrose prior to processing by the membrane bound SuSy (Delmer and Haigler, 2002; Haigler *et al.*, 2001). The fact that these enzymes have shared substrates/products, and are generally capable of catalyzing both forward and reverse reactions, clearly suggests (1) a provision for cyclic metabolic processing within this system, and (2) the existence of regulatory mechanisms that balance flux through this cycle according to environmental factors,

developmental cues, cell fate, feedback inhibition, *etc*. Further work is necessary, however, to help determine the biochemical mechanisms and enzymatic interplay that underpin the provision of substrate for cellulose biosynthesis.

1.4.3 Hemicellulose

Hemicellulose is a heterogeneous glycan polymer that is derived from glucose, mannose, galactose, rhamnose, arabinose, and xylose. In contrast to cellulose, the polymer chains are branched, and achieve comparatively short lengths of 500 – 3000 glycan residues. The biosynthesis of hemicellulose requires glycan synthase and glycosyltransferase enzymes for polymer backbone and sidechain formation, respectively (Li *et al.*, 2006). Cellulose synthase-like proteins (CSLs) are also believed to be involved, and functional genomics approaches have begun to reveal gene families for these enzymes in *Arabidopsis* and poplar (reviewed by Li *et al.*, 2006; Mellerowicz and Sundberg, 2008), and more recently in loblolly pine (Nairn *et al.*, 2008).

In different plants the structure of hemicellulose varies in terms of sugar composition and linkage patterns. In dicots and many monocots the main hemicellulose of the primary wall is xyloglucan. In contrast, glucuronoxylan is the principal hemicellulose in dicot secondary cell walls, while glucomannan and others are minor contributors, notably in poplar (Mellerowicz *et al.*, 2001; Sjostrom, 1993; York and O'Neill, 2008). In gymnosperm species the hemicelluloses of secondary cell walls are mainly galactoglucomannans, as well as a small proportion of others, such as arabinoglucuronoxylan and arabinogalactan (Sjostrom, 1993). In terms of metabolomics analysis, the significance of this variability in hemicellulose composition is that in order for particular polymeric structures to be assembled, there must be a flux of carbon into activated monomer precursors of that structure. It might be expected that phosphorylated and UDP-conjugated forms of particular pentoses and hexoses would be generated to fill this need in specific species, with a prevalence of xylose and mannose related molecules in angiosperms, and galactose, mannose, arabinose and xylose related molecules in gymnosperms.

1.4.4 Lignin

Lignin is an aromatic heterobiopolymer formed primarily in the secondary xylem of vascular plants, as one of a wide variety of products of the phenylpropanoid pathway. It

is a principal structural component of woody tissue, and contributes significantly to vascular integrity and wood strength (Donaldson, 2001). Research has sought to understand the mechanisms by which lignin is formed in vascular plants, and can be summarised into three main areas: 1) the ultrastructure and topochemistry of lignin deposition (reviewed by Donaldson, 2001), 2) the identification and characterisation of genes, enzymes and regulatory elements involved in monolignol synthesis (reviewed by Anterola and Lewis, 2002; Dixon *et al.*, 2001; Humphreys and Chapple, 2002), and 3) the elucidation of the mechanisms by which lignin polymers are assembled from precursor monomer units (reviewed by Hatfield and Vermerris, 2001). More recently, a key review brought the results and models from all three areas together (Boerjan *et al.*, 2003).

The core aspects of lignin biosynthesis appear to be conserved within vascular plants, with the monomeric units of lignin being modified products of the phenylpropanoid pathway. The constituents of the lignin polymer in gymnosperms are primarily derived from *p*-coumaryl and coniferyl alcohols, whereas in angiosperms a third, sinapyl alcohol, is also involved (Lewis and Yamamoto, 1990). The lignin constituents derived from these three 'monolignols' are known as *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units, respectively, and combinations of these monomeric components are incorporated into lignin with species, tissue and developmental specificity (Donaldson, 2001). In addition to the three monolignols, other phenylpropanoids, such as hydroxycinnamyl aldehydes, acetates, *p*-hydroxybenzoates, *p*-coumarates and hydroxycinnamate esters are incorporated into the polymer (Ralph *et al.*, 2001).

1.4.4.1 Monolignol biosynthesis

Lignin biosynthesis has its origin in the shikimate pathway, which is the reaction series primarily responsible for linking carbohydrate metabolism to the biosynthesis of aromatic compounds in plants. The shikimate pathway consists of seven metabolic steps taking place in plastids, beginning with the condensation of erythrose 4-phosphate and phosphoenolpyruvate, and terminating with the synthesis of chorismate (precursor for phenylalanine, tyrosine and tryptophan) (Herrmann and Weaver, 1999). In photosynthetically active cells the erythrose 4-phosphate and phosphoenolpyruvate comes directly from photosynthesis in chloroplasts, via the pentose phosphate and

glycolysis pathways, respectively. Alternatively, in non-photosynthetic cells such as those found in developing xylem, these substrates arise from the breakdown of carbon source molecules delivered to the cell by source-sink translocation (e.g. carbohydrates such as sucrose) (Amthor, 2003). The initial conversion of transported carbohydrates into monomeric sugar phosphates occurs in the cytosol (the generation of the hexose phosphate pool is likely common to cellulose biosynthesis), with subsequent conversion of glucose 6-phosphate into erythrose 4-phosphate, the ensuing phosphoglycerates into phosphoenolpyruvate, and the subsequent reactions of the shikimate pathway occurring in non-photosynthetic plastids (Amthor, 2003). Presumably, the translocated sugars also act as substrates for the regeneration of S-adenosylmethionine (the methyl donor consumed in monolignol biosynthesis), and the production of the ATP and NADPH (via respiration) that are required for monolignol transport and subsequent polymerisation (Amthor, 2003).

The core phenylpropanoid pathway is common to the biosynthesis of a diverse range of phenolic compounds, notably the monolignols, coumarins, flavonoids, stilbenes and tannins. The reaction series begins with the conversion of phenylalanine to cinnamate, via a deamination of the side-chain catalyzed by phenylalanine ammonia-lyase (PAL). Subsequent conversion of cinnamate to *p*-coumarate is catalyzed by cinnamate 4-hydroxylase (C4H), which hydroxylates C4 of the benzene ring. Finally, the addition of co-enzyme A (CoA) to the acid-propane side-chain, by 4-coumaroyl CoA-ligase (4CL), yields an activated form of the molecule (Dixon and Paiva, 1995; Hahlbrock and Scheel, 1989; Holton and Cornish, 1995).

Until fairly recently, the model for the monolignol-specific phenylpropanoid pathway included a series of hydroxylation and O-methylation reactions on the aromatic ring, which converted cinnamate into a set of hydroxycinnamic acids (caffeate, ferulate, 5-hydroxyferulate and sinapate). *p*-Coumarate, ferulate and sinapate were then thought to be converted into monolignols via a series of reactions in which the side-chain carboxyl group was substituted with CoA, then an aldehyde, and finally a hydroxyl group to yield *p*-coumaryl, coniferyl and sinapyl alcohols, respectively (Freudenberg and Neish, 1968). However, with the identification of a set of enzymes capable of mediating the molecular conversions in this pathway, and the discovery that these enzyme are responsible for hydroxylation and methylation of hydroxycinnamic acids as well as the *in*

vitro identification of enzymes responsible for conversions at the CoA level, the pathway became represented by a 'metabolic grid' (Whetten and Sederoff, 1995).

The metabolic grid of monolignol biosynthesis was the product of *in vitro* enzyme analyses that involved single enzymes, substrates and products. In its entirety, however, this grid constituted an unlikely representation of a biological process in which a high degree of spatial and temporal regulation occurs. With new evidence from many sources, reviewers assessed the *in vitro* grid and scrutinized the original model (Anterola and Lewis, 2002; Dixon *et al.*, 2001; Humphreys and Chapple, 2002), and concluded that a number of the reactions and chemical intermediates they contained were unlikely to play significant roles in monolignol biosynthesis *in vivo*. The opinion, which continues to be favoured, was that monolignol biosynthesis does not involve substitutions of the aromatic ring at the level of hydroxycinnamic acids, and that contrary to the "grid" hypothesis, the pathway is more linear, with flux favouring certain spatially and energetically preferable reactions.

The current model proposes a conventional pathway, which represents the general trend, but which may not be entirely correct in particular situations or for specific species. Following the final reaction of the core phenylpropanoid pathway (conversion of *p*-coumarate to *p*-coumaroyl CoA by 4CL), *p*-coumaroyl CoA is converted into caffeoyl CoA via shikimate (primarily) and quinate ester intermediates. The substitution of CoA with a shikimate or quinate group is catalyzed by hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase (HCT) (Franke *et al.*, 2002; Hoffmann *et al.*, 2003; Nair *et al.*, 2002; Schoch *et al.*, 2001). This provides coumaroyl shikimate and quinate substrates for coumarate 3′-hydroxylase (C3′H), which generates caffeoyl shikimate and quinate by hydroxylation of the aromatic C3 (Schoch *et al.*, 2001; Ulbrich and Zenk, 1980). HCT is a 'reversible' acyltransferase, and as such also catalyses the resubstitution of shikimate/quinate for CoA to give caffeoyl CoA, thus creating the substrate of caffeoyl CoA O-methyl transferase (CCoAOMT), which methylates the hydroxyl group on the aromatic C3 to produce feruloyl CoA (Parvathi *et al.*, 2001; Ye, 1997). The CoA moiety of this intermediate would then be cleaved by cinnamoyl CoA reductase (CCR) to generate coniferaldehyde (Li *et al.*, 2005), which can then be converted to coniferyl alcohol by cinnamyl alcohol dehydrogenase (CAD) (Sibout *et al.*, 2005; Sibout *et al.*, 2003) and/or possibly sinapyl alcohol dehydrogenase

(SAD). Coniferyl alcohol is the precursor to guaiacyl lignin monomers. Coniferaldehyde and coniferyl alcohol are also likely intermediates in the biosynthesis of sinapyl alcohol, the precursor to syringyl lignin monomers in angiosperms. The aromatic C5 position of both molecules may be hydroxylated by ferulate 5-hydroxylase (F5H) (Humphreys *et al.*, 1999; Osakabe *et al.*, 1999), which yields a 5-hydroxylated form that can then be methylated by caffeic acid O-methyl transferase (COMT) (Humphreys *et al.*, 1999; Li *et al.*, 2000; Osakabe *et al.*, 1999; Parvathi *et al.*, 2001). When coniferaldehyde, the preferred substrate of F5H, is processed, the product of these reactions is sinapaldehyde, which is then converted to sinapyl alcohol by SAD (Li *et al.*, 2001) and/or CAD. Alternatively, when coniferyl alcohol is the initial substrate, sinapyl alcohol would be the direct product of COMT activity.

Several systems appear to co-regulate lignin monomer biosynthesis. Many of the genes encoding biosynthetic enzymes (notably PAL, 4CL, CAD and F5H of *Arabidopsis*) belong to multigene families, so specific isoforms may be expressed in different cell types, at different developmental stages, or in response to environmental conditions (Goujon *et al.*, 2003). This presumably affords a substantial degree of flexibility, allowing the pathway to vary around the constitutive backbone, and possibly incorporate other aspects of the metabolic grid as required. Transcription factors, specifically the R2R3 type MYB proteins, are implicated as regulators of gene expression for lignin biosynthetic enzymes. MYB proteins bind *cis*-acting AC elements, which are DNA motifs found in the promoter regions of many genes encoding lignin biosynthetic enzymes. Elevated expression of more than ten of these transcription factors has been associated specifically with developing xylem in *Arabidopsis* (Oh *et al.*, 2003), and a loblolly pine MYB (ptMYB1) has been shown to activate transcription from the PAL2 promoter (Patzlaff *et al.*, 2003). Gene expression for the monolignol pathway is also sensitive to the abundance of substrate and intermediate metabolites. In lignifying suspension cultures of loblolly pine, the transcriptional levels of PAL, 4CL, CCoAOMT, CCR and CAD are highly positively correlated to phenylalanine availability, while C4H and C3′H are largely stable (Anterola *et al.*, 2002). Cinnamate inhibits PAL at the transcriptional and post-translational levels, and possibly induces the activity of HCT (Anterola *et al.*, 2002).

An early hypothesis regarding the spatial organisation of monolignol biosynthesis was that the core phenylpropanoid pathway is tightly associated with the endomembrane network, whereas divergent pathways have only a loose association. This idea arose from sub-cellular fractionation studies during the late 1970's and 1980's, when the first discussions concerning the spatial organisation of phenylpropanoid metabolism took place (Czichi and Kindl, 1975; 1977; Hrazdina and Wagner, 1985; Hrazdina *et al.*, 1987). Recent studies have shown that although some PAL subunits are indeed associated with the lumen face of the endoplasmic reticulum (ER), others appear to be associated with the cytosol, Golgi-derived vesicles, or even with the lignifying secondary cell wall (Nakashima *et al.*, 1997; Smith *et al.*, 1994). Aside from this development, the most recent work continued to support the original hypothesis. C4H appears to be embedded in the ER membrane, and in French bean is particularly concentrated in the Golgi bodies (Smith *et al.*, 1994). 4CL, CCoAOMT, COMT, CCR and CAD appear to be mainly cytosolic in both monocots and dicots (Hrazdina and Wagner, 1985; Kersey *et al.*, 1999), although in cells of *Zinnia elegans*, Nakashima *et al*. (1997) also detected CAD in the Golgi vesicles and secondary walls.

Furthermore, the enzymes of monolignol biosynthesis undoubtedly participate in complex interactions with other enzymes and/or structural components, in order to bring the necessary efficiency to the biosynthetic process. Evidence suggests that the pathway is not simply comprised of a series of isolated single enzyme-assisted modifications that produce pools of pathway intermediates. Rather, the intermediates are covalently bound to, and passed between sequential active sites of multi-enzyme complexes, and as such no free pools of chemical intermediates are generated. This arrangement is referred to as 'metabolite channelling', and typically occurs where intermediates have no other cellular function except in a single biosynthetic pathway. It can be seen as a strategy for sparing cellular solvent capacity for the regulation and efficiency of the metabolic sequence, and also for the containment of molecules that have cytotoxic properties. Evidence of metabolite channelling exists for a multitude of metabolic pathways (Hrazdina *et al.*, 1987; Srere, 1987), now including monolignol biosynthesis (Anterola *et al.*, 1999; Rasmussen and Dixon, 1999; Winkel-Shirley, 1999). Anterola *et al*. (1999) supplied exogenous phenylalanine to cell suspension cultures of (gymnosperm) loblolly pine, and observed increases in the intracellular pools of

cinnamate and *p*-coumarate, as well as secreted pools of 4-courmaryl alcohol and coniferyl alcohol. There was no evidence of accumulation of any of the other predicted intermediates of the pathway. Additionally, when cells were fed with cinnamate, *p*-coumarate, caffeate or ferulate, these were not metabolised, but instead accumulated in the cells as glucosides – a conversion that may be part of compound detoxification. Together, these results suggest that the series of reactions between *p*-coumarate and ultimately monolignol synthesis are structured as a metabolic channel, and spatially distinct from any cytosolic pools that may exist. This has profound implications for the arrangement of the monolignol biosynthetic pathway, in that the existence of channels should provide another level of pathway control via specific ordering of sequentially acting enzymes, or orchestrated inclusion/exclusion of specific isozymes in order to achieve set biosynthetic outcomes.

1.4.4.2 Lignin polymerisation

Following their synthesis, lignin precursor monomers (monolignols) are transported to the cell wall where they are oxidised and polymerised. Monolignol transport remains one of the most poorly defined aspects of lignin biosynthesis, especially in angiosperms. It has long been held that 4-O-β-D-glucosides of the monomers are used for storage and/or transportation of these relatively toxic and unstable compounds. Genes encoding several UDPG-glycosyl transferases (UGTs) capable of catalyzing the transfer of glucose from UDP glucose to the phenolic hydroxyl group of *p*-coumaryl, coniferyl and sinapyl alcohols have been isolated from pine and *Arabidopsis* (Lim *et al.*, 2001; Steeves *et al.*, 2001). Similarly, genes encoding β-glycosidases that are able to cleave the glucose residue prior to polymerisation have been identified in pine (Dharmawardhana *et al.*, 1995; 1999; Leinhos *et al.*, 1994), and the enzymes localised to the secondary walls of lignifying cells (Samuels *et al.*, 2002). Although large pools of coniferin (glycosylated coniferyl alcohol) are readily detectable in gymnosperms, pools of similar size have been detected in only some angiosperms (for example, Magnolia species) (reviewed by Whetten and Sederoff, 1995). It is therefore speculated that in gymnosperms coniferin is held in the vacuole prior to being transported to the apoplast, either in Golgi-derived vesicles or by direct plasma membrane pumping by specific transporters (Samuels *et al.*, 2002). However, aside from the confirmed existence of *Arabidopsis* glycosyl transferases capable of generating sinapyl alcohol-4-O-glucoside

(Lim *et al.*, 2001; Steeves *et al.*, 2001), there is no clear indication of the corresponding angiosperm mechanism.

After transport of the monolignols to the cell wall, lignin is formed through dehydrogenative polymerisation of the monolignols. Although not proven outright, it is generally agreed that monolignols cleaned of any transport/storage related carbohydrate residues freely diffuse though the wall matrix, until they encounter cell wall-bound laccases or peroxidases and hydrogen peroxide, which generate radicals at the phenolic 4-OH position (Boerjan *et al.*, 2003). The best-supported mechanism for polymerisation of these radicals is known as the 'random coupling' model, which was reviewed effectively by Hatfield and Vermerris (2001). In this model, lignin arises primarily from the stepwise addition of monolignol radicals to the continually expanding polymer. This process is controlled by the diffusion of monolignols through the cell wall matrix itself; therefore the type and quantity of monolignols at the lignification site determine lignin composition. While another model for polymerisation, known as the "dirigent protein" model, has been proposed (Burlat *et al.*, 2001; Davin and Lewis, 2000; Gang *et al.*, 1999), to date there is no compelling evidence that dirigent proteins play roles in either the initiation or control of lignin polymerisation. In any case, it would appear that the structure of lignin is such that the random coupling model adequately explains its formation, while the dirigent protein model is improbable. There appears to be no requirement for a protein-assisted coupling mechanism, which would be exceedingly elaborate. In order to cover the range of bond types between the three monolignols and account for the lack of optical specificity, it has been estimated that approximately 100 different dirigent proteins with unique activities would be required (Hatfield and Vermerris, 2001).

1.5 Goals and hypotheses

1.5.1 A metabolomics platform for wood biology

The overarching technical goal of this research project was to establish a platform for the effective metabolomics analysis of wood properties in tree species. As described, such a platform is comprised of a series of elements, related to sample collection, preparation and analysis, and subsequent data processing, statistical analysis and the presentation of results. To this end, the intention was to employ liquid solvent-based

extraction protocols and analytical technologies such as GC/MS and LC/MS to generate broad-scale metabolite profiles from developing xylem tissue. The data generated were handled in a non-targeted manner; they were collated using semi-automated computer software, and subsequently analysed in conjunction with relevant genetic and phenotypic data via uni- and multi-variate statistical approaches. The requirement and development of this metabolomics infrastructure as part of this research should be apparent in the experiments described herein.

1.5.2 Metabolomics analysis of wood traits in industrially cultivated tree species

The goal of these experiments (Chapters 2 and 3) was to define relationships between the metabolite profiles of developing xylem tissue and physico-chemical wood traits in industrially relevant tree species. The subjects of this research, specific cultivated populations of *Pseudotsuga menziesii* (Douglas-fir) and *Pinus radiata*, were studied in isolation. The Douglas-fir population included a series of high-performance, full-sib families replicated on environmentally distinct sites, while the radiata pine population included a series of lines exhibiting varying severity in a value-limiting (internal checking), heritable wood phenotype. It was postulated that the heritable and/or environmentally influenced variation observed in wood traits would correlate with variable elements in the metabolite profiles of the developing xylem tissue from which the wood arises.

1.5.3 Metabolomics analysis of wood traits in genetically modified hybrid poplar

The goal of these experiments (Chapters 4 and 5) was to investigate the interaction between metabolite profiles of developing xylem and phenotypic wood traits, as influenced by genetic modification and genetic background. This involved an analysis and comparison between transgenic lines of two distinct poplar hybrids (*Populus grandidentata* × *alba* and *Populus tremula* × *alba*) harbouring the same wood-altering genetic construct. The influences of two transgenes were studied: the *Arabidopsis thaliana* ferulate 5-hydroxylase (F5H) under the control of the *Arabidopsis thaliana* cinnamate 4-hydroxylase promoter (C4H), and a hairpin-loop for RNAi suppression targeting *p*-coumaroyl-CoA 3′-hydroxylase (C3′H) under the control of the tobacco mosaic virus 35S promoter. The postulation was that transformation with wood composition-altering genetic constructs would induce detectable and equivalent

metabolic shifts in related, yet distinct genetic backgrounds. Furthermore, it was proposed that linear, predictive relationships would exist between elements of the metabolite profile and the severity of construct-induced phenotypic disturbance.

1.6 References

Alvarez, S., Marsh, E.L., Schroeder, S.G., & Schachtman, D.P. (2008). Metabolomic and proteomic changes in the xylem sap of maize under drought. *Plant Cell Environ.* 31**,** 325-340.

Amor, Y., Haigler, C.H., Johnson, S., Wainscott, M., & Delmer, D.P. (1995). A membrane-associated form of sucrose synthase and its potential role in synthesis of cellulose and callose in plants. *Proc. Natl. Acad. Sci. U. S. A.* 92**,** 9353-9357.

Amthor, J.S. (2003). Efficiency of lignin biosynthesis: A quantitative analysis. *Ann. Bot. (London)* 91**,** 673-695.

Andersson-Gunneras, S., Mellerowicz, E.J., Love, J.*, et al.* (2006). Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J.* 45**,** 144-165.

Anterola, A.M., van Rensburg, H., van Heerden, P.S., Davin, L.B., & Lewis, N.G. (1999). Multi-site modulation of flux during monolignol formation in loblolly pine (Pinus taeda). *Biochem. Biophys. Res. Commun.* 261**,** 652-657.

Anterola, A.M., Jeon, J.H., Davin, L.B., & Lewis, N.G. (2002). Transcriptional control of monolignol biosynthesis in Pinus taeda. Factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. *J. Biol. Chem.* 277**,** 18272-18280.

Anterola, A.M. & Lewis, N.G. (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* 61**,** 221-294.

Arioli, T., Peng, L.C., Betzner, A.S.*, et al.* (1998). Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science* 279**,** 717-720.

Baker, J.M., Hawkins, N.D., Ward, J.L.*, et al.* (2006). A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnol. J.* 4**,** 381-392.

Batagelj, V. & Mrvar, A. (2002). Pajek - Analysis and visualization of large networks. in Di Battista, G., Eades, P., Tamassia, R., & Tollis, I.G. (Eds), Graph Drawing. pp. 477-478.

Bligny, R. & Douce, R. (2001). NMR and plant metabolism. *Curr. Opin. Plant Biol.* 4**,** 191-196.

Boerjan, W., Ralph, J., & Baucher, M. (2003). Lignin biosynthesis. *Annu. Rev. Plant Biol.* 54**,** 519-546.

Brett, C.T. (2000). Cellulose microfibrils in plants: Biosynthesis, deposition, and integration into the cell wall. in Brett, C.T. & Waldron, K.W. (Eds), International Review of Cytology - a Survey of Cell Biology, Vol 199. Chapman & Hall, London, pp. 161-199.

Brown, R.M. (2004). Cellulose structure and biosynthesis: What is in store for the 21st century? *J. Polym. Sci., Part A: Polym. Chem.* 42**,** 487-495.

Burlat, V., Kwon, M., Davin, L.B., & Lewis, N.G. (2001). Dirigent proteins and dirigent sites in lignifying tissues. *Phytochemistry* 57**,** 883-897.

Bylesjo, M., Eriksson, D., Kusano, M., Moritz, T., & Trygg, J. (2007). Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J.* 52**,** 1181-1191.

Canam, T., Mak, S.W.Y., & Mansfield, S.D. (2008). Spatial and temporal expression profiling of cell-wall invertase genes during early development in hybrid poplar. *Tree Physiol.* 28**,** 1059-1067.

Carpita, N.C. & Delmer, D.P. (1981). Concentration and metabolic turnover of UDP-glucose in developing cotton fibers. *J. Biol. Chem.* 256**,** 308-315.

Carrari, F., Baxter, C., Usadel, B.*, et al.* (2006). Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiol.* 142**,** 1380-1396.

Charlton, A., Allnutt, T., Holmes, S.*, et al.* (2004). NMR profiling of transgenic peas. *Plant Biotechnol. J.* 2**,** 27-35.

Chen, F., Duran, A.L., Blount, J.W., Sumner, L.W., & Dixon, R.A. (2003). Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* 64**,** 1013-1021.

Christin, C., Smilde, A.K., Hoefsloot, H.C.J., Suits, F., Bischoff, R., & Horvatovich, P.L. (2008). Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* 80**,** 7012-7021.

Colebatch, G., Desbrosses, G., Ott, T.*, et al.* (2004). Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in Lotus japonicus. *Plant J.* 39**,** 487-512.

Coleman, H.D., Canam, T., Kang, K.Y., Ellis, D.D., & Mansfield, S.D. (2007). Over-expression of UDP-gluclose pyrophosphorylase in hybrid poplar affects carbon allocation. *J. Exp. Bot.* 58**,** 4257-4268.

Cook, D., Fowler, S., Fiehn, O., & Thomashow, M.F. (2004). A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 101**,** 15243-15248.

Cramer, G.R., Ergul, A., Grimplet, J.*, et al.* (2007). Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Funct. Integr. Genomics.* 7**,** 111-134.

Czichi, U. & Kindl, H. (1975). Formation of p-coumaric acid and O-coumaric acid from L-phenyl alanine by microsomal membrane-fractions from potato - evidence of membrane bound enzyme complexes. *Planta (Heidelberg)* 125**,** 115-125.

Czichi, U. & Kindl, H. (1977). Phenylalanine ammonia-lyase and cinnamic acid hydroxylases as assembled consecutive enzymes on microsomal-membranes of cucumber cotyledons - cooperation and sub cellular-distribution. *Planta (Heidelberg)* 134**,** 133-143.

Damiani, I., Morreel, K., Danoun, S.*, et al.* (2005). Metabolite profiling reveals a role for atypical cinnamyl alcohol dehydrogenase CAD1 in the synthesis of coniferyl alcohol in tobacco xylem. *Plant Mol. Biol.* 59**,** 753-769.

Daub, C.O., Kloska, S., & Selbig, J. (2003). MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics* 19**,** 2332-2333.

Dauwe, R., Morreel, K., Goeminne, G.*, et al.* (2007). Molecular phenotyping of lignin-modified tobacco reveals associated changes in cell-wall metabolism, primary metabolism, stress metabolism and photorespiration. *Plant J.* 52**,** 263-285.

Davin, L.B. & Lewis, N.G. (2000). Dirigent proteins and dirigent sites explain the mystery of specificity of radical precursor coupling in lignan and lignin biosynthesis. *Plant Physiol. (Rockville)* 123**,** 453-461.

Delmer, D.P. & Haigler, C.H. (2002). The regulation of metabolic flux to cellulose, a major sink for carbon in plants. *Metab. Eng.* 4**,** 22-28.

Dharmawardhana, D.P., Ellis, B.E., & Carlson, J.E. (1995). A beta-glucosidase from lodgepole pine xylem specific for the lignin precursor coniferin. *Plant Physiol.* 107**,** 331-339.

Dharmawardhana, D.P., Ellis, B.E., & Carlson, J.E. (1999). cDNA cloning and heterologous expression of coniferin beta-glucosidase. *Plant Mol. Biol.* 40**,** 365-372.

Dijksterhuis, G., Martens, H., & Martens, M. (2005). Combined Procrustes analysis and PLSR for internal and external mapping of data from multiple sources. *Comput. Stat. Data Anal.* 48**,** 47-62.

Dixon, R.A. & Paiva, N.L. (1995). Stress-induced phenylpropanoid metabolism. *Plant Cell* 7**,** 1085-1097.

Dixon, R.A., Chen, F., Guo, D., & Parvathi, K. (2001). The biosynthesis of monolignols: A "metabolic grid", or independent pathways to guaiacyl and syringyl units? *Phytochemistry* 57**,** 1069-1084.

Donaldson, L.A. (2001). Lignification and lignin topochemistry: An ultrastructural view. *Phytochemistry* 57**,** 859-873.

Duran, A.L., Yang, J., Wang, L.J., & Sumner, L.W. (2003). Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* 19**,** 2283-2293.

Fernie, A.R., Trethewey, R.N., Krotzky, A.J., & Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* 5**,** 763-769.

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., & Willmitzer, L. (2000a). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18**,** 1157-1161.

Fiehn, O., Kopka, J., Trethewey, R.N., & Willmitzer, L. (2000b). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72**,** 3573-3580.

Fiehn, O., Kloska, S., & Altmann, T. (2001). Integrated studies on plant biology using multiparallel techniques. *Curr. Opin. Biotechnol.* 12**,** 82-86.

Fiehn, O. (2003). Metabolic networks of Cucurbita maxima phloem. *Phytochemistry* 62**,** 875-886.

Franke, R., Humphreys, J.M., Hemm, M.R*., et al.* (2002). The Arabidopsis REF8 gene encodes the 3-hydroxylase of phenylpropanoid metabolism. *Plant J.* 30**,** 33-45.

Freudenberg, K. & Neish, A.C. Eds (1968). Constitution and biosynthesis of lignin: Molecular biology, biochemistry and biophysics. Vol.2 Springer-Verlag, New York.

Gang, D.R., Costa, M.A., Fujita, M*., et al.* (1999). Regiochemical control of monolignol radical coupling: A new paradigm for lignin and lignan biosynthesis. *Chem. Biol.* 6**,** 143-151.

Giuliani, A., Zbilut, J.P., Conti, F., Manetti, C., & Miccheli, A. (2004). Invariant features of metabolic networks: a data analysis application on scaling properties of biochemical pathways. *Physica A* 337**,** 157-170.

Gong, Q.Q., Li, P.H., Ma, S.S., Rupassara, S.I., & Bohnert, H.J. (2005). Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *Plant J.* 44**,** 826-839.

Goujon, T., Sibout, R., Eudes, A., MacKay, J., & Joulanin, L. (2003). Genes involved in the biosynthesis of lignin precursors in Arabidopsis thaliana. *Plant Physiol. Biochem.* 41**,** 677-687.

Grata, E., Boccard, J., Guillarme, D.*, et al.* (2008). UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in Arabidopsis thaliana. *J. Chromatogr. B* 871**,** 261-270.

Gray, G.R. & Heath, D. (2005). A global reorganization of the metabolome in Arabidopsis during cold acclimation is revealed by metabolic fingerprinting. *Physiol. Plantarum.* 124**,** 236-248.

Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M., & Moritz, T. (2004). Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* 331**,** 283-295.

Hahlbrock, K. & Scheel, D. (1989). Physiology and molecular-biology of phenylpropanoid metabolism. *Annu. Rev. Plant Phys.* 40**,** 347-369.

Haigler, C.H., Ivanova-Datcheva, M., Hogan, P.S.*, et al.* (2001). Carbon partitioning to cellulose synthesis. *Plant Mol. Biol.* 47**,** 29-51.

Hannah, M.A., Wiese, D., Freund, S., Fiehn, O., Heyer, A.G., & Hincha, D.K. (2006). Natural genetic variation of freezing tolerance in Arabidopsis. *Plant Physiol.* 142**,** 98-112.

Hatfield, R. & Vermerris, W. (2001). Lignin formation in plants. The dilemma of linkage specificity. *Plant Physiol.* 126**,** 1351-1357.

Helm, R.F. (2000). Lignin-polysaccharide interactions in woody plants. in Glasser, W.G., Northey, R.A., & Schultz, T.P. (Eds), Lignin : historical, biological, and materials perspectives. pp. 161-171.

Herrmann, K.M. & Weaver, L.M. (1999). The shikimate pathway. *Annu. Rev. Plant Phys.* 50**,** 473-503.

Herth, W. (1983). Arrays of plasma-membrane rosettes involved in cellulose microfibril formation of spirogyra. *Planta* 159**,** 347-356.

Hirai, M.Y., Yano, M., Goodenowe, D.B.*, et al.* (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 101**,** 10205-10210.

Hoffmann, L., Maury, S., Martz, F., Geoffroy, P., & Legrand, M. (2003). Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. *J. Biol. Chem.* 278**,** 95-103.

Holland, N., Holland, D., Helentjaris, T., Dhugga, K.S., Xoconostle-Cazares, B., & Delmer, D.P. (2000). A comparative analysis of the plant cellulose synthase (CesA) gene family. *Plant Physiol.* 123**,** 1313-1323.

Holton, T.A. & Cornish, E.C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* 7**,** 1071-1083.

Hrazdina, G. & Wagner, G.J. (1985). Metabolic pathways as enzyme complexes - evidence for the synthesis of phenylpropanoids and flavonoids on membrane-associated enzyme complexes. *Arch. Biochem. Biophys.* 237**,** 88-100.

Hrazdina, G., Zobel, A.M., & Hoch, H.C. (1987). Biochemical, immunological and immunocytochemical evidence for the association of chalcone synthase with endoplasmic-reticulum membranes. *Proc. Natl. Acad. Sci. U. S. A.* 84**,** 8966-8970.

Humphreys, J.M., Hemm, M.R., & Chapple, C. (1999). New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. *Proc. Natl. Acad. Sci. U. S. A.* 96**,** 10045-10050.

Humphreys, J.M. & Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5**,** 224-229.

Jeong, M.L., Jiang, H.Y., Chen, H.S., Tsai, C.J., & Harding, S.A. (2004). Metabolic profiling of the sink-to-source transition in developing leaves of quaking aspen. *Plant Physiol.* 136**,** 3364-3375.

Joshi, C.P., Bhandari, S., Ranjan, P.*, et al.* (2004). Genomics of cellulose biosynthesis in poplars. *New Phytol.* 164**,** 53-61.

Kaplan, F., Kopka, J., Sung, D.Y.*, et al.* (2007). Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant J.* 50**,** 967-981.

Kersey, R., Inoue, K., Schubert, K.R., & Dixon, R.A. (1999). Immunolocalization of two lignin O-methyltransferases in stems of alfalfa (Medicago sativa L.). *Protoplasma* 209**,** 46-57.

Kim, J.K., Bamba, T., Harada, K., Fukusaki, E., & Kobayashi, A. (2007). Time-course metabolic profiling in Arabidopsis thaliana cell cultures after salt stress treatment. *J. Exp. Bot.* 58**,** 415-424.

Kjalstrand, J., Ramnas, O., & Petersson, G. (1998). Gas chromatographic and mass spectrometric analysis of 36 lignin-related methoxyphenols from uncontrolled combustion of wood. *Journal of Chromatography. A.* 824**,** 205-210.

Klukas, C., Junker, B.H., & Schreiber, F. (2006). The VANTED software system for transcriptomics, proteomics and metabolomics analysis. *J. Pestic. Sci.* 31**,** 289-292.

Kopka, J., Schauer, N., Krueger, S.*, et al.* (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21**,** 1635-1638.

Koyama, M., Helbert, W., Imai, T., Sugiyama, J., & Henrissat, B. (1997). Parallel-up structure evidences the molecular directionality during biosynthesis of bacterial cellulose. *Proc. Natl. Acad. Sci. U. S. A.* 94**,** 9091-9095.

Kuiper, H.A., Kok, E.J., & Engel, K.-H. (2003). Exploitation of molecular profiling techniques for GM food safety assessment. *Curr. Opin. Biotechnol.* 14**,** 238-243.

Laurentin, H., Ratzinger, A., & Karlovsky, P. (2008). Relationship between metabolic and genomic diversity in sesame (Sesamum indicum L.). *BMC Genomics* 9**,** 11.

Lawoko, M., Henriksson, G., & Gellerstedt, G. (2006). Characterisation of lignin-carbohydrate complexes (LCCs) of spruce wood (Picea abies L.) isolated with two methods. *Holzforschung* 60**,** 156-161.

Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J., & Verhoeyen, M.E. (2003). Metabolite profiling of tomato (Lycopersicon esculentum) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *J. Agric. Food Chem.* 51**,** 2447-2456.

Leinhos, V., Udagamarandeniya, P.V., & Savidge, R.A. (1994). Purification of an acidic coniferin-hydrolyzing beta-glucosidase from developing xylem of Pinus-banksiana. *Phytochemistry* 37**,** 311-315.

Leple, J.C., Dauwe, R., Morreel, K.*, et al.* (2007). Downregulation of cinnamoyl-coenzyme a reductase in poplar: Multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* 19**,** 3669-3691.

Levandi, T., Leon, C., Kaljurand, M., Garcia-Canas, V., & Cifuentes, A. (2008). Capillary electrophoresis time-of-flight mass spectrometry for comparative metabolomics of transgenic versus conventional maize. *Anal. Chem.* 80**,** 6329-6335.

Lewis, N.G. & Yamamoto, E. (1990). Lignin occurrence biogenesis and biodegradation. *Annu. Rev. Plant Phys.* 41**,** 455-496.

Li, L., Popko, J.L., Umezawa, T., & Chiang, V.L. (2000). 5-hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosynthesis in angiosperms. *J. Biol. Chem.* 275**,** 6537-6545.

Li, L.G., Cheng, X.F., Leshkevich, J., Umezawa, T., Harding, S.A., & Chiang, V.L. (2001). The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. *Plant Cell* 13**,** 1567-1585.

Li, L.G., Cheng, X.F., Lu, S.F., Nakatsubo, T., Umezawa, T., & Chiang, V.L. (2005). Clarification of cinnamoyl co-enzyme a reductase catalysis in monolignol biosynthesis of aspen. *Plant Cell Physiol.* 46**,** 1073-1082.

Li, L.G., Lu, S.F., & Chiang, V. (2006). A genomic and molecular view of wood formation. *Crit. Rev. Plant Sci.* 25**,** 215-233.

Li, W.T. & Nyholt, D.R. (2001). Marker selection by Akaike information criterion and Bayesian information criterion. *Genet. Epidemiol.* 21**,** S272-S277.

Lim, E.-K., Li, Y., Parr, A., Jackson, R., Ashford, D.A., & Bowles, D.J. (2001). Identification of glucosyltransferase genes involved in sinapate metabolism and lignin synthesis in Arabidopsis. *J. Biol. Chem.* 276**,** 4344-4349.

Lisec, J., Meyer, R.C., Steinfath, M.*, et al.* (2008). Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations. *Plant J.* 53**,** 960-972.

Long, M., Millar, D.J., Kimura, Y.*, et al.* (2006). Metabolite profiling of carotenoid and phenolic pathways in mutant and transgenic lines of tomato: Identification of a high antioxidant fruit line. *Phytochemistry* 67**,** 1750-1757.

Mellerowicz, E.J., Baucher, M., Sundberg, B., & Boerjan, W. (2001). Unravelling cell wall formation in the woody dicot stem. *Plant Mol. Biol.* 47**,** 239-274.

Mellerowicz, E.J. & Sundberg, B. (2008). Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. *Curr. Opin. Plant Biol.* 11**,** 293-300.

Merchant, A., Richter, A., Popp, M., & Adams, M. (2006). Targeted metabolite profiling provides a functional link among eucalypt taxonomy, physiology and evolution. *Phytochemistry* 67**,** 402-408.

Meyer, R.C., Steinfath, M., Lisec, J.*, et al.* (2007). The metabolic signature related to high plant growth rate in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 104**,** 4759-4764.

Morcuende, R., Bari, R., Gibon, Y.*, et al.* (2007). Genome-wide reprogramming of metabolism and regulatory networks of Arabidopsis in response to phosphorus. *Plant Cell Environ.* 30**,** 85-112.

Morreel, K., Goeminne, G., Storme, V., *et al.* (2006). Genetical metabolomics of flavonoid biosynthesis in Populus: a case study. *Plant J.* 47**,** 224-237.

Morris, C.R., Scott, J.T., Chang, H.-M., Sederoff, R.R., O'Malley, D., & Kadla, J.F. (2004). Metabolic profiling: A new tool in the study of wood formation. *J. Agric. Food Chem.* 52**,** 1427-1434.

Nair, R.B., Xia, Q., Kartha, C.J., *et al.* (2002). Arabidopsis CYP98A3 mediating aromatic 3-hydroxylation. Developmental regulation of the gene, and expression in yeast. *Plant Physiol.* 130**,** 210-220.

Nairn, C.J., Lennon, D.M., Wood-Jones, A., Nairn, A.V., & Dean, J.F.D. (2008). Carbohydrate-related genes and cell wall biosynthesis in vascular tissues of loblolly pine (Pinus taeda). *Tree Physiol.* 28**,** 1099-1110.

Nakamura, Y., Kimura, A., Saga, H., *et al.* (2007). Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture. *Planta* 227**,** 57-66.

Nakashima, J., Awano, T., Takabe, K., Fujita, M., & Saiki, H. (1997). Immunocytochemical localization of phenylalanine ammonia-lyase and cinnamyl alcohol dehydrogenase in differentiating tracheary elements derived from Zinnia mesophyll cells. *Plant Cell Physiol.* 38**,** 113-123.

Nielsen, N.P.V., Carstensen, J.M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A.* 805**,** 17-35.

Nikiforova, V.J., Daub, C.O., Hesse, H., Willmitzer, L., & Hoefgen, R. (2005a). Integrative gene-metabolite network with implemented causality deciphers informational fluxes of sulphur stress response. *J. Exp. Bot.* 56**,** 1887-1896.

Nikiforova, V.J., Kopka, J., Tolstikov, V., *et al.* (2005b). Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. *Plant Physiol.* 138**,** 304-318.

Oh, S., Park, S., & Han, K.H. (2003). Transcriptional regulation of secondary growth in Arabidopsis thaliana. *J. Exp. Bot.* 54**,** 2709-2722.

Osakabe, K., Tsao, C.C., Li, L., *et al.* (1999). Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 96**,** 8955-8960.

Osuna, D., Usadel, B., Morcuende, R., *et al.* (2007). Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings. *Plant J.* 49**,** 463-491.

Ott, K.-H., Aranibar, N., Singh, B., & Stockton, G.W. (2003). Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* 62**,** 971-985.

Park, A.Y., Canam, T., Kang, K.Y., Ellis, D.D., & Mansfield, S.D. (2008). Over-expression of an arabidopsis family A sucrose phosphate synthase (SPS) gene alters plant growth and fibre development. *Transgenic Res.* 17**,** 181-192.

Parvathi, K., Chen, F., Guo, D., Blount, J.W., & Dixon, R.A. (2001). Substrate preferences of O-methyltransferases in alfalfa suggest new pathways for 3-O-methylation of monolignols. *Plant J.* 25**,** 193-202.

Patzlaff, A., Newman, L.J., Dubos, C.*, et al.* (2003). Characterisation of PtMYB1, an R2R3-MYB from pine xylem. *Plant Mol. Biol.* 53**,** 597-608.

Popper, Z.A. & Fry, S.C. (2008). Xyloglucan-pectin linkages are formed intra-protoplasmically, contribute to wall-assembly, and remain stable in the cell wall. *Planta* 227**,** 781-794.

Ralph, J., Lapierre, C., Marita, J.M.*, et al.* (2001). Elucidation of new structures in lignins of CAD- and COMT- deficient plants by NMR. *Phytochemistry* 57**,** 993-1003.

Rasmussen, S. & Dixon, R.A. (1999). Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. *Plant Cell* 11**,** 1537-1551.

Ratcliffe, R.G. & Shachar-Hill, Y. (2001). Probing plant metabolism with NMR. in Jones, R.L., Bohnert, H.J., & Delmar, D.P. (Eds), Annu. Rev. Plant Phys. Annual Reviews, Palo Alto, pp. 499-526.

Rider, S.D., Hemm, M.R., Hostetler, H.A., Li, H.C., Chapple, C., & Ogas, J. (2004). Metabolic profiling of the Arabidopsis pkl mutant reveals selective derepression of embryonic traits. *Planta* 219**,** 489-499.

Rizhsky, L., Liang, H.J., Shuman, J., Shulaev, V., Davletova, S., & Mittler, R. (2004). When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiol.* 134**,** 1683-1696.

Robinson, D.G. (1996). SuSy ergo GluSy: New developments in the field of cellulose biosynthesis. *Bot. Acta* 109**,** 261-263.

Roepenack-Lahaye, E.v., Degenkolb, T., Zerjeski, M.*, et al.* (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol. (Rockville)* 134**,** 548-559.

Roessner-Tunali, U., Hegemann, B., Lytovchenko, A.*, et al.* (2003). Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence

of hexose phosphorylation diminishes during fruit development. *Plant Physiol.* 133**,** 84-99.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N., & Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23**,** 131-142.

Roessner, U., Luedemann, A., Brust, D.*, et al.* (2001a). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13**,** 11-29.

Roessner, U., Willmitzer, L., & Fernie, A.R. (2001b). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127**,** 749-764.

Rohde, A., Morreel, K., Ralph, J.*, et al.* (2004). Molecular phenotyping of the pal1 and pal2 mutants of Arabidopsis thaliana reveals far-reaching consequences on and carbohydrate metabolism. *Plant Cell* 16**,** 2749-2771.

Samuels, A.L., Rensing, K.H., Douglas, C.J., Mansfield, S.D., Dharmawardhana, D.P., & Ellis, B.E. (2002). Cellular machinery of wood production: Differentiation of secondary xylem in Pinus contorta var. latifolia. *Planta (Berlin)* 216**,** 72-82.

Sanchez, D.H., Siahpoosh, M.R., Roessner, U., Udvardi, M., & Kopka, J. (2008). Plant metabolomics reveals conserved and divergent metabolic responses to salinity. *Physiol. Plantarum.* 132**,** 209-219.

Schad, M., Mungur, R., Fiehn, O., & Kehr, J. (2005). Metabolic profiling of laser microdissected vascular bundles of Arabidopsis thaliana. *Plant Methods* 1**,** 2.

Schauer, N. & Fernie, A.R. (2006). Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci.* 11**,** 508-516.

Schoch, G., Goepfert, S., Morant, M.*, et al.* (2001). CYP98A3 from Arabidopsis thaliana is a 3'-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. *J. Biol. Chem.* 276**,** 36566-36574.

Shepherd, T., Dobson, G., Verrall, S.R.*, et al.* (2007). Potato metabolomics by GC-MS: what are the limiting factors? *Metabolomics* 3**,** 475-488.

Sibout, R., Eudes, A., Pollet, B.*, et al.* (2003). Expression pattern of two paralogs encoding cinnamyl alcohol dehydrogenases in Arabidopsis. Isolation and characterization of the corresponding mutants. *Plant Physiol.* 132**,** 848-860.

Sibout, R., Eudes, A., Mouille, G.*, et al.* (2005). Cinnamyl alcohol dehydrogenase-C and -D are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis. *Plant Cell* 17**,** 2059-2076.

Sjostrom, E. (1993). Wood chemistry. Academic Press, Inc., San Diego.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78**,** 779-787.

Smith, C.G., Rodgers, M.W., Zimmerlin, A., Ferdinando, D., & Bolwell, G.P. (1994). Tissue and subcellular immunolocalisation of enzymes of lignin synthesis in differentiating and wounded hypocotyl tissue of French bean (Phaseolus vulgaris L.). *Planta (Heidelberg)* 192**,** 155-164.

Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., & Nishioka, T. (2003). Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* 2**,** 488-494.

Srere, P.A. (1987). Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.* 56**,** 89-124.

Steeves, V., Forster, H., Pommer, U., & Savidge, R. (2001). Coniferyl alcohol metabolism in conifers - I. Glucosidic turnover of cinnamyl aldehydes by UDPG: coniferyl alcohol glucosyltransferase from pine cambium. *Phytochemistry* 57**,** 1085-1093.

Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19**,** 1019-1026.

t'Kindt, R., De Veylder, L., Storme, M., Deforce, D., & Van Bocxlaer, J. (2008). LC-MS metabolic profiling of Arabidopsis thaliana plant leaves and cell cultures: Optimization of pre-LC-MS procedure parameters. *J. Chromatogr. B* 871**,** 37-43.

Terskikh, V.V., Feurtado, J.A., Borchardt, S., Giblin, M., Abrams, S.R., & Kermode, A.R. (2005). In vivo C-13 NMR metabolite profiling: potential for understanding and assessing conifer seed quality. *J. Exp. Bot.* 56**,** 2253-2265.

Thimm, O., Blasing, O., Gibon, Y.*, et al.* (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37**,** 914-939.

Tikunov, Y., Lommen, A., de Vos, C.H.R.*, et al.* (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* 139**,** 1125-1137.

Tohge, T., Nishiyama, Y., Hirai, M.Y.*, et al.* (2005). Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant J.* 42**,** 218-235.

Tolstikov, V.V. & Fiehn, O. (2002). Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* 301**,** 298-307.

Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N., & Fiehn, O. (2003). Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* 75**,** 6737-6740.

Ulbrich, B. & Zenk, M.H. (1980). Partial-purification and properties of para-hydroxycinnamoyl-CoA - shikimate-para-hydroxycinnamoyl transferase from higher-plants. *Phytochemistry* 19**,** 1625-1629.

Uraki, Y., Nakamura, A., Kishimoto, T., & Ubukata, M. (2007). Interaction of hemicelluloses with monolignols. *J. Wood Chem. Technol.* 27**,** 9-21.

Urbanczyk-Wochniak, E., Baxter, C., Kolbe, A., Kopka, J., Sweetlove, L.J., & Fernie, A.R. (2005). Profiling of diurnal patterns of metabolite and transcript abundance in potato (Solanum tuberosum) leaves. *Planta* 221**,** 891-903.

van Riel, N.A.W. (2006). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.* 7**,** 364-374.

Wafler, U. & Meier, H. (1994). Enzyme-activities in developing cotton fibers. *Plant Physiol. Biochem.* 32**,** 697-702.

Wang, S.Y., Wang, Y.S., Tseng, Y.H., Lin, C.T., & Liu, C.P. (2006). Analysis of fragrance compositions of precious coniferous woods grown in Taiwan. *Holzforschung* 60**,** 528-532.

Weckwerth, W. (2003). Metabolomics in systems biology. *Annu. Rev. Plant Biol.* 54**,** 669-689.

Weckwerth, W., Loureiro, M.E., Wenzel, K., & Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* 101**,** 7809-7814.

Welti, R., Li, W.Q., Li, M.Y.*, et al.* (2002). Profiling membrane lipids in plant stress responses - Role of phospholipase D alpha in freezing-induced lipid changes in Arabidopsis. *J. Biol. Chem.* 277**,** 31994-32002.

Whetten, R. & Sederoff, R. (1995). Lignin biosynthesis. *Plant Cell* 7**,** 1001-1013.

Winkel-Shirley, B. (1999). Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiol. Plantarum.* 107**,** 142-149.

Wolfender, J.-L., Ndjoko, K., & Hostettmann, K. (2003). Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic

resonance spectroscopy: A powerful combination for the on-line structural investigation of plant metabolites. *J. Chromatogr. A.* 1000**,** 437-455.

Wurtele, E.S., Li, J., Diao, L.*, et al.* (2003). MetNet: software to build and model the biogenetic lattice of Arabidopsis. *Comp. Funct. Genom.* 4**,** 239-245.

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Commun. Stat.-Theory Methods* 36**,** 2395-2403.

Yang, S., Qiao, B., Lu, S.H., & Yuan, Y.J. (2007). Comparative lipidomics analysis of cellular development and apoptosis in two Taxus cell lines. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1771**,** 600-612.

Ye, Z.-H. (1997). Association of caffeoyl coenzyme A 3-O-methyltransferase expression with lignifying tissues in several dicot plants. *Plant Physiol. (Rockville)* 115**,** 1341-1350.

Yeh, T.F., Morris, C.R., Goldfarb, B., Chang, H.M., & Kadla, J.F. (2006). Utilization of polar metabolite profiling in the comparison of juvenile wood and compression wood in loblolly pine (Pinus taeda). *Tree Physiol.* 26**,** 1497-1503.

York, W.S. & O'Neill, M.A. (2008). Biochemical control of xylan biosynthesis - which end is up? *Curr. Opin. Plant Biol.* 11**,** 258-265.

CHAPTER 2

Metabolite profiling of Douglas-fir (*Pseudotsuga menziesii*) field trials reveals strong environmental and weak genetic variation

A version of this chapter has been published and the original publication is available at www3.interscience.wiley.com. Robinson, A.R. Ukrainetz; N.K. Kang, K.-Y and Mansfield, S.D. 2007. Metabolite profiling of Douglas-fir (*Pseudotsuga menziesii*) field trials reveals strong environmental and weak genetic variation. New Phytologist. 174:762-773.

2.1 Introduction

Recently, non-targeted metabolite analysis (metabolomics) has evolved into a new branch of functional genomics, which complements transcriptomics and proteomics technologies. Ideally, metabolomics aims to identify and quantify the full complement of small molecular weight, soluble metabolites in actively metabolising tissues (Fiehn and Weckwerth, 2003). However, in practicality, the narrow molecular specificity of individual analytical techniques, and difficulties in amalgamating substantial data sets acquired using multiple techniques, have thus far generally restricted analyses to "targeted" subsets of the greater metabolite pool (e.g. phenolics, carbohydrates, anthocyanins). Once collected, such data may be associated with measurements of plant genetic and overt quantitative or qualitative phenotypic traits, permitting correlative associations to be drawn between plants' metabolite "pools" and their genetic background, inherent phenotypic characteristics, responses to biotic and abiotic stress and/or genetic mutations (e.g. the *Arabidopsis* 'pkl' mutant in which seedlings retain some metabolic traits of embryos (Rider *et al.*, 2004)). Through this connectivity, metabolomic data may assist in establishing causal relationships between genetic, metabolic and phenotypic phenomena. In recent years metabolomics has been used successfully on numerous plant genera including *Arabidopsis* (Fiehn *et al.*, 2000a; Roepenack-Lahaye *et al.*, 2004), *Populus* (Jeong *et al.*, 2004; Robinson *et al.*, 2005), *Medicargo* (Huhman and Sumner, 2002), *Solanum* (Roessner *et al.*, 2001a; Szopa, 2002), *Cucurbita* (Fiehn, 2003), *Pinus* (Morris *et al.*, 2004) and most recently *Triticum* (Baker *et al.*, 2006).

Metabolomics has demonstrated relationships between plant metabolite pools, genotype and phenotype, and helped to elucidate biological processes involving abiotic and biotic plant interactions in a variety of species. It is clear that metabolomics is a useful approach and promises to further contribute to our understanding of plant systems – specifically in the fields of tree growth and development. To date, most comparative metabolomics investigations have focussed on model plant systems that have been subjected to environmental extremes (Rizhsky *et al.*, 2004; Urbanczyk-Wochniak and Fernie, 2005), mutation, and/or targeted genetic modification (Le Gall *et al.*, 2003; Robinson *et al.*, 2005; Roessner *et al.*, 2001a). This approach has been effective to the extent that well-defined systems, which exhibit single-gene alterations

and corresponding phenotypes, or acute responses to specific nutritional scenarios or environmental stresses, have allowed the underlying concepts and utility of metabolomics to be evaluated. However, experiments involving model systems and extreme, controlled conditions bare limited resemblance to the development of plant populations in "real-world" contexts. It is under exposure to variable genetic and environmental factors that the plastic nature of plant development is revealed, giving rise to observed variability in phenotypic parameters. Presumably, such variation is accompanied by corresponding shifts in metabolism that may be detected in the metabolite pools. Therefore, broad-scale elucidation of metabolic structure and the association of this with the genotypic, phenotypic and/or environmental characteristics of plant populations may aid in linking these aspects and furthering our understanding of plant development as a whole.

The research described herein evaluated a global metabolomics approach to investigating natural variability due to the influence of family and site on wood formation and tree growth in multiple full-sib Douglas-fir (*Pseudotsuga menziesii*) trees selected from an advanced second generation breeding population, duplicated by site. This research represents a fundamental, non-targeted assessment of one of the newest branches of functional genomics for discerning biological variation in tree species. It demonstrates a technical ability to reveal the expected coherency between metabolic traits and other biotic and abiotic parameters, in the context of tree populations.

2.2 Materials and methods

2.2.1 Plant material and sampling

Ten, full-sib, 26-year-old Douglas-fir families from the British Columbia Ministry of Forests second generation breeding program were employed in this study. The families represent a subset of trees from an extensive multi-family, multi-site progeny study, breeding predominantly for superior growth performance. Each family is represented by ten (of a possible 16) individuals randomly selected from four blocks with four-tree row plots, randomly planted on each of two sites (total 200 trees). The two sites, Adam River (AR) and Gold River (GR), are located on Vancouver Island, British Columbia, and represent a more productive and less productive site, respectively, as defined by

Douglas-fir height growth classification. Nineteen random samples were lost during transit and processing resulting in a total of 181 samples over the ten families.

Sampling was conducted over a four-day period in late summer (August 6th-9th 2003). This period represents the latter part of the growing season, when latewood formation is occurring, and the cambial tissue was very fluid during sampling and thus suggested that wood-forming metabolism was still active. The developing xylem tissue was obtained from each tree by first peeling a section of bark/phloem/outer cambium from the main bole of the trunk at breast height, and then scraping the inner cambium with a fresh razor blade. The collected material was immediately transferred to a cryovial, snap-frozen in liquid nitrogen and maintained in a cooled liquid nitrogen vapour tank in the field, and then at -80°C in the laboratory. At the same time, a 10 mm increment core was extracted at breast height for wood fibre evaluation, and the diameter at breast height (DBH) and absolute tree height were recorded.

2.2.2 Quantitative wood traits

A concurrent study focussing on genetic mapping of phenotypic growth and wood traits used tree measurements and the increment core wood from each sample tree to measure a set of 16 quantitative traits, including: tree diameter at breast height (DBH), height (HT), and volume (VOL); wood microfibril angle (MFA), fibre length (FL), fibre coarseness (Cs), earlywood density (ED), latewood density (LWD), average density – density of entire increment core (AD), latewood proportion (LWP); wood chemistry traits including total lignin content (TL) 6.3Appendix D.1), and arabinose (Ara), galactose (Gal), glucose (Glu), mannose (Man) and xylose (Xyl) contents (6.3Appendix D.2).

2.2.3 Calculation of site index

Site index, a measure of site productivity, was employed to characterise each site by estimating the height of dominant and co-dominant trees at age 50. Thirty trees with the largest diameter at breast height (DBH) of the sample population at each site were used to estimate site index. The breast height age was estimated using increment cores and the top height was estimated using a Vertex instrument (Vertex III; Haglöf, Sweden). Site index was then assessed for each tree using British Columbia Ministry of Forests growth intercept tables for Coastal Douglas-fir (Nigh, 1997). The individual tree site index values were then averaged for each site to estimate site productivity.

2.2.4 Metabolite sample preparation

Frozen tissue was macerated to a fine powder with a 15 s burst using a dental amalgam mixer, employing a liquid N_2-chilled copper/plastic capsule and steel ball bearings. Samples were kept frozen at all times and, once ground, were returned to -80°C.

Metabolites were extracted from tissue samples and prepared for GC/MS using a two-phase methanol/chloroform method developed for metabolite extraction from Populus cambium and developing xylem (Fiehn *et al.*, 2000a; Fiehn *et al.*, 2000b; Robinson *et al.*, 2005). Approximately 100 mg frozen, ground cambium was accurately weighed into a pre-chilled 2 mL lock-cap centrifuge tube. To this, 600 µL HPLC-grade methanol (CH_3OH) was immediately added and vortexed for 10 s to halt biological activity and minimise degradation. In addition, 40 µL distilled, deionised water and 10 µL internal standard mixture (10 mg/mL ribitol in H_2O) were added. The sample was then incubated for 15 min at 70°C with constant agitation, and centrifuged at 13 000 rpm for 5 min. The supernatant, containing extracted metabolites, was retained. $CHCl_3$ (800 µL) was then added to the pellet, vortexed for 10 s to re-suspend, and incubated for 5 min at 35°C with constant agitation. The resultant supernatant recovered, following a second 5 min centrifugation at 13 000 rpm, was pooled with the supernatant from the initial CH_3OH extraction. H_2O (600 µL) was added to the combined supernatant, vortexed for 10 s, and then centrifuged for 15 min at 4000 rpm to permit the separation of polar (methanol/water) and non-polar (methanol/chloroform) phases. This combination and separation of phases allowed metabolites extracted in one phase but with greater affinity for the other to repartition. A 1 mL aliquot of the polar (upper) phase was taken, and either processed immediately or stored at -20°C until further analysis. Metabolites in the non-polar phase were not analysed in this study.

The soluble polar metabolite samples were derivatised prior to GC/MS analyses. A 900 µL aliquot of the methanol/water phase was dried using a Vacufuge (Eppendorf) (3-4 h, 30°C), and first methoxylated by re-suspending the pellet in 50 µL methoxyamine hydrochloride solution (20 mg/mL in pyridine) and incubating with constant agitation for 2 h at 60°C in order to protect carbonyl moieties. Acidic protons were then trimethylsilylated with 200 µL N-methyl-N-trimethylsilyltrifluoro acetamide (MSTFA) and incubated at 60°C with constant agitation for 30 min. Samples were left

to stand at room temperature overnight to ensure complete derivatisation, and then filtered through compacted tissue paper prior to GC/MS analysis.

2.2.5 GC/MS analysis

GC/MS analysis was conducted on a ThermoFinnigan Trace GC-PolarisQ ion trap system fit with an AS2000 auto-sampler and a split/splitless injector. The GC was equipped with a low-bleed Restek Rtx-5MS column (fused silica, 30 m, 0.25 mm ID, stationary phase diphenyl 5% dimethyl 95% polysiloxane). The GC conditions were set as follows: inlet temperature 250°C, helium carrier gas flow at constant 1 mL/min, injector split ratio 10:1, resting oven temp 70°C, and GC/MS transfer line temperature 300°C. Following injection of a 1 µL aliquot of sample, the oven was held at 70°C for 2 min and then ramped to 325°C at a rate of 8°C/min. The temperature was held at 325°C for an additional 6 min before being cooled rapidly to 70°C in preparation for the next run.

Mass spectrometry analysis was conducted in positive electron ionisation (EI) mode, with the fore-line evacuated to approximately 40 mTorr, and with helium gas flow into the chamber set at 0.3 mL/min. The source temperature was held at 250°C, with an electron ionisation potential of 70 eV. The detector signal was recorded from 3.35 min after injection until 35.5 min, and ions were scanned across the range of 50-650 mass units (mu) with a total scan time of 0.58 s.

2.2.6 Data acquisition and processing

ThermoFinnigan 'Xcalibur' (v1.3) software was used for both GC/MS data collection and peak determination and measurement. GC/MS total ion chromatograms (TIC) of TMS-derivatives from the developing xylem at breast height, were collected for all full-sib Douglas-fir families replicated on two sites, in order to elucidate the common "metabolite pools" present in the actively metabolising developing xylem tissue of each tree.

To normalise the raw TIC peak data, the area of each peak in a chromatogram was expressed relative to the area of the ribitol internal standard peak, and then again standardised across all chromatograms by adjusting for the precise amount of tissue (mg fresh weight) used in each sample extraction.

The alignment of peaks that represented the same compound in multiple chromatograms was automated using purpose-built 'PeakMatch' software (Robinson *et*

al., 2005). Once compiled, the dataset consisted of 251 distinct compound peaks across all 181 samples (an array of ~45 000). Peaks are consequently labelled 1-251. As a means of minimising artefacts caused by sample processing and analysis, the dataset was further reduced to only those peaks that appeared in at least 10% of the samples from each site. This yielded a dataset of 139 peaks across the 181 samples (an array of ~25 000), which was used in all statistical analyses.

Intermediate data handling and manipulation were carried out using Microsoft Excel 2000 and Corel Quattro Pro 12.

2.2.7 Multivariate statistical analyses

Further reductions of the metabolite and quantitative phenotypic trait datasets were carried out by Multivariate Discriminant Analysis (MDA), Factor Analysis (FA), and Canonical Correlation Analysis (CCA) and Canonical Discriminant Analysis using the 'proc discrim', 'proc factor', 'proc cancor' and 'proc candisc' procedures of the SAS v9.1 software (SAS Institute, Inc., Cary, N.C.), respectively.

Multivariate discriminant analysis, a statistical approach that assesses the variation in pre-classified multivariate data and is capable of generating predictive models, was applied to the metabolite data array. The data structure of this research project allowed MDA models to be developed using two different classification schemes: by site (Adams River, Gold River) or by family (2, 26, 38, 46, 62, 75, 92, 130, 151, 156). During the site analysis, the data were split into four equal subsets to build the predictive model. Four models were generated; each model was developed using three of the four datasets. The fourth dataset was used as an independent validation array to assess the accuracy of the model. This process was repeated until all combinations of the four datasets were used and a final accuracy calculated as the average of the four models. For family analysis, the data were equally split into two, rather than four sets due to the limited number of samples per class (at most 10 replicates per site). In this case, two models were generated, tested, and the average accuracy calculated. For the two-class site model and the 10-class family model, prior probabilities of 0.5 and 0.1 (50% and 10%) are expected, respectively. Higher model accuracy than the prior probabilities implies that the MDA is able to distinguish between classes at a higher probability than random chance.

Factor analysis (FA) allows the variation in metabolite and quantitative trait data arrays to be explored without the constraints of data pre-classification (as is the case with MDA). Initial exploratory analyses were carried out without limiting the number of factors generated (essentially making the factor analysis a principal components analysis). The Eigenvalues and scree plot slope shifts (Tabachnick and Fidell, 2001) were used to select factors that represented significant portions of the variation in a dataset. The FA was then rerun specifying an orthogonal 'varimax' rotation and the number of factors to be used in the rotation. Factor scores were plotted on the axes of scatter plots to generate a graphical representation of the variation in the original data captured by the analysis. The separation of sample clusters is considered to illustrate differences between distinct metabolic systems (Chen *et al.*, 2003; Fiehn, 2003; Fiehn *et al.*, 2000a; Morris *et al.*, 2004; Roessner *et al.*, 2001a; Roessner *et al.*, 2001b).

Canonical correlation analysis is used to investigate the relationship between two groups of variables (X and Y), and transforms the data into canonical variables in such a way as to maximise the covariance between groups. Specifically, in our study, this technique was used to explore the relationships between the metabolite array and quantitative phenotypic traits having relevance to tree growth and wood quality characteristics. The first group of variables was comprised of 139 metabolites for both the Adams River and Gold River sites, while the second consisted of the 16 quantitative phenotypic traits described above. Canonical variables were considered important if the canonical correlation was large, and significant at an alpha value of 0.05. It was also necessary for the transformed variables to explain a considerable proportion of the standardised variation in the original data, as described by canonical redundancy analysis. The structure correlation coefficients (between canonical variables and original metabolites or growth trait variables) were used to identify variables in the two sets that were related via the canonical correlation. Variables with correlations >0.3 explained 10% or more of the variance, and were considered to be part of the canonical variable.

Canonical discriminant analysis (CDA) is a multivariate statistical technique that derives linear combinations of groups of variables (metabolites) in such a way that maximizes the variation between classes (families or sites). The multivariate analysis of

variance (MANOVA) output generated by CDA was used to test the ability to distinguish families and sites based on metabolite data and confirm results generated by MDA.

2.2.8 Calculation of heritabilities

Broad-sense heritability is an estimate of the total amount of variation that can be explained by genetics (additive, dominance and epistatic variation) and is measured on a scale between 0 (little genetic control) to 1 (entirely controlled by genetics). These estimates are an indication of the amount of variation caused by family versus environmental (site) effects.

SAS was used to generate components of variance for the calculation of metabolite heritability values and to test model parameters for family, site and family-by-site interaction. 'Proc GLM' was used to conduct analysis of variance for all metabolites using the following components of variance and linear model:

	df	Components of Variance
Family	(f-1)	σ2E + nσ2FB + bnσ2FS + bcnσ2F
Site	(s-1)	σ2E + fnσ2B + fbnσ2S
Family*Site	(f-1)(s-1)	σ2E + nσ2FB + bnσ2FS
Block(site)	s(b-1)	σ2E + fnσ2B
Family*Block(Site)	s(b-1)(f-1)	σ2E + nσ2FB
Sampling Error	sfb(n-1)	σ2E

F = family; B = block; S = site; f = # of families; b = # of blocks; s = # of sites; n = # of trees

$$Yijlp = \mu + Fi + Sl + Bj(l) + FBij(l) + FSil + Ep(ijl)$$

where, Yijlp is the individual phenotypic observation, μ is the overall mean, Fi is the fixed family effect, Sl is the random site effect, Bj(l) is the random block effect, FBij(l) is

the random family-by-block interaction nested within site, FSil is the random family-by-site interaction and Ep(ijl) is the random residual effect.

Variance components for broad-sense heritability calculations were estimated using the REML method of 'proc VARCOMP'. Broad-sense heritability was calculated for all metabolites showing significant family variation (F-test, $\alpha = 0.05$) using the following formula:

$$H^2 = \frac{2\sigma 2F}{\sigma 2F + \sigma 2FS + \sigma 2FB + \sigma 2E}$$

where, σ2F is family variance, σ2FS is the variance of family-by-site interaction, σ2FB is the variance of family-by-block nested within site and σ2E is the residual variance.

2.2.9 Compound identification

National Institute of Standards and Technology (NIST) MS-Search software equipped with the NIST mass spectra, as well as the Max Planck Institute Trimethylsilane (TMS) (http://www.mpimp-Golm.mpg.de/mms-library/index-e.html), Gölm Metabolome Database (http://csbdb.mpimp-Golm.mpg.de/csbdb/gmd/gmd.html) (Kopka *et al.*, 2005) and our own (Mansfield UBC laboratory) TMS mass spectral libraries were collectively used to identify metabolites of interest, as highlighted by the statistical analyses.

2.3 Results and discussion

2.3.1 Family-related variation

Factor analysis (FA) and multivariate discriminant analysis (MDA) were performed on the metabolite dataset (181 trees, 10 families, 2 sites, 139 metabolites), focusing on family variation. In the factor analysis, five factors that collectively accounted for 51% of the total variance were included in the varimax rotation. Although marked clustering and separation of samples was observed in certain factors, this was not family related (Figure 2.1a). In light of the apparent dominance of site over other effects when both sites were analysed together, separate FAs for samples from each site individually were conducted as a potential means of revealing distinctions between families, free of the complexities of site interactions. In these analyses, some individual family clusters did separate from one another in factor score plots of various factor pairs (data not shown).

A family-based factor analysis was also conducted on the data for a set of 16 quantitative phenotypic traits, which gave very similar results to the metabolite FA (Figure 2.2a). The first four factors, which accounted for 67% of the variation in that data, were used in the varimax rotation. When both sites were analysed together, no separation of family clusters was evident. However, when each site was analysed separately, some family separation was apparent, but as with the metabolites, no clear distinctions were observed (data not shown).

When the dataset included samples from both sites (Adam River and Gold River) the MDA was only 18% accurate on average and 37% accurate at best (Table 2.2a); this represents an improvement over the 10% probability of random chance, and implies that family variation can be distinguished. These findings were supported by the results of a canonical discriminant analysis (CDA) which was used to analyze the same data, and showed that the MANOVA results could distinguish clearly between families (Figure 2.3) based on the 139 metabolites used in the analysis ($p<0.05$). MDA accuracy was further improved when samples from the two locations were analysed separately, with a moderate improvement for Adams River (37% on average, 67% at best) and a more pronounced improvement for Gold River (65% on average, 90% at best). The improvement observed when samples from each location were analysed separately is noteworthy and alludes to a confounding influence of site when investigating genetic variation in this and other tree populations (*i.e.* family $\times$ site interactions).

2.3.2 Site-related variation

Analyses that focused on site-based variation were conducted as a complement to those relating to the family variation, described above. Adam River and Gold River differed in site productivity: Adam River was a more productive site with a site index of 39.7 m, and Gold River was a less productive site with a site index of 35.4 m. Adam River and Gold River are both located on Vancouver Island in the CWHvm and CWHxm biogeoclimatic subzones, respectively. Adam River (latitude: 50º 24’ 00; longitude: 126º 10’ 00) is 576 m above sea level and has very little under story vegetation, while Gold River (latitude: 49º 51’ 30; longitude: 126º 04’ 45) is 561 m above sea level and has an understory composed primarily of *Vaccinium* spp. The largest difference in site is related to the precipitation regime, with Adam River being classified as a “very wet” (v) environment and Gold River located in the “very dry” (x) biogeoclimatic region. Both

sites were on relatively flat terrain, free of stumps and were surrounded by even-age stands that did not restrict light access and protected the stands from wind damage. The major biogeoclimatic difference was in water availability, which will also influence both understory and soil composition. The site-related factor analysis of the metabolite dataset was the same as that used in the family analysis (above), however, the samples were labelled by site rather than family (Figure 2.2b-d). The three highest-ranking factors (F-1, F-2 and F-3, accounting for 16.7%, 12.1% and 11.5% of the dataset variance, respectively) were responsible for clustering and separation of the samples, with site being the dominant influence (Figure 2.1b-d). F-1 was the primary source of separation between site clusters, and a positive relationship between scores in F-1 and F-3 improved the separation (Figure 2.2c). A small cluster of four AR samples that grouped with the GR cluster in F-1 is effectively isolated by F-2 (Figures 2.1b and 2.2c), and these samples presumably represent a variant metabolic subset.

The site-related factor analysis of the phenotypic trait dataset was also the same as that used in the family analysis (above), involving a varimax rotation of the first four factors, which collectively accounted for 0.67 of the total variance. In this analysis, F-1 was primarily responsible for clustering and separating the trees based on site, with some improvement offered by F-3 and F-4 (Figure 2.2b-d). These factors accounted for 25.5%, 14.0% and 8.9% of the dataset variance, respectively.

The MDA for site, based on the metabolite dataset, showed strong predictive accuracy (Table 2.1b), which is indicative of large and/or consistent metabolic differences between populations from the two sites. MANOVA results derived from the CDA confirm that sites can be distinguished based on the 139 metabolites used in the analysis ($p<0.05$).

The results from the MDA and CDA of GC/MS metabolite profiles of developing xylem and FA of metabolite profiles and quantitative phenotypic traits indicate that in this Douglas-fir population, a much clearer distinction can be made between trees based on site, compared to genetic origin (family), however, both can be differentiated. It is apparent from the metabolite profiles that differences between sites have had a detectable influence on the wood-forming metabolism of the trees. Although it is generally accepted that growing conditions can significantly influence metabolism and phenotypic traits in trees, to date there have been few demonstrations of the influences

of uncontrolled site (climatic and environmental) factors on global metabolism in plant species. The findings of this study are consistent with those of Baker *et al*. (2006), for whom PCA of NMR-derived metabolic profiles demonstrated a much clearer distinction between transgenic and control wheat lines on the basis of site, rather than genotype.

2.3.3 Interaction between genetic and environmental elements

The determination of metabolites exhibiting significant family- or site-related variation, and subsequent calculation of the broad-sense heritabilities of metabolite pools, provided a quantitative representation of the trends observed in MDA, CDA and FA. Of the complete set of 139 metabolites, seventy-eight (56.1%) showed significant family variation, 108 (77.7%) had significant site variation, while 53 (38.19%) showed significant family-by-site interaction (ANOVA, $\alpha = 0.05$). Broad-sense heritability estimates of the individual metabolites ranged from 0 to 0.67 with only one being >0.5. The generally low values of these estimates (mean = 0.12) further suggests that genetics (family) has a smaller influence on the observed variation in cambial metabolism, than environmental (site) factors. Furthermore, greater than 1/3 of all the metabolites showed significant family-by-site interaction indicating that families often produce different metabolic responses to similar environmental cues. This analysis clearly illustrates that cambial metabolism is a complex response to both genetic and environmental stimuli, and the interaction of the two. This result agrees with a previous study of the relative influence of genetic and specific environmental factors in *Pinus sylvestris*, in which significant family $\times$ temperature and family $\times$ temperature $\times$ water interactions were observed, in the absence of significant family main effects (Sonesson and Eriksson, 2000). Furthermore, this helps to explain why the MDA family predictions were improved when the Adams River and Gold River sites were analysed separately. Of the 108 metabolites with significant site variation, 64 (59.3%) showed significant family variation but no family-by-site interaction. For this subset of site-distinguishing metabolites, heritability was only slightly lower than that of the complete set (ranging from 0.00 to 0.67 and with a mean of 0.11), lending further support to the hypothesis that environment (site) was more largely responsible for the observed metabolic variation, than genetics (family) origin. For a complete list of these 64 compounds, with mass spectral data and possible chemical class assignments, see the supplemental material (Appendix A.1).

It was possible to assign positive identities to approximately half of the 64 compounds that exhibited significant site and family variation with no family-by-site interaction, based on GC retention time and mass-spectral matches (Table 2.2a). Several aspects of metabolism are represented, with some notable inclusions from branches of metabolism involved in wood formation. The list includes participants in the tricarboxylic acid (TCA) cycle (fumaric and malic acids), the major sugar pools and pentose phosphates (sucrose, fructose, Fructose-6P, glucose and Glucose-6P), and metabolites related to lignin biosynthesis (coniferin and quinic acid). The identities of metabolites with the highest heritabilities are those related to carbohydrate metabolism. This is in agreement with the heritabilities calculated for quantitative traits, in which the glucan (*i.e.* cellulose), arabinose and xylose contents of wood were high relative to others traits.

Heritabilities were also calculated for the 16 quantitative phenotypic traits, and although they were larger on average than for the metabolites, the estimates were still fairly low (Table 2.2b). Of the heritable traits measured, tree height, arabinose, xylose and glucose content, all had heritabilities greater than 0.35. In particular, arabinose and glucose contents were greater than 50%. The broad-sense heritability estimate for glucose (1.28) is an over-estimation that is likely a result of the small number of families used in the calculations. It is an indication that these values be used in a relative sense for comparison with each other, rather than absolute values. However, despite this, the generally low heritabilities observed for the phenotypic traits should still be applicable. As with metabolites, genetics (family) does not appear to have much influence on the observed variation in phenotypic traits.

2.3.4 Interaction between metabolic and phenotypic elements

A canonical correlation analysis (CCA) including 139 metabolites and 16 phenotypic traits was conducted. In this analysis, the first pair of canonical variables (Metabolite 1 and Growth 1) was the only relevant set. The canonical correlations for all 16 variate pairs were high (ranging from 0.99 to 0.74), yet only variates one and two were significant at an alpha of 0.05 (0.0006 and 0.0282, respectively). In addition, canonical redundancy analysis showed that only the first variate exhibited predictive power with regard to both sets of original variables, and that this was limited to prediction of variance in growth traits only. The transformed canonical variables Metabolite 1 and

Growth 1 accounted for a small proportion of the variation of the original data (0.2165 and 0.2207, respectively). Although low, these values are considerably higher than those for the second and subsequent sets of canonical variables.

The metabolites' and growth traits' canonical correlation coefficients (canonical factor loadings) for the first canonical variate have been assembled in Tables 2.3a and 2.3b. In total, 52 of 139 metabolites and 10 of 16 growth traits were significantly correlated with their canonical variate (Metabolite 1 and Growth 1, respectively), although the correlation for latewood density was barely below the 0.3 cut-off. Due to space limitations and to aid clarity, only metabolites whose correlation was significant (>0.3) and whose identity could be positively determined have been presented here. For a complete list of 52 compounds, with mass spectral data and possible chemical class assignments see the supplemental material (Appendix A.2).

For the phenotypic traits (Table 2.3b), measures of wood yield (tree diameter at breast height, volume and height) were highly correlated with Growth 1. Similarly, indicators of wood fibre quality (microfibril angle, fibre length and coarseness) were also highly correlated. This suggests that Growth 1 is strongly related to wood yield and wood cell morphology. Additionally, the contents of primary chemical constituents of wood (total lignin, glucose, mannose, xylose) show less influence on Growth 1, with lower, but significant correlation coefficients. Correlation coefficients for traits related to wood density (latewood and earlywood density, average density, and latewood proportion) were less than 0.3, and as such did not significantly influence Growth 1.

Many metabolite pools are correlated well with Metabolite 1. A spread of metabolites associated with the tricarboxylic acid (TCA) cycle (Fumaric acid), ascorbate and aldarate metabolism (threonic acid), amino acid metabolism (glyceric acid, pyroglutamic acid, alanine), carbohydrate storage (rhamnose), and stress tolerance (pinitol) are present. Significant correlations are apparent for major (glucose and fructose) and minor (xylose, arabinose, and maltose) sugar pools. The pools of glucose and fructose are catabolite products of sucrose the major transportable photoassimilate, and represent a starting point for many branches of metabolism, the most notable of which is cell wall biosynthesis. The minor pools observed are involved in ascorbate, nucleotide and more specific aspects of cell wall metabolism. All three have structural roles in cell walls, while xylose in particular is a key cell wall carbohydrate associated

with primary wall deposition and a component of wood hemicellulose. Precursors to lignin biosynthesis (shikimic acid, coniferin, quinic acid) also correlate well with Metabolite 1. Coniferin is believed to be involved in the transportation and storage of the monolignol coniferyl alcohol, and consequently plays an integral role in the process of cell wall lignification in softwoods (Samuels *et al.*, 2002). On the other hand, shikimic and quinic acids are more broadly associated precursors, acting as intermediates in the synthesis of aromatic amino acids, flavonoids, and a range of other secondary metabolites aside from their involvement in lignin biosynthesis. It is, therefore, fitting that both shikimic and quinic acids are seen forming pools in the developing xylem of Douglas-fir, a phenomenon frequently associated with roles in alternative downstream pathways (Srere, 1987). Aside from their roles in the broadly-serving shikimate pathway (reviewed by Herrmann and Weaver, 1999), there is support for their participation in the formation of shikimate and quinate esters of *p*-coumarate, as part of the metahydroxylation of that molecule in the phenylpropanoid pathway specifically responsible for monolignol biosynthesis (Humphreys and Chapple, 2002). The lignin-related metabolites, shikimic acid and coniferin, are among those most highly correlated to Metabolite 1, along with a number of amino acid metabolites and pinitol. These compounds predominate over the precursors of structural carbohydrates, which, although relevant, do not have as strong an influence as Metabolite 1.

Collectively, the correlations between metabolites and growth traits and their highly-correlated canonical variates indicate a clear link between wood yield and fibre quality of a tree, and the pooling of a series of metabolites related directly to wood biosynthesis in the developing xylem. Firstly, there is an inverse relationship between pools of metabolite precursors to significant carbohydrate components of wood and the presence of the structural components themselves (glucose, mannose, and xylose). This suggests that increased pooling of these metabolites occurs as a consequence of limited metabolic flux beyond the pool, and that reduced incorporation into the cell wall matrix is not a consequence of limited precursor availability, but rather of low demand. A similar but stronger inverse relationship also exists between the pools of lignin precursors and the total lignin content of wood, whereby the pools of coniferin, shikimic acid and quinic acid become larger as the total lignin content of wood is reduced. Again, such a relationship implies that the limiting factor in lignin biosynthesis is

deposition, rather than precursor supply. Finally, there is a simple inverse, but perhaps tell-tale relationship between the measures of yield (DBH, VOL, HT) and the pool of pinitol. As a “compatible osmoticum” that has been associated with a response to drought stress (Griffin *et al.*, 2004; Keller and Ludlow, 1993), the observed negative correlation between this metabolite and wood yield is understandable.

There is another, more unified relationship that exists within the data of Tables 2.3a and 2.3b. Where variability in the chemical composition of wood of a specific species is observed, there is typically an inverse relationship between the major carbohydrate and lignin contents. This appears to be the case in these Douglas-fir trees, as total lignin content is positively correlated with the growth canonical variate, while mannose, xylose and glucose content are all negatively correlated. Interestingly, similar (but opposite) correlations can be seen for the metabolite canonical variate in the pools of metabolite precursors to the carbohydrate and lignin polymers. The metabolomics approach applied here has allowed observation of a set of wood formation-related phenotypic and metabolic traits broadly reflecting one another. The observation of broad relationships such as this undoubtedly provides a starting point from which detailed understanding of specific interactions between metabolism and phenotype may be developed.

The relationships demonstrated by the CCA seem to be rooted in the metabolic and phenotypic variation associated with site differences. Almost all of the metabolites that correlated highly in the CCA are high loaders in one or more of the factors responsible for site-related sample clustering and separation in the FA. Furthermore, it was possible to calculate broad-sense heritabilities for half of the high-correlating metabolites in the CCA (Table 2.3a). A similar trend is seen for the phenotypic traits in the CCA (Table 2.3b), where all traits aside from average density, latewood proportion, arabinose and galactose contents load high in at least one of the site-differentiating factors F-1, F-3 and F-4. On average heritabilities were greater than they had been for the metabolites, but in general remained low. These observations all point toward the importance of site over family in the CCA, directly supporting the qualitative, visual evidence provided by the FA factor score plots (Figures 2.1 and 2.2).

In summary, this study demonstrates that broad-scale, non-targeted metabolic profiles of actively metabolizing cambium can be correlated with extensive phenotypic

data that define aspects of tree growth and wood properties in populations of siblings from high-growth performance families of Douglas-fir. Further, a strong relationship between associated metabolic and phenotypic variation and environmental (site) factors exists, while a similar genetics (family) relationship exists, but is comparatively weak.

Additionally, significant correlations were observed between phenotypic indicators of tree growth (diameter at breast height, tree height and volume), cell morphology (microfibril angle, fibre length, fibre coarseness) and cell wall chemistry, and metabolite pools related to major components of cell wall biosynthesis including cellulose (glucose, fructose), hemicellulose (xylose, arabinose, and maltose), and lignin (quinic acid, shikimic acid and coniferin). The existence of linear, quantitative relationships between tree and wood phenotype and wood-forming metabolism, as well as associations between the relative influences of family (genetics) and site (environment) on phenotype and the metabolite pools in actively growing tissue establish a clear biological connection between genetics, metabolism, phenotype and the impact or growing environment. And, as such illustrates the importance of metabolomics within the framework of functional biology, and demonstrates the potential of metabolic data in a unified approach to studying processes involved in tree/plant growth and wood biosynthesis.

Future studies should aim to increase the sampling population (number of families) to better satisfy the requirements of quantitative genetic calculations, as well as replication of site conditions to allow relationships between geo-climatic and biotic factors to be more clearly defined. A notable outcome of this research was the weaker correlation between genetics (*i.e.* family) and metabolic or phenotypic traits. Whether this result accurately reflects the situation in tree populations in general, or was simply due to characteristics of the specific sample population used in this study is not clear. As such, future attempts to demonstrate links between genetic and metabolic factors should look to tree populations that include families which exhibit a wider range of genetic and/or phenotypic diversity, rather than a somewhat narrow selection of “high performance” families, as was employed in the current study. Alternatively, the use of clonal lines in place of full-sib families may be useful in controlling dataset variation, although taking this approach would lead away from any goal of understanding tree and wood development in situations where genetic variability within families exists. Further

resource-intensive, yet potentially enlightening studies could also involve the tracking of wood-forming metabolism in multiple families or clones, under a variety of geo-climatic conditions, throughout the growing season. The metabolic data could then be related back to other biotic and abiotic factors as was undertaken in the current study, to establish a more complete picture of wood-forming metabolism and how it relates to these associated factors. It is, however, apparent that broad scale metabolic profiling of "global" plant metabolism can contribute to our understanding of biological processes in trees/plants or be used to diagnose specific genetic or phenotypic characteristics or responses.

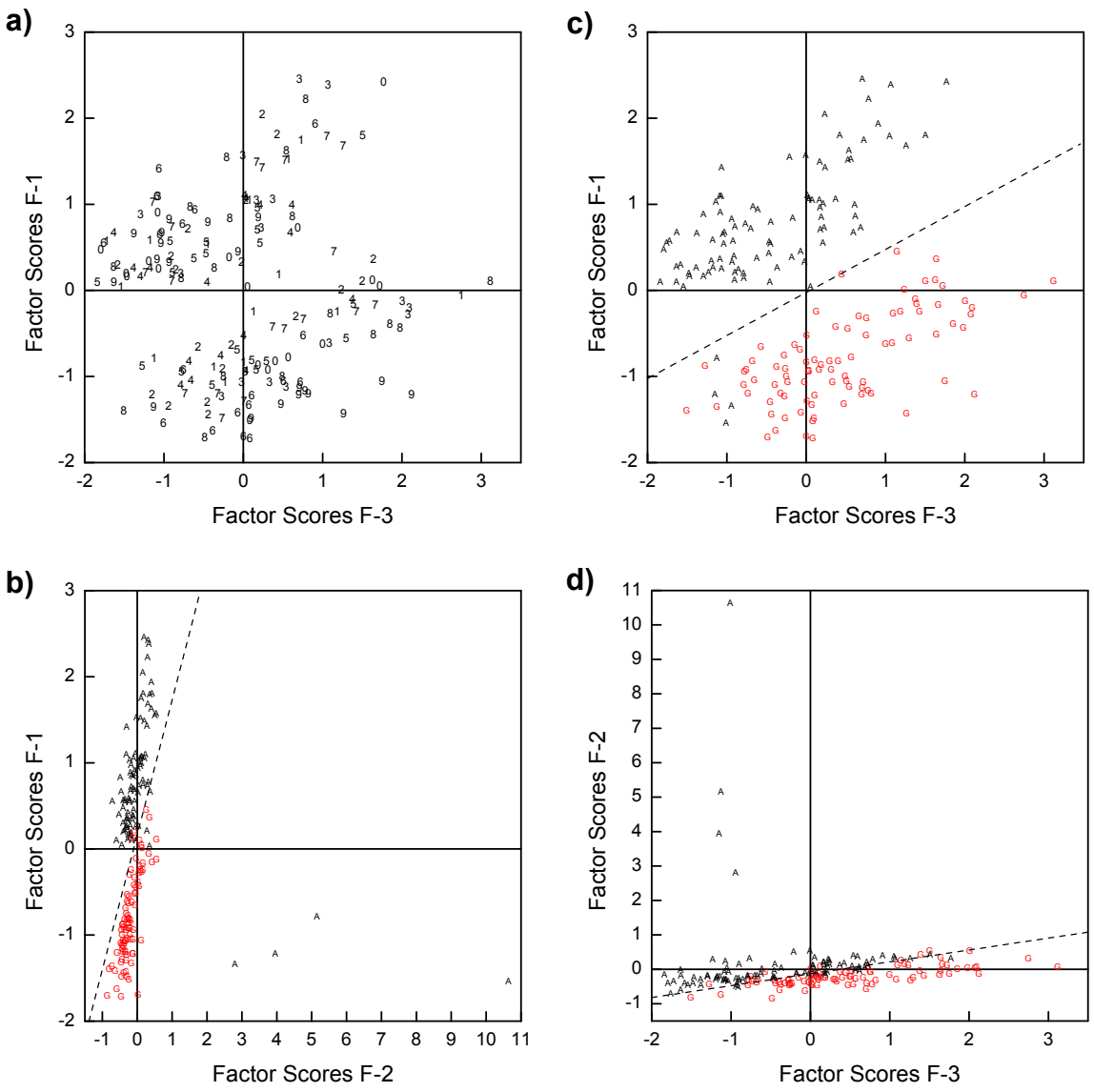


Figure 2.1 Scatter plots of factor analysis (FA) factor scores for metabolite profiles of developing xylem from Douglas-fir trees, with plot axes derived from FA factors 1-3. Analysis represents the differentiation of 181 individual trees (93× AR and 88× GR), across 139 metabolites, and clearly demonstrates the clustering and separation of samples based on site. Dashed lines suggest plane of separation only. a) samples classified by family, representing individuals from families 2, 26, 38, 46, 62, 75, 92, 130, 151, and 156, designated by 0 – 9, respectively. b-d) samples classified by site, "A" indicating Adam River (AR) and "G" indicating Gold River (GR).

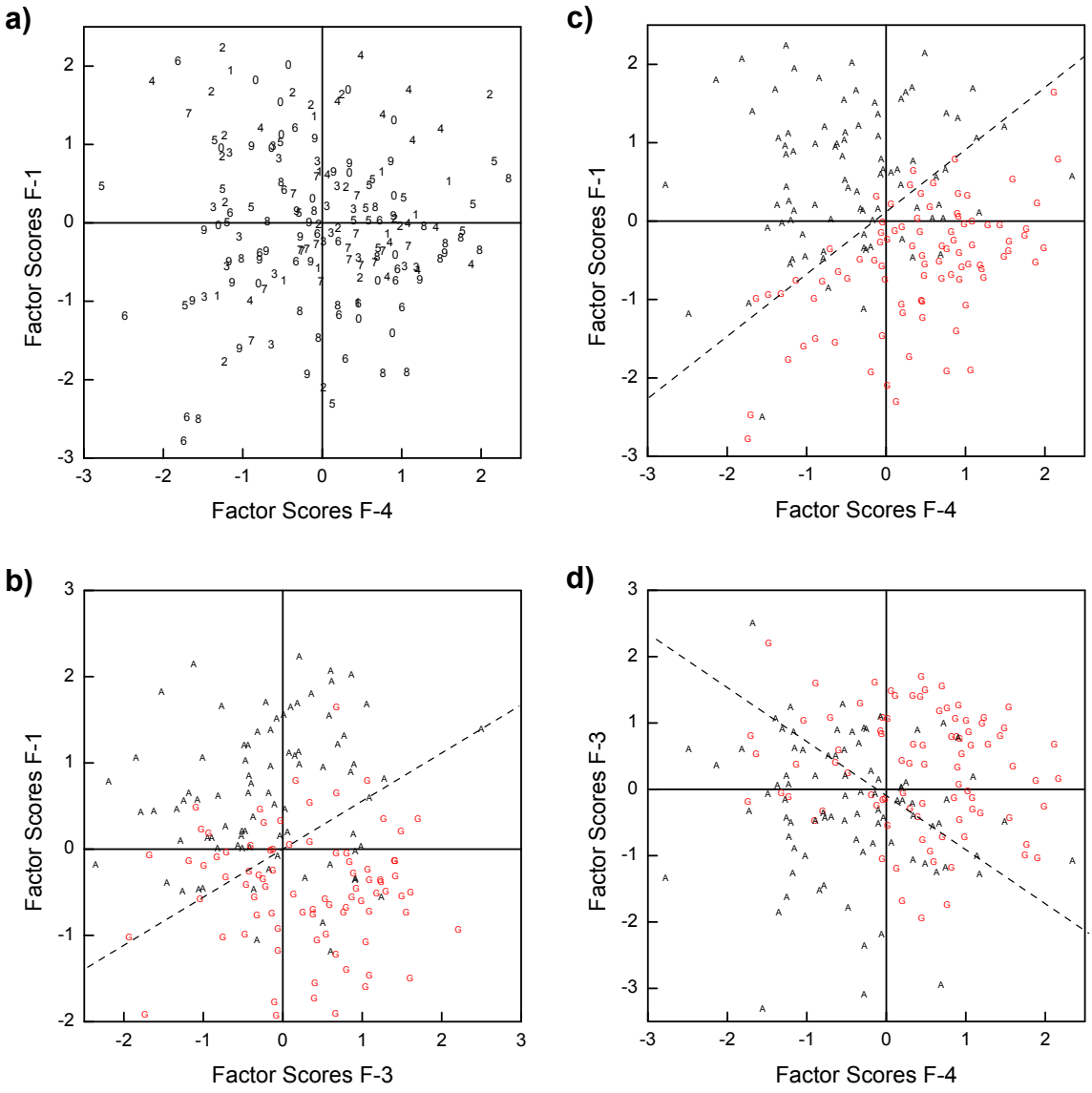


Figure 2.2 Scatter plots of factor analysis (FA) factor scores for quantitative phenotypic traits from Douglas-fir trees, with plot axes derived from FA factors 1-3. Analysis represents the differentiation of 181 individual trees (93× AR and 88× GR), across 16 quantitative phenotypic traits, and clearly demonstrates the clustering and separation of samples based on site. Dashed lines suggest plane of separation only. a) samples classified by family, representing individuals from families 2, 26, 38, 46, 62, 75, 92, 130, 151, and 156, designated by 0 – 9, respectively. b-d) samples classified by site, "A" indicating Adam River (AR) and "G" indicating Gold River (GR).

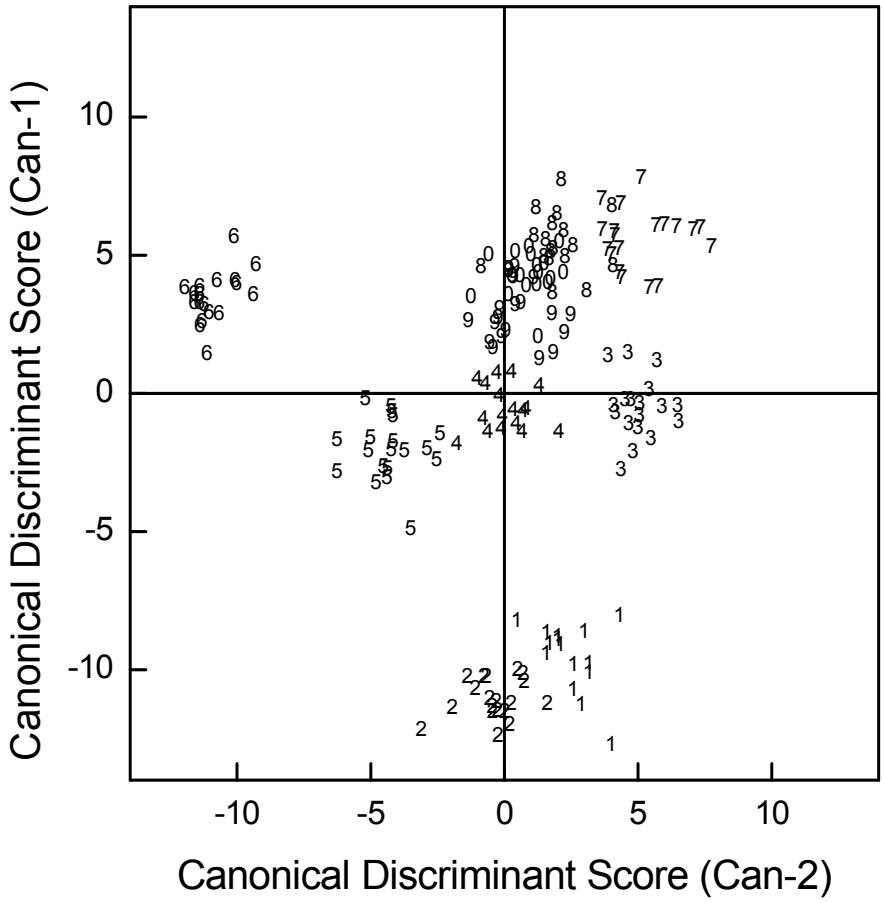


Figure 2.3 Scatter plots of canonical discriminant analysis (CDA) canonical scores for metabolite profiles of developing xylem from Douglas-fir trees, with plot axes derived from canonical factors 1 and 2. Analysis represents the differentiation of 181 individual trees (93× AR and 88× GR), across 139 metabolites, and clearly demonstrates the clustering and separation of samples based on genetics (family). Samples classified by family, representing individuals from families 2, 26, 38, 46, 62, 75, 92, 130, 151, and 156, designated by 0 – 9, respectively.

Table 2.1. Prediction accuracies of multiple discriminant analyses of metabolite profiles of developing xylem from 181 Douglas-fir trees. The "percent accuracy" represents the average frequency with which the discriminant model accurately predicted a) family (out of a possible ten) or b) growth site (out of a possible two) of individual known trees, based on their metabolite profiles (139 metabolites).

a)

	Ave. Prediction Accuracy of MDA by family (%)									
Sites	**F2**	**F26**	**F38**	**F46**	**F62**	**F75**	**F92**	**F130**	**F151**	**F156**
AR and GR	0	17	12	12	17	37	0	25	37	25
Adam River	40	45	70	10	10	20	53	67	12	40
Gold River	37	46	90	70	46	77	70	90	57	70

b)

	Ave. Prediction Accuracy of MDA by site (%)	
Sites	**Adam River**	**Gold River**
AR and GR	80	92

Table 2.2. a) Positively identified metabolites exhibiting significant site variation, for which broad-sense heritabilities could be calculated. b) Broad-sense heritabilities of quantitative phenotypic traits.

a)

Metabolite Information[#]		Heritability[+]
Peak#	Compound ID	H^2
20	Acetic Acid	0.00
54	Acetic Acid, bisoxyl	0.00
60	Phosphoric acid	0.00
92	Alanine, B-	0.00
117	Erythronic acid	0.00
141	Ribose	0.00
169	Pinitol	0.00
173	Quinic acid	0.00
182	Glucose {BP}	0.00
230	Sucrose	0.03
221	Fructose 6P	0.05
175	Fructose	0.06
120	Threonic acid	0.09
73	Glyceric acid	0.10
177	Fructose {BP}	0.11
104	Malic acid	0.11
244	Coniferin	0.13
209	Inositol	0.13
160	Ribonic acid	0.14
74	Fumaric acid	0.15
67	Maleic Acid	0.16
229	Adenosine	0.17
222	Glucose 6P	0.18
223	Glucose 6P {BP}	0.21
178	Glucopyranose	0.23
138	Arabinose	0.31
135	Xylose {BP}	0.34
137	Xylose	0.42

Compound identity determined through mass-spectral and GC retention time matches with standard compounds. {BP} indicates metabolite by-product, as suggested by the Gölm Metabolite Database. + Of the metabolites for which significant site and family variation existed (ANOVA $\alpha = 0.05$) in the absence of site-by-family interaction (*i.e.* G×E effects), allowing for calculation of broad-sense heritability (64 of 139), only metabolites for which it was possible to assign positive identities (28) are presented.

b)

Quantitative trait	Heritability[+]
	H^2
Total Lignin	0.00
Fibre Coarseness	0.00
Fibre Length	0.16
Dia Breast Height	0.20
Latewood Porosity	0.21
Latewood Density	0.22
Tree Volume	0.22
Galactose	0.25
Average Density	0.25
Earlywood Density	0.30
Mannose	0.31
Microfibril Angle	0.34
Xylose	0.37
Tree Height	0.40
Arabinose	0.69
Glucose	1.28

[+] Quantitative traits are sorted according to their heritability score. Significant family-related variation in the absence of site-by-family interaction (*i.e.* G×E effects) was observed in all traits, permitting broad-sense heritability to be calculated for each.

Table 2.3. a) Positively identified metabolites exhibiting significant canonical correlation coefficients, presented in conjunction with factor analysis scores and broad-sense heritabilities values for the same compounds. b) Canonical correlation coefficients of quantitative traits presented in conjunction with factor analysis scores and broad-sense heritabilities.

a)

Metabolite Information[#]		CCA[$]	Factor analyses*			Heritability[+]
Peak#	Compound ID	Metabolite1	F-1	F-2	F-3	H^2
92	Alanine, B-	0.509	0.71			0.00
178	Glucopyranose	0.397		0.65		0.23
54	Acetic Acid, bisoxyl	0.372	0.65			0.00
138	Arabinose	0.363	0.46	0.41		0.31
24	Ammonium	0.353	0.41	-0.35	-0.43	
120	Threonic acid	0.349	0.49			0.09
182	Glucose {BP}	0.346				0.00
135	Xylose {BP}	0.337	0.51			0.34
111	Pyroglutamic acid	0.333	0.54			
20	Acetic Acid	0.323	0.59			0.00
175	Fructose	0.310				0.06
177	Fructose {BP}	0.308				0.11
235	Maltose	0.302	0.43			
179	Glucose	0.300				
173	Quinic acid	-0.304	-0.31	-0.49		0.00
150	Rhamnose	-0.308			0.31	
74	Fumaric acid	-0.373			0.35	0.15
244	Coniferin	-0.459			0.66	0.13
164	Shikimic acid	-0.487	-0.45	0.39	0.56	
73	Glyceric acid	-0.546	-0.52			0.10
169	Pinitol	-0.659	-0.70		0.32	0.00

[#] Compound identity determined through mass-spectral and retention time matches with standard compounds. Compounds sorted by correlation coefficient. Peak# is the unique numerical identity of a metabolite in the 251 compound set originally resolved from chromatographic data. {BP} indicates metabolite by-product, as suggested by the Gölm Metabolite Database. [$] Of the 51 metabolites with significant (>+/- 0.3) canonical correlation coefficients across all 139 metabolites analysed, 21 were positively identified and presented in this table. * For metabolites presented, factor scores in the site-differentiating factors F-1 F-2 and F-3 are presented only where significant (>+/- 0.3). [+] Broad-sense heritabilities were calculated only for metabolites exhibiting significant family and site variation (ANOVA $\alpha = 0.05$) in the absence of family-by-site-interaction (*i.e.* G×E effects).

Continued on following page…

b)

Quantitative trait	CCA[$]	Factor analysis[*]			Heritability[+]
	Growth1	F-1	F-3	F-4	
Dia Breast Height	**0.867**	0.90			0.20
Tree Volume	**0.825**	0.91			0.22
Tree Height	**0.783**	0.91			0.40
Microfibril Angle	**0.575**			0.81	0.34
Total Lignin	**0.484**			0.76	0.00
Arabinose	**0.153**				0.69
Galactose	**0.107**				0.25
Earlywood Density	**0.069**		0.33	0.52	0.30
Latewood Proportion	**0.012**				0.21
Average Density	**-0.076**				0.25
Latewood Density	**-0.295**		-0.69		0.22
Glucose	**-0.309**		0.73		1.28
Xylose	**-0.342**		0.90		0.37
Fibre Coarseness	**-0.412**	-0.49			0.00
Mannose	**-0.418**		-0.80		0.31
Fibre Length	**-0.481**	0.43		-0.52	0.16

[$] Quantitative traits are sorted according to canonical coefficient. [*] Factor scores in the site-differentiating factors F-1, F-3 and F-4 are presented only where significant (>+/- 0.3). [+] Broad-sense heritabilities are presented for each quantitative trait.

2.4 References

Baker, J.M., Hawkins, N.D., Ward, J.L., *et al.* (2006). A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnol. J.* 4, 381-392.

Chen, F., Duran, A.L., Blount, J.W., Sumner, L.W., & Dixon, R.A. (2003). Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* 64, 1013-1021.

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., & Willmitzer, L. (2000a). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157-1161.

Fiehn, O., Kopka, J., Trethewey, R.N., & Willmitzer, L. (2000b). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72, 3573-3580.

Fiehn, O. (2003). Metabolic networks of Cucurbita maxima phloem. *Phytochemistry* 62, 875-886.

Fiehn, O. & Weckwerth, W. (2003). Deciphering metabolic networks. *Eur. J. Biochem.* 270, 579-588.

Griffin, J.J., Ranney, T.G., & Pharr, D.M. (2004). Heat and drought influence photosynthesis, water relations, and soluble carbohydrates of two ecotypes of redbud (Cercis canadensis). *J. Am. Soc. Hortic. Sci.* 129, 497-502.

Herrmann, K.M. & Weaver, L.M. (1999). The shikimate pathway. *Annu. Rev. Plant Phys.* 50, 473-503.

Huhman, D.V. & Sumner, L.W. (2002). Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* 59, 347-360.

Humphreys, J.M. & Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5, 224-229.

Jeong, M.L., Jiang, H.Y., Chen, H.S., Tsai, C.J., & Harding, S.A. (2004). Metabolic profiling of the sink-to-source transition in developing leaves of quaking aspen. *Plant Physiol.* 136, 3364-3375.

Keller, F. & Ludlow, M.M. (1993). Carbohydrate-metabolism in drought-stressed leaves of pigeonpea (Cajanus-cajan). *J. Exp. Bot.* 44, 1351-1359.

Kopka, J., Schauer, N., Krueger, S., *et al.* (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21, 1635-1638.

Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J., & Verhoeyen, M.E. (2003). Metabolite profiling of tomato (Lycopersicon esculentum) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *J. Agric. Food Chem.* 51, 2447-2456.

Morris, C.R., Scott, J.T., Chang, H.-M., Sederoff, R.R., O'Malley, D., & Kadla, J.F. (2004). Metabolic profiling: A new tool in the study of wood formation. *J. Agric. Food Chem.* 52, 1427-1434.

Nigh, G.D. (1997). A growth intercept model for coastal douglas-fir. B.C. Min. For., Res. Br., Victoria, B.C.

Rider, S.D., Hemm, M.R., Hostetler, H.A., Li, H.C., Chapple, C., & Ogas, J. (2004). Metabolic profiling of the Arabidopsis pkl mutant reveals selective derepression of embryonic traits. *Planta* 219, 489-499.

Rizhsky, L., Liang, H.J., Shuman, J., Shulaev, V., Davletova, S., & Mittler, R. (2004). When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiol.* 134, 1683-1696.

Robinson, A.R., Gheneim, R., Kozak, R.A., Ellis, D.D., & Mansfield, S.D. (2005). The potential of metabolite profiling as a selection tool for genotype discrimination in Populus. *J. Exp. Bot.* 56, 2807-2819.

Roepenack-Lahaye, E.v., Degenkolb, T., Zerjeski, M., *et al.* (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol. (Rockville)* 134, 548-559.

Roessner, U., Luedemann, A., Brust, D., *et al.* (2001a). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11-29.

Roessner, U., Willmitzer, L., & Fernie, A.R. (2001b). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127, 749-764.

Samuels, A.L., Rensing, K.H., Douglas, C.J., Mansfield, S.D., Dharmawardhana, D.P., & Ellis, B.E. (2002). Cellular machinery of wood production: Differentiation of secondary xylem in Pinus contorta var. latifolia. *Planta (Berlin)* 216, 72-82.

Sonesson, J. & Eriksson, G. (2000). Genotypic stability and genetic parameters for growth and biomass traits in a water x temperature factorial experiment with Pinus sylvestris L. seedlings. *Forest Sci.* 46, 487-495.

Srere, P.A. (1987). Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.* 56**,** 89-124.

Szopa, J. (2002). Transgenic 14-3-3 isoforms in plants: The metabolite profiling of repressed 14-3-3 protein synthesis in transgenic potato plants. *Biochem. Soc. Trans.* 30**,** 405-410.

Tabachnick, B.G. & Fidell, L.S. (2001). Using multivariate statistics. Allyn & Bacon, Boston.

Urbanczyk-Wochniak, E. & Fernie, A.R. (2005). Metabolic profiling reveals altered nitrogen nutrient regimes have diverse effects on the metabolism of hydroponically-grown tomato (Solanum lycopersicum) plants. *J. Exp. Bot.* 56**,** 309-321.

CHAPTER 3

Metabolite profiling reveals complex relationship between developing xylem metabolism and intra-ring internal checking in *Pinus radiata*

A version of this chapter is to be submitted for publication. Robinson, A.R., Ukrainetz, N.K., Samuels, A.L., and Mansfield, S.D. Metabolite profiling reveals complex relationship between developing xylem metabolism and intra-ring internal checking in *Pinus radiata*.

3.1 Introduction

The forest industry has traditionally relied on natural forest resources as its primary source of wood; however, due to long-term depletion and intensive harvesting, recent efforts have witnessed a shift towards plantation-grown forests as alternative wood sources. Unfortunately, plantation-based forestry, involving mass cultivation of select genotypes, has brought about wood quality issues that are only beginning to become apparent, as early generations of plantation forests have matured and been processed.

Intra-ring (a.k.a. within-ring) internal checking is a structural wood quality defect that limits the suitability of wood for high-value or value-added applications. The phenomenon is prevalent in the faster-grown, younger-pruned trees from cultivated crops of softwood species such as *Pinus radiata* (radiata pine), and consequently can have a significant negative impact on the commercial value of the ensuing lumber. In radiata pine, the intra-ring internal checking phenotype is characterised by the formation of longitudinal cracks and voids within the earlywood portion of sapwood annular growth rings and is more prevalent in rings adjacent to the heartwood/sapwood boundary. These cracks typically occur during rapid kiln drying, but in severe cases may already be present in green wood (Ball *et al.*, 2005; Williams, 1981). It has been proposed that the development of internal checks during drying is related to interactions between water and the cell wall matrix, and the subsequent exposure of the matrix to tensional stresses (Booker *et al.*, 2000). It appears that certain physico-chemical cell wall properties can predispose cells to checking, with the initial failure of the cell wall at the interface of the compound middle lamella and the S_1 layer of the secondary cell wall culminating in a check (Putoczki *et al.*, 2007). Specifically, it has been proposed that a loss of striations in the S_1 wall layer accompanies an increased incidence of checking, which may be associated with locally altered microfibril organisation and lignin distribution that weaken the wall structure (Donaldson, 1995; 1997; Putoczki *et al.*, 2007). Less well understood are the genetic, gene regulatory, and metabolic factors that underlie and contribute to this predisposition. There is good evidence, however, that genetics are a major contributing factor to checking susceptibility, as the trait exhibits fairly strong heritability (Ball and McConchie, 2001; Kumar, 2004).

It is recognised that an accurate method for predicting the potential for wood to check, either at the breeding/sapling stage, immediately pre-harvest, or prior to post-

harvest processing, would be an invaluable tool in the plantation forestry industry. The possibility of this prompted a SilviScan-based assessment of wood properties for checking individuals, and the subsequent development of a generalised linear model in which tracheid radial diameter and cell wall thickness could be used to accurately predict checking status for these individuals (Ball *et al.*, 2005); however, this method required destructive harvesting, making it less ideal for progeny test scenarios in breeding trials.

The research described herein tested a metabolomics approach to elucidate the metabolic elements of xylem biosynthesis that may be related to the predisposition for wood to check in families of radiata pine. Integral to this goal was the use of these distinctive metabolic elements to distinguish between trees exhibiting distinct levels of checking severity, and to predict the severity of the internal checking phenotype in a non-destructive manner. Furthermore, the potential role of a key metabolite, coniferin, in the manifestation of the internal checking defect was investigated and its association discussed.

3.2 Materials and methods

3.2.1 Plant material and sampling

The radiata pine tree population sampled in this research consisted of seven-year-old full siblings from multiple families grown in the Puriki trial forest near Rotorua, New Zealand. These families exhibited a range of severity in an "internal checking" wood phenotype, determined by previous analysis of other related siblings. These families were taken from the same field trial as used in other contemporary checking-related research (Booker *et al.*, 2000; Putoczki *et al.*, 2007). Sampling was conducted during the early growing season, in late October 2004. Samples were acquired in random order between 10 am and 3 pm, under overcast conditions. A sample of developing xylem tissue was obtained from the north-facing side of each tree by first peeling a section of bark/phloem/outer cambium from the main bole of the trunk at ~50 cm from the base, and then scraping the inner cambium with a clean razor blade. The collected material was immediately transferred to a cryovial, snap-frozen in liquid nitrogen, and stored at -80°C until further processing.

3.2.2 Metabolite sample preparation

Frozen tissue was macerated to a fine powder with a 15 s burst using a dental amalgam mixer, employing a liquid N_2-chilled copper/plastic capsule and steel ball bearings. Samples were kept frozen at all times and, once ground, were returned to -80°C.

Metabolites were extracted from tissue samples and prepared for GC/MS using a two-phase methanol/chloroform method developed for metabolite extraction from *Populus* cambium and developing xylem (Robinson *et al.*, 2005). Approximately 60 mg frozen, ground developing xylem was accurately weighed into a pre-chilled 2 mL lock-cap centrifuge tube. To this, 600 µL HPLC-grade methanol (CH_3OH) was immediately added and vortexed for 10 s to halt biological activity and minimise degradation. In addition, 40 µL distilled, deionised water (H_2O) and 10 µL internal standard mixture (10 mg/mL ribitol in H_2O) were added. The sample was then incubated for 15 min at 70°C with constant agitation, and centrifuged at 13 000 rpm for 5 min. The supernatant, containing extracted metabolites, was retained. $CHCl_3$ (800 µL) was then added to the pellet, vortexed for 10 s to re-suspend, and incubated for 5 min at 35°C with constant agitation. The resultant supernatant recovered following a second 5 min centrifugation at 13 000 rpm, was pooled with the supernatant from the initial CH_3OH extraction. H_2O (600 µL) was added to the combined supernatant, vortexed for 10 s, and then centrifuged for 15 min at 4000 rpm to permit the separation of polar (methanol/water) and non-polar (methanol/chloroform) phases. This combination and separation of phases allowed metabolites extracted in one phase but with greater affinity for the other to repartition. The polar (upper) methanol/water phase was taken, and either processed immediately or stored at -20°C until further analysis. Metabolites in the non-polar phase were not analysed in this study. The soluble polar metabolite samples were derivatised prior to analysis by GC/MS. An aliquot (900 µL) of the methanol/water phase was dried using a Vacufuge (Eppendorf) (3-4 h, 30°C), and methoxylated by re-suspending the pellet in 50 µL methoxyamine hydrochloride solution (20 mg/mL in pyridine) and incubating with constant agitation for 2 h at 60°C in order to protect carbonyl moieties. Acidic protons were then trimethylsilylated with 200 µL N-methyl-N-trimethylsilyltrifluoro acetamide (MSTFA) and incubated at 60°C with constant agitation for 30 min. Samples were left to stand at room temperature overnight to ensure complete derivatisation, and then filtered through compacted tissue paper prior to GC/MS analysis.

For HPLC, approximately 100 mg frozen, ground developing xylem was weighed accurately into a pre-chilled 2 mL lock-cap centrifuge tube. To this, 2 mL extraction solvent (48.5%vv CH_3OH, 48.5%vv H_2O, 1.5%vv glacial acetic acid) was added. The extraction was allowed to proceed for 5 h at 40°C with constant agitation at 200 rpm. Samples were then centrifuged at 13 000 rpm for 10 min, after which the supernatant was transferred to a clean centrifuge tube. For HPLC, samples were concentrated 5-fold, reducing 1 mL supernatant to <200 µL in a Vacufuge (Eppendorf), and then making it back up to 200 µL with CH_3OH.

3.2.3 HPLC-based analysis

Analysis of metabolite extracts by HPLC was conducted with a Dionex Summit HPLC fitted with a Phenomenex Spherosill ODF C18 250 mm × 4.6 mm column and a UV/VIS photodiode array detector. A 35 µL aliquot of the methanol/water/acetic acid extraction described above was injected, and subjected to the chromatographic separation. Gradients were based on mixtures of eluents "A" (0.1% trifluoroacetic acid in water) and "B" (0.1% trifluoroacetic acid in 75:25 methanol:acetonitrile mix). Zero min (95% "A", 5% "B"), gradient to 38 min (40% "A", 60% "B"), gradient to 40 min (100% "B"). Ten min column wash (100% "B"), followed by rapid return to 95% "A", 5% "B" in preparation for the next run. The identity of the coniferin peak was confirmed by the comparison of the retention time and UV spectrum with a confirmed chemical standard compound. Samples were analysed in random order. Peak areas were normalised against the amount of tissue included in each extraction, and the average relative abundance of coniferin in samples from low and severe checkers compared by two-tailed t-test ($\alpha = 0.05$).

3.2.4 GC/MS-based analysis

3.2.4.1 GC/MS conditions

GC/MS analysis was conducted on a ThermoFinnigan Trace GC-PolarisQ ion trap system fit with an AS2000 auto-sampler and a split/splitless injector. The GC was equipped with a low-bleed Restek Rtx-5MS column (fused silica, 30 m, 0.25 mm ID, stationary phase diphenyl 5% dimethyl 95% polysiloxane). The GC conditions were set as follows: inlet temperature 250°C, helium carrier gas flow at constant 1 mL/min, injector split ratio 10:1, resting oven temp 70°C, and GC/MS transfer line temperature 300°C. Following injection of a 1 µL aliquot of sample, the oven was held at 70°C for 2

min and then ramped to 325°C at a rate of 8°C/min. The temperature was held at 325°C for an additional 6 min before being cooled rapidly to 70°C in preparation for the next run.

Mass spectrometry analysis was conducted in positive electron ionisation (EI) mode, the fore-line was evacuated to approximately 40 mTorr, with helium gas flow into the chamber set at 0.3 mL/min. The source temperature was held at 250°C, with an electron ionisation potential of 70 eV. The detector signal was recorded from 3.35 min after injection until 35.5 min, and ions were scanned across the range of 50-650 mass units (mu) with a total scan time of 0.58 s.

3.2.4.2 Data acquisition and processing

ThermoFinnigan Xcalibur v1.3 software was used for both GC/MS data collection and peak determination and measurement. GC/MS total ion chromatograms (TIC) of TMS-derivatised extracts were recorded for all samples, in order to elucidate the common "metabolite pools" present in the developing xylem tissue of each tree. To normalise raw TIC peak data, metabolite peak areas were expressed relative to that of the ribitol internal standard peak, and then further standardised across all chromatograms by adjusting for the precise amount of tissue (mg fresh weight) used in each sample extraction. A compiled peak data set was generated through semi-automated alignment of peaks that represented the same compound in multiple chromatograms by the purpose-built 'PeakMatch' software (Robinson *et al.*, 2005). As a means of minimising artefacts caused by sample processing and erroneous non-detection of small metabolite peaks proximal to baseline noise, all peaks that did not appear in at least 5% of the samples were removed from the dataset.

3.2.4.3 Data reduction by univariate analysis

Reduction of the metabolite peak set to include only metabolites that showed differences between trees having different severity in the internal checking phenotype was achieved using Bonferroni F-tests. This test constitutes the core of ANOVA, and allows the comparison of multiple means while taking into account the degrees of freedom (*i.e.* the number of means being compared). A peak that shows significance in this test is considered to be different in at least one of the submitted classes. Various thresholds for significance were set in these tests, as indicated in the results.

3.2.4.4 Multivariate statistical analyses

Other reductions of the metabolite data were carried out by Multiple Discriminant Analysis (MDA), and Principal Components Analysis (PCA) using the 'proc discrim' and 'proc factor' procedures of the SAS v9.1 software (SAS Institute, Inc., Cary, N.C.), respectively.

Multiple discriminant analysis, a statistical approach that assesses the variation in pre-classified multivariate data and is capable of generating predictive models, was applied to the metabolite data array. To generate a cross-validated analysis, the data in each class were split into four equal (number of samples) subsets. Then, three of the four subsets in each class are used to build the model, and the fourth used as an independent validation array to assess its predictive accuracy. This building/testing process was repeated four times using the different possible combinations of "builders" and "testers", and a final predictive accuracy rate for each class calculated as the average over the four models. For a dataset having three classes, prior probabilities of 0.333 (33.3%) are expected. For two-class data these probabilities are 0.5 (50%). Higher model accuracy than the prior probability implies that the MDA is able to distinguish between classes better than would be the case in random prediction.

Principal components analysis (PCA) allows the variation in metabolite and quantitative trait data arrays to be explored without the constraints of data pre-classification. Factor scores of individual samples from selected principal components were plotted as co-ordinates on the axes of two-dimensional scatter plots. This generates a graphical representation of the variation in the original data captured by the analysis, and of the relationship between individual samples. In metabolomics analyses, the separation of sample clusters in such plots is considered to illustrate differences between distinct metabolic systems, (Chen *et al.*, 2003; Fiehn *et al.*, 2000; Fiehn and Weckwerth, 2003; Roessner *et al.*, 2001a; Roessner *et al.*, 2001b).

3.2.4.5 Compound identification

National Institute of Standards and Technology (NIST) MS-Search software equipped with the NIST mass spectra, as well as the Max Planck Institute Trimethylsilane (TMS) (http://www.mpimp-Golm.mpg.de/mms-library/index-e.html), Gölm Metabolome Database (http://csbdb.mpimp-Golm.mpg.de/csbdb/gmd/gmd.html) (Kopka *et al.*, 2005)

and our own (Mansfield UBC laboratory) TMS mass spectral libraries were collectively used to identify metabolites of interest, as highlighted by the statistical analyses.

3.2.5 Scanning electron microscopy

Air-dried mature secondary xylem was dissected transversely with razor blades and wood samples were attached to scanning electron microscope (SEM) stubs using double-sided stick tape. Following gold coating, samples were viewed using a Hitachi S7600 at 3 kV and images captured digitally.

3.3 Results and discussion

3.3.1 Sampling, data acquisition and pre-processing

Samples of developing xylem were collected from five, seven-year-old full-siblings from each of 24 families of radiata pine; a total of 120 trees. Of those families sampled, 8 were defined as non or very low checkers (0-13 checks per cross section), 8 medium level checkers (30-55 checks), and 8 high level checkers (90-140 checks), as previously determined by destructive harvests, drying and checking counts in basal cross sections of other related siblings. Taking sample losses during processing into account, the final set of analysed samples consisted of 40 low, 39 medium and 40 high checking individuals. Once GC/MS data were collated, the complete dataset consisted of 228 distinct compound peaks across 119 samples (an array of ~27 000).

At this point, it bears mentioning that in the tree families studied, the internal checking phenomenon has arisen in the course of intensive breeding for tree form and growth rate, rather than as a result of transgenic induction. Additionally, the individual tree subjects were siblings, rather than line clones. As a consequence, the basis of the internal checking trait may very well be polygenic, while the phenotypic severity trait is unquestionably continuous in nature. In any case, there is an obvious phenotypic distinction between the wood of non-checking and high-checking individuals (Figure 3.1).

3.3.2 Analysis of GC/MS metabolite profiles

Data mining using statistical analyses was conducted on the compiled metabolite array with the intention of a) distinguishing the metabolism of tree families displaying a range of checking severity, and b) developing a model(s) capable of predicting checking severity, on the basis of metabolite profile data. Statistical analyses were carried out on

the complete data array, as well as on array subsets in which the number of checking classes and/or metabolite variables were reduced by logical statistical means.

3.3.2.1 Complete metabolite profiles

The initial approach was to subject the data array to principal components analysis (PCA), a multivariate tool commonly applied in metabolomics that can be used to visualise the distribution of metabolic variation within sample sets, by sample. In a PCA involving all three checking classes (high, medium and low) and all 228 variables, the best distinction between classes was seen with principal components 1 and 2 (PC-1, PC-2), which accounted for 25% and 16% of the variance in the dataset, respectively (Figure 3.2a). However, the clustering and separation of classes was loose and incomplete. A similar PCA involving the same metabolite variable set, but only samples in the extreme classes (high and low), yielded improved clustering and separation with PC-4 and PC-5 (accounting for 6% and 5% of the variance, respectively), although this also remained incomplete (Figure 3.2b). The results from PCA indicated that differences do exist between the metabolisms of families exhibiting different levels of checking severity, however, the resolution of this was only moderate when profile data were analysed in their entirety, by this means.

Multiple discriminant analysis (MDA) was then implemented as a tool for modeling and predicting checking class on the basis of metabolite profiles. Cross-validated MDA models yield an average error rate for the predictive classification of samples in each class, and this may be used as an indicator of the overall accuracy of the model. Initially, MDA prediction models were generated using the complete set of 228 metabolite variables (Table 3.1). The model built around all three classes was ~85% accurate. It is notable however, that the accuracy was not consistent across all classes, and that greater error was seen in classifying low and medium checkers. Three additional models were generated using the complete set of metabolites, each including only two of the three checking classes (Table 3.1). Accuracy was also high in these models, at ~90% overall.

Although predictive accuracy was generally high in the two-class models, the greater accuracy of the high-low model, the lesser accuracy of medium class prediction in the high-medium model, and the much poorer overall accuracy of the low-medium model all suggest that MDA had difficulty classifying the medium checkers. These

specific results suggest that the metabolite profiles (and, by extension, metabolisms) of medium-checking genotypes resemble those of genotypes in the low checking class. However, additional trends are also apparent in the results, which complicate interpretation. The increased accuracy of the high checker prediction in the high-medium model, compared to the high-low model, suggests that there is greater similarity between low and high checkers than there is between medium and high. To confound this, the decrease in the accuracy of low prediction in the medium-low model compared to the high-low model suggests that there is greater similarity between medium and low checkers than there is between high and low.

The most attractive interpretation of the apparently conflicting results in the two-class models is that there are some strong common elements between high and low checkers, and other also strong elements in common between medium and low checkers. Although it would be convenient if the relationships between patterns in metabolite abundances and the severity of the checking phenotype were universally simple, these results suggest that this is unlikely. It is also important to note that the three classifications used in this study were based on arbitrary criteria (ranges of check number in the cross section) that attempted to represent continuous data as discrete. The actual ranges used for classification likely have an impact on the statistical analysis. It should not, then, be surprising that the MDA models had the highest error rate when predicting medium checkers' identity, nor that the two-classification models perform a little better than the model that included all three.

3.3.2.2 Reduced metabolite profiles

A logical progression was to investigate whether it was possible to reduce the number of metabolite variables used to generate MDA models, yet retain accuracy. Although the accuracy of models generated using the complete metabolite profiles was excellent given the limited sample size and the complexity of the biological system, it was felt that a subset of the metabolites might have played a primary role in distinguishing between checking phenotypes in the MDAs. Therefore, statistical tests were employed to select metabolites believed to play important roles in the discriminant models. The datasets generated were tested in the three-class MDA model as well as the three, two-class models, to assess the effectiveness of the reductions. However, in the interests of

clarity and owing to their greater relevance, only the results from the three-class and high-low two-class models will be presented and discussed.

For the three-class model, F-tests (part of ANOVA) at a given threshold were used to identify peaks that changed significantly between at least one of the three sample classifications (99% and 99.9% confidence are represented by α values of 0.01 and 0.001, respectively). Those that showed significant differences were included in the reduced dataset and subsequently included in the MDA model. The accuracy of the prediction model suffered with sequential reduction, particularly in the case of the medium checkers (Table 3.2). Again, the considerable difficulty that the model had with this class is indicative of the true continuous nature of checking severity data that has been classified as discrete.

The process of logical reduction for the high-low two-class model was the same as for the three-class model except that t-tests instead of F-tests were used to identify peaks that were significantly different between the two sample classifications (95%, 99%, and 99.9% confidence are represented by α values of 0.05, 0.01 and 0.001, respectively). In these MDAs, the accuracy of prediction remained high, and was maintained at ~90% under an extreme reduction to 16 of the most clearly different metabolites (Table 3.2 and Figure 3.3). The continued performance of these analyses is a clear indication that the critical, distinctive aspects of the metabolite profiles had been retained, despite the disposal of a large portion of the original data. The identification of these metabolites as highly differential between high and low checkers, and the high accuracy of the ensuing MDA model both suggest that the metabolites contained in this small set may be closely related to the generation of the internal checking phenomenon.

Attempts to identify the set of metabolites resulted in positive identities for five, and tentative molecular class assignment for another five of the total 16 (Table 3.3). Succinic acid is a participant of the tricarboxylic acid cycle, which, in the case of developing xylem, generates usable energy from translocated carbohydrates as well as precursors for many amino acids. The hydrophilic amino acid, serine, participates in the biosynthesis of purines and pyrimidines and is also a precursor of several other amino acids. Several carbohydrate-based molecules were also found in the list, although only inositol could be positively identified. Additionally, several phenolic metabolites with

links to phenylpropanoid metabolism and lignin biosynthesis were identified. Shikimic acid is a participant in the shikimate metabolic pathway, which is ultimately responsible for funnelling carbon toward phenylpropanoid metabolism and the lignin-specific biosynthetic pathway (Herrmann and Weaver, 1999; Humphreys and Chapple, 2002). Also of interest is an unidentified derivative of quinic acid. Although this metabolite is not large enough to be the *p*-coumaroyl or caffeoyl quinate esters involved in phenylpropanoid biosynthesis (Franke *et al.*, 2002; Hoffmann *et al.*, 2003; Schoch *et al.*, 2001), it could potentially be associated with the generation of quinic acid via the shikimate pathway, or the cycling of quinic acid in and out of phenylpropanoid esters pools. Finally, 4-hydroxy benzoic acid, a breakdown product of hydroxycinnamic acids or, indirectly, their -CoA derivatives, is another identified metabolite that participates in the 'fringe' aspects of phenylpropanoid metabolism. In light of these results, it is apparent that several molecule classes are represented in the list of sixteen differential metabolites, indicating that a propensity to check is associated with physico-chemical wood properties arising via the interaction of elements from distinct aspects of the cellular metabolism for wood formation.

A simple test was conducted to assess coherency between MDA and PCA under the conditions of logical reduction. The reduced peak set that gave the highest accuracy in MDA for both the three-class (106 peaks) and high-low two-class (16 peaks) models were subjected to PCA, and the two components that showed greatest clustering and separation of checking classes were plotted for each (Figure 3.4). In the three-class analysis, PC-3 and PC-5 were selected (accounting for 8% and 4% of total variance, respectively), whereas in the high-low two-class analysis PC-1 and PC-2 were selected (accounting for 30% and 15% of total variance, respectively). It was immediately apparent that the cluster patterns observed in PCA reflected the performance of the MDA. For the three-class analysis, clustering and separation of the different classes remained incomplete (Figure 3.4a), although there is a gradient from high to low checkers that is sensible in terms of a graduated phenotype. The high and low checking genotypes almost completely separate, while medium checkers are scattered, and this provides a visual description as to why the three-class MDA models were comparatively inaccurate. For the high-low two-class analysis, all but complete

distinction between checking classes was observed (Figure 3.4b), and this was a result that reflected the extremely high accuracy of the MDA.

The use of MDA modeling in applied, real-world screening situations would involve building an accurate prediction model using a set of samples of known checking class, and then using this model to predict the checking class of samples of unknown phenotype. Therefore, as a simple assessment of the performance of MDA models in this type of situation, the 16 metabolite high-low two-class model was used to predict the checking phenotype of the set of 39 medium checkers. Under this model, 30 of the 39 samples were classified as low checkers, and the certainty of classification was generally very high (average 92%). This classification spread is in general agreement with the results in the two-class MDAs based on complete metabolite profiles (Table 3.1), which also indicated that in terms of metabolite profiles, medium checkers are more similar to low checkers than they are to high checkers. However, as discussed previously, this pattern is likely an artefact of the classification criteria used during evaluation. 'Medium' checking has a range that errs heavily on the low side of the classification scheme, so it is not that surprising that on average the medium samples were found to be more similar to the low than the high.

3.3.3 Reflection on structure of phenotypic data

In this research, logistical constraints dictated that the phenotypic data could only be collected and analysed in a discrete, classified form. The analysis of the continuous internal checking variable as if it were discrete clearly had drawbacks. The most significant of these was that the limits of phenotypic severity included in each class approached those included in the fringe of the neighbouring class(es). It transpired that adjoining classes likely shared metabolic properties to some extent, which confounded efforts to consistently distinguish between samples in neighbouring classes via MDA of metabolite profile data, and precluded the possibility of cluster separation in the three-class PCA's. Ideally, further metabolomic analysis of the internal checking trait will make provisions for the collection of continuous quantitative data based on the individual tree (rather that a "per family" class-based assessment), which will open other avenues of multivariate statistical analysis such as canonical correlation analysis (CCA), partial least squares regression (PLSR), and stepwise modeling.

3.3.4 The relationship between coniferin and the internal checking phenotype

The metabolite coniferin is implicated in the process of cell wall development as the proposed glucosylated storage/transport form of coniferyl alcohol (Samuels *et al.*, 2002; Savidge, 1989), which is itself the structural precursor of guaiacyl lignin (Humphreys and Chapple, 2002). The observation of localised, subcellular changes in lignin composition in checking-prone cell clusters hase been reported (Donaldson, 1995; Putoczki *et al.*, 2007), and given the heavy bias towards guaiacyl lignin in coniferous species, it seems possible that the metabolism of coniferin could be playing a role in the manifestation of the checking phenotype.

Metabolite extracts of developing xylem from 30 trees from low checking families, and 28 trees from severe checking families were analysed by HPLC. Student's t-test of normalised peak area confirmed a statistically significant increase in the average abundance of coniferin in the samples from severe checking families. The proportional increase over low checkers was 0.16 on average which, although mild, was significant at $\alpha = 0.05$. This apparent increase in the pool of coniferin, in developing xylem from families susceptible to severe checking, could be interpreted as evidence of some sort of "block" or reduced efficiency in downstream mechanisms of the lignin biosynthetic pathway; however, this would not be in agreement with the determination by Putoczki *et al* (2007) that, overall, total lignin content was not reduced in the wood of checking-prone individuals.

Instead, the study of scanning electron micrographs showing the fine cellular detail of check structure encourages another interpretation. It appears that checks consistently originate from ray cell files, and expand longitudinally from that origin (Figures 3.5 and 3.6). Rays in wood of radiata pine are of both the uniseriate (occasionally part biseriate) and fusiform (integrating epithelial cells and a resin canal) types, which are typically 1-12 and 1-21 cells in height, respectively (Maddern-Harris, 1991). The parenchyma cells that constitute ray tissue are actively involved in the transport and storage of metabolites in the sapwood structure, including compounds such as starch and, during cambial activity in gymnosperms, coniferin (Savidge, 1989). It seems possible that rays may constitute a weak point in the wood structure, so that increased ray density would be associated with increased checking, in situations where wood composition promotes the defect. In such case, the putative increase in ray

density would be responsible for the increase in coniferin concentration observed in high checking individuals, and the metabolism and fate of coniferin itself may not be directly involved in generating the physico-chemical wood properties that lead to internal checking. In order for this claim to be substantiated, an assessment and comparison of ray density in low and high checkers will be required.

3.4 Concluding remarks

It is clear that a relationship exists between the metabolism of wood formation and the internal checking phenotype in juvenile radiata pine. Both of the multivariate analytical techniques employed were able to distinguish between the metabolite profiles of trees having different levels of internal checking severity, with MDA predicting the checking severity class of individual trees considerably better than would be the case with random assignment, and PCA differentiating between the low, medium and severe checking classes. Additionally, the combined evidence that the concentration of coniferin in the developing xylem of high checking individuals is greater than in low checkers, and that checks appear to originate at the ray files where coniferin is stored, suggests a role for ray density in the propensity to check, and therefore only an indirect association between coniferin concentration and furthermore phenylpropanoid metabolism, and this wood defect.

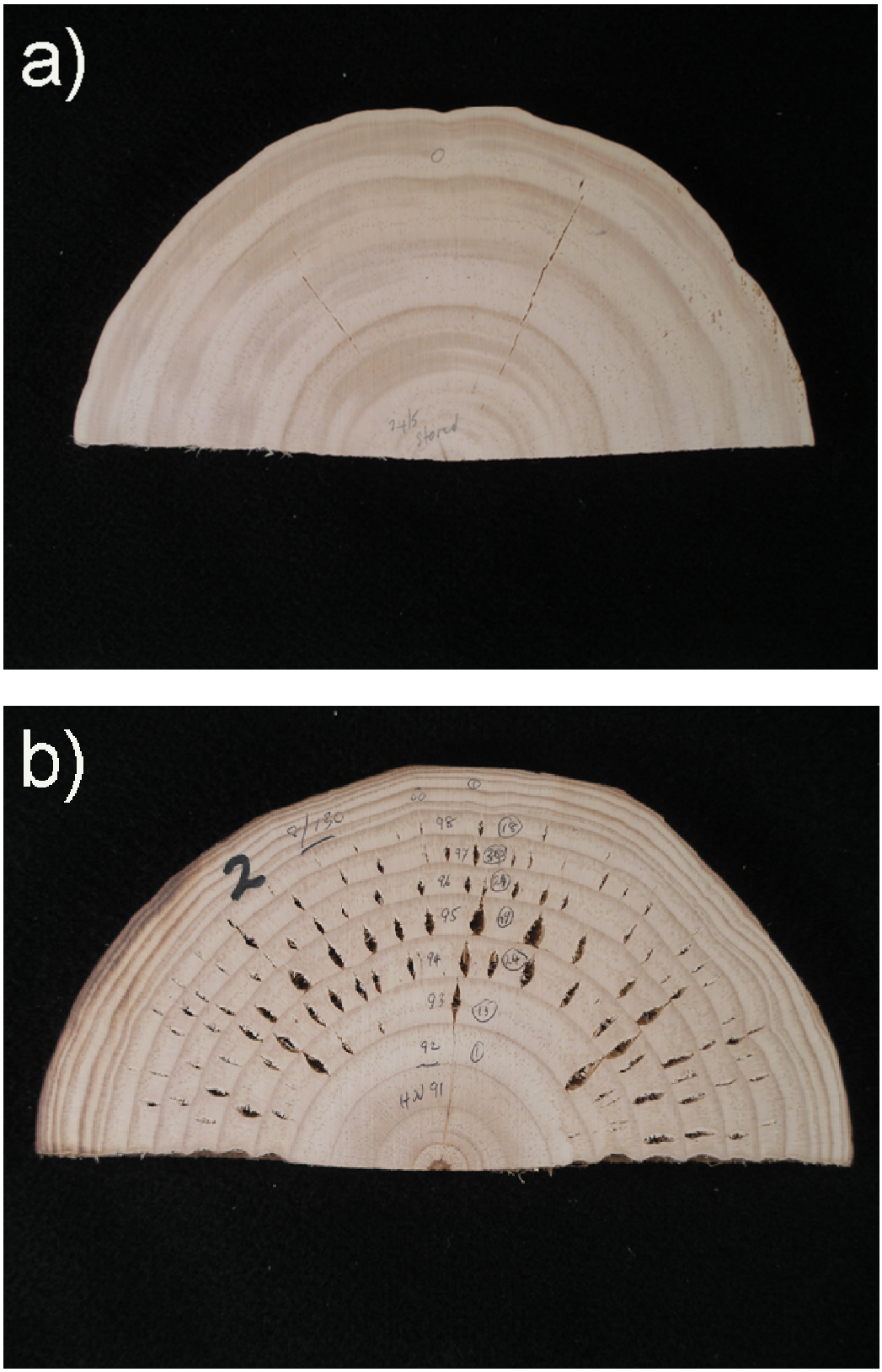


Figure 3.1 Radial cross-sections of juvenile radiata pine post-drying from: a) a non-checking individual, and b) a high-checking individual.

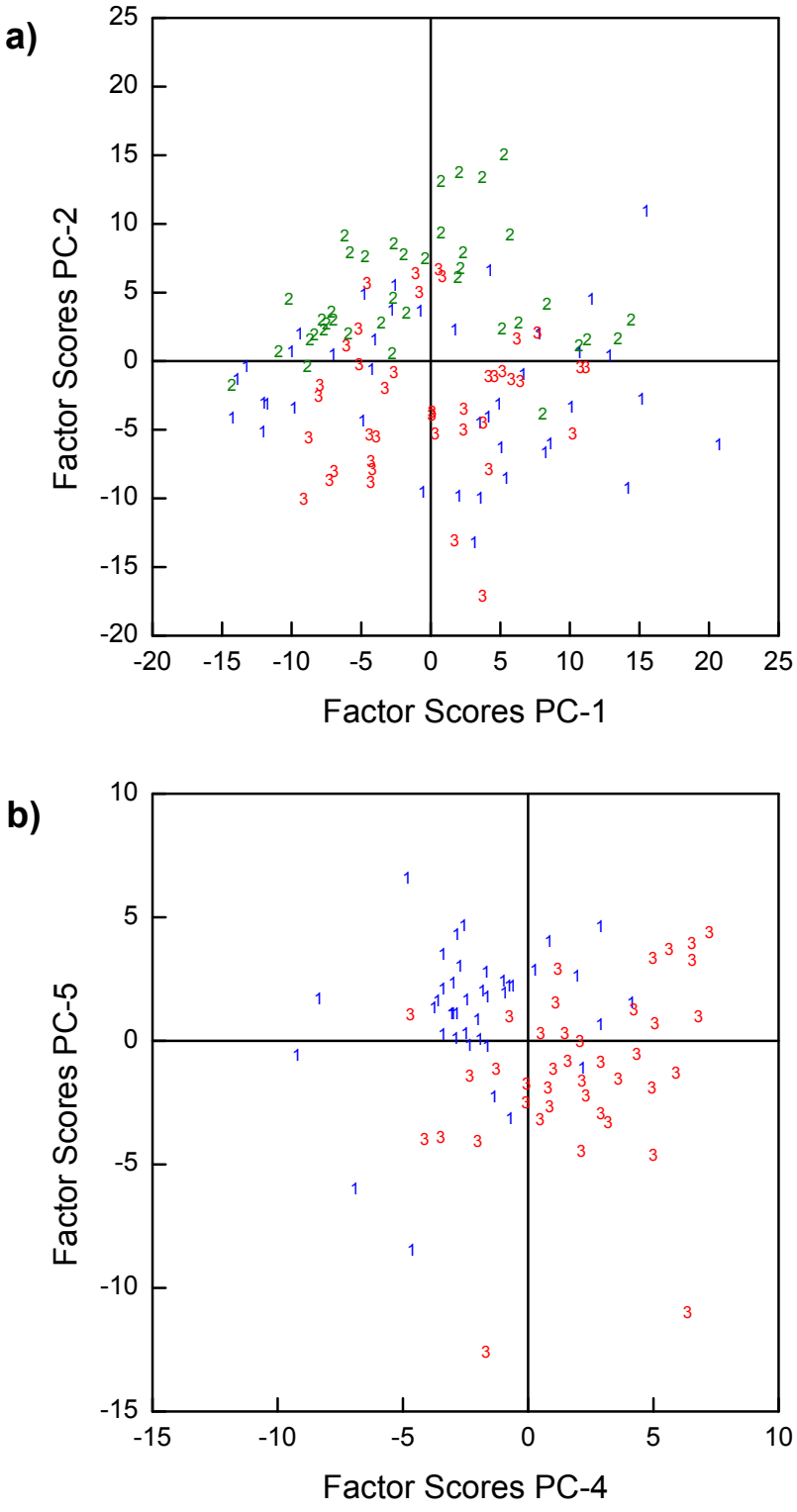


Figure 3.2 Factor score plots from PCA of complete metabolite profiles (228 metabolites) for: a) high, medium and low checking families (119 individuals), and b) high and low checking families (80 individuals). Markers each represent one sample with 1 (blue) representing non or low checkers, 2 (green) medium checkers, and 3 (red) severe checkers.

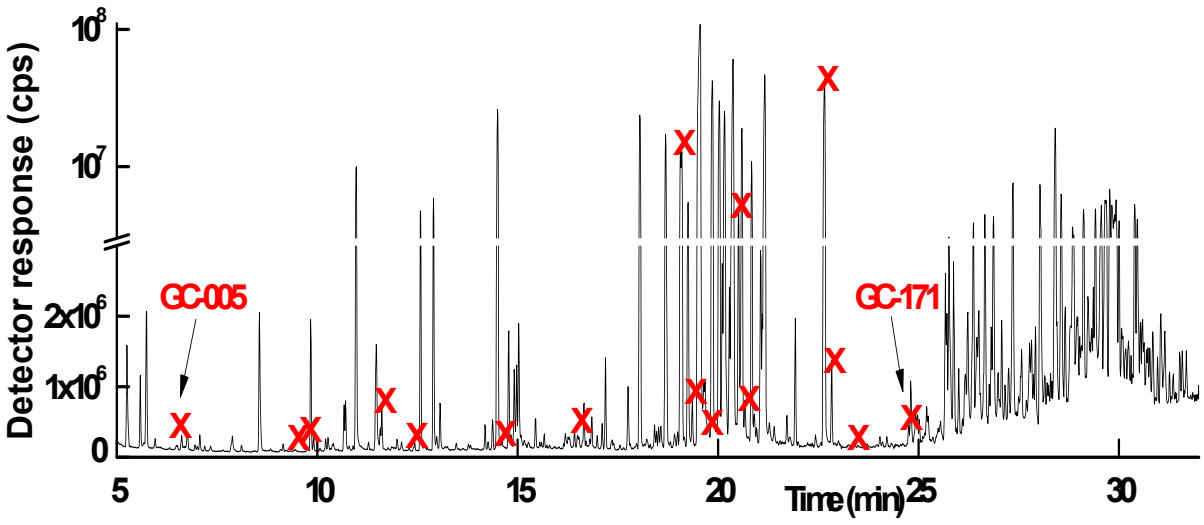


Figure 3.3 Representative GC/MS chromatogram demonstrating the complexity of the metabolite profile and the location of the 16 highly differential metabolites used in the MDA model based on high and low checkers. “cps” = counts per second.

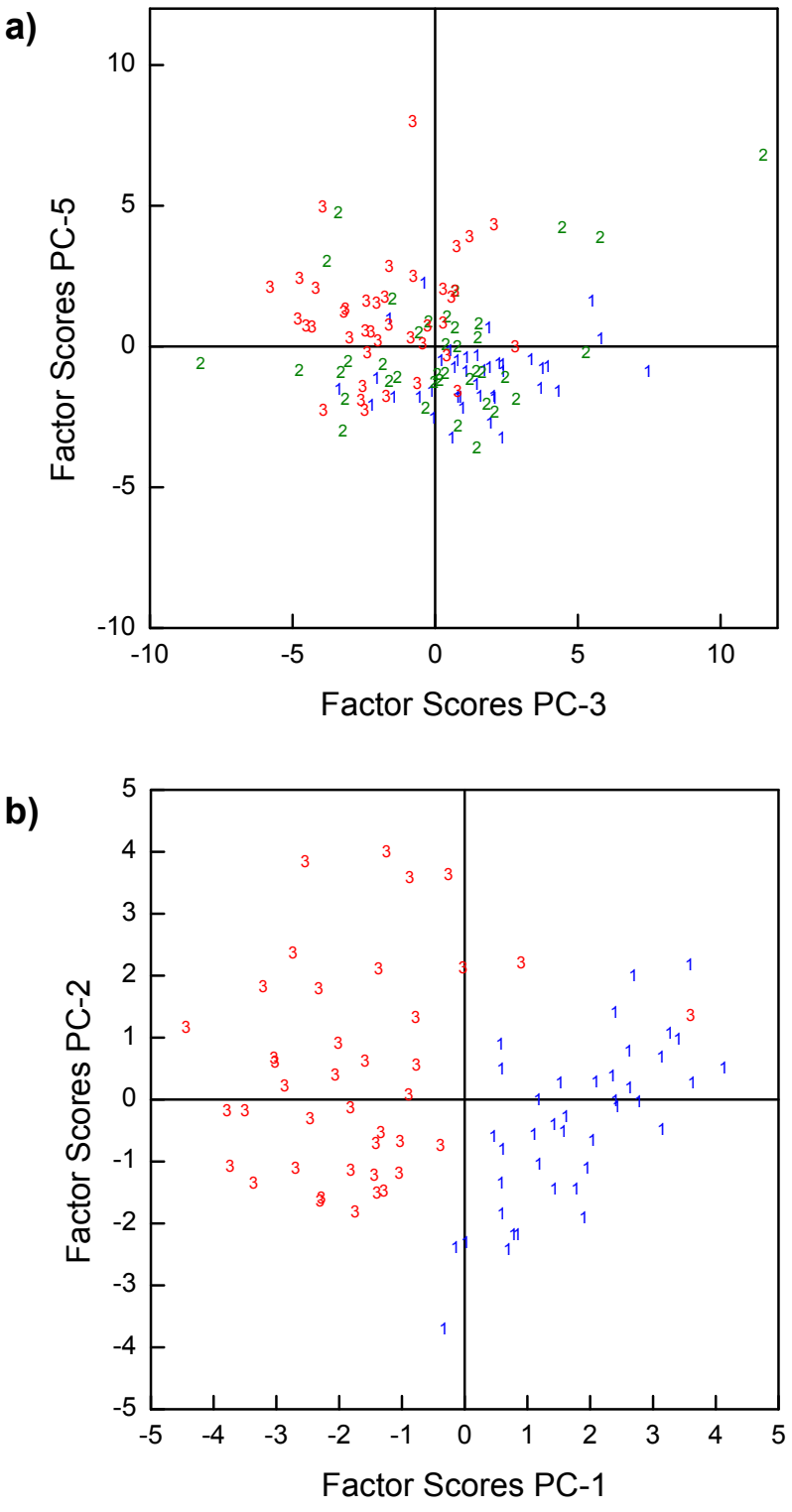


Figure 3.4 Factor score plots from PCA of reduced metabolite profiles for: a) high, medium and low checking families (106 metabolites, 119 individuals), and b) high and low checking families (16 metabolites, 80 individuals). Markers each represent one sample with 1 (blue) representing non or low checkers, 2 (green) medium checkers, and 3 (red) severe checkers.

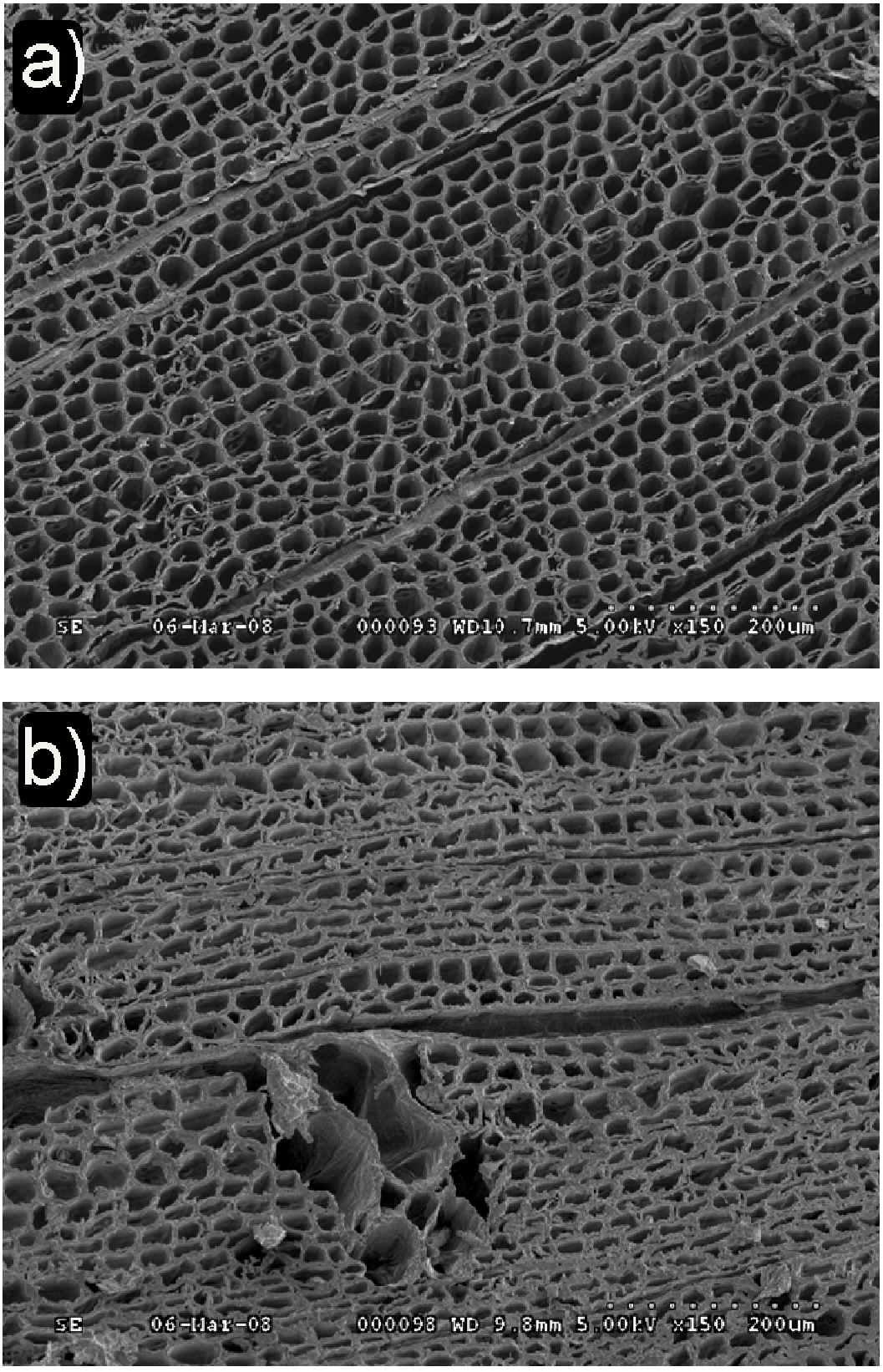


Figure 3.5 Scanning electron micrographs of radial cross-sections of juvenile radiata pine from a) a non-checking individual, and b) a high-checking individual.

Figure 3.6 Scanning electron micrograph showing the detail of an internal check originating at a ray file, in juvenile radiata pine.

Table 3.1. Summaries of cross-validated MDA models for the prediction of internal checking severity based on complete GC/MS metabolite profiles of developing xylem. "Class Set" specifies the sample classes, "Sample#" indicates number of samples, and "Metab." Indicates number of metabolites included in each model.

				Class prediction accuray (%)		
Class Set	**Dataset**	**Sample#**	**Metab.**	**Low**	**Medium**	**High**
High Medium Low	Complete	119	228	97.5	81.4	95.0
High Low	Complete	80	228	95.0	-	92.0
High Medium	Complete	79	228	-	89.2	97.5
Medium Low	Complete	79	228	87.5	86.9	-

Table 3.2. Summaries of cross-validated MDA models for the prediction of internal checking severity, based on reduced profiles including only those metabolites exhibiting significantly different abundances in the sample classes analysed, as judged by logical tests. "Class Set" specifies the sample classes included in each model. Reduction test and criteria for significance specified in "Dataset" and "α". "Metab." indicates number of metabolites included in each model.

				Class prediction accuracy (%)		
Class Set	**Dataset**	**α**	**Metab.**	**Low**	**Medium**	**High**
High Medium Low	Complete		228	77.5	81.4	95.0
119 samples	Logical F-	0.01	106	60.0	35.6	72.5
		0.001	61	55.0	44.4	60.0
High Low	Complete		228	95.0	-	92.5
80 samples	Logical t-	0.05	55	77.5	-	82.5
		0.01	30	85.0	-	85.0
		0.001	16	97.5	-	90.0

Table 3.3. Detailed list of metabolites having significant difference in abundance between high and low checkers (t-test α = 0.001). "RT avg" indicates average retention time of the metabolite in gas chromatography, "Rel abundance" indicates the abundance of the metabolite in high checkers, expressed relative to the abundance in low checkers.

Peak #	RT avg (min)	Metabolite Identity (mass fragments)	Rel abundance
GC-005	6.54	Unidentified	2.64
GC-021	9.47	Unidentified	0.69
GC-022	9.76	Unidentified	1.22
GC-036	11.60	Succinic acid 2TMS	0.67
GC-043	12.48	L-Serine 3TMS	1.30
GC-060	14.62	Unidentified	0.76
GC-082	16.51	4-hydroxybenzoic acid 2TMS	0.70
GC-117	19.10	Shikimic acid 4TMS	1.21
GC-121	19.37	Unidentified; carbohydrate	1.31
GC-124	19.70	Unidentified; sugar alcohol	0.69
GC-135	20.55	Unidentified	0.83
GC-137	20.70	Unidentified; carbohydrate	0.66
GC-158	22.66	Inositol 6TMS	1.23
GC-159	22.83	Unidentified; carbohydrate	1.32
GC-162	23.37	Unidentified; Quinic acid derivative	1.20
GC-171	24.72	Unidentified	0.82

For unidentified metabolites, fragment mass(relative abundance) of 10 most abundant fragments are:

RT(avg)	Fragment mass, rel abundance (base peak 100) |
6.54	131 100 | 73 82 | 147 70 | 149 23 | 75 16 | 132 12 | 148 10 | 133 10 | 74 9 | 219 5 |
9.47	73 100 | 147 60 | 188 20 | 204 16 | 149 12 | 89 11 | 74 9 | 131 8 | 177 7 | 104 7 |
9.76	124 100 | 214 62 | 107 21 | 157 16 | 73 15 | 125 11 | 215 11 | 98 10 | 114 9 | 158 6 |
14.62	73 100 | 147 67 | 202 66 | 230 57 | 229 53 | 215 29 | 227 21 | 149 16 | 235 12 | 155 12 |
19.37	73 100 | 257 95 | 289 90 | 204 64 | 147 37 | 217 29 | 258 23 | 189 23 | 290 23 | 379 23 |
19.70	73 100 | 217 63 | 147 47 | 159 42 | 318 32 | 129 22 | 247 22 | 163 22 | 188 17 | 191 15 |
20.55	147 100 | 73 72 | 205 38 | 189 29 | 149 26 | 148 18 | 117 18 | 273 13 | 133 9 | 89 8 |
20.70	147 100 | 73 64 | 205 53 | 189 39 | 149 34 | 285 26 | 204 21 | 117 15 | 148 15 | 273 13 |
22.83	73 100 | 191 96 | 147 75 | 343 62 | 433 43 | 204 35 | 318 28 | 149 19 | 434 18 | 192 18 |
23.37	345 100 | 255 54 | 346 29 | 73 19 | 191 17 | 256 15 | 239 14 | 147 13 | 347 12 | 217 108 |
24.72	73 100 | 147 67 | 173 61 | 217 57 | 129 31 | 305 25 | 335 22 | 149 18 | 218 15 | 263 14 |

3.5 References

Ball, R.D. & McConchie, M.S. (2001). Heritability of internal checking in Pinus radiata - evidence and premiminary estimates. *N.Z.J.For.Sci.* 31**,** 78-87.

Ball, R.D., McConchie, M.S., & Cown, D.J. (2005). Evidence for associations between SilviScan-measured wood properties and intraring checking in a study of twenty-nine 6-year-old Pinus radiata. *Can. J. Forest Res.* 35**,** 1156-1172.

Booker, R.E., Haslett, T.N., & Sole, J.A. (2000). Acoustic emission study of within-ring internal checking in radiata pine. *The 12th international symposium on nondestructive testing of wood. http://www.ultrasonic.de/article/v06n03/booker/booker.htm*

Chen, F., Duran, A.L., Blount, J.W., Sumner, L.W., & Dixon, R.A. (2003). Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* 64**,** 1013-1021.

Donaldson, L.A. (1995). Cell-wall fracture properties in relation to lignin distribution and cell dimensions among three genetic groups of radiata pine. *Wood Sci. Technol.* 29**,** 51-63.

Donaldson, L.A. (1997). Ultrastructure of transwall fracture surfaces in radiata pine wood using transmission electron microscopy and digital image processing. *Holzforschung* 51**,** 303-308.

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18**,** 1157-1161.

Fiehn, O. & Weckwerth, W. (2003). Deciphering metabolic networks. *Eur. J. Biochem.* 270**,** 579-588.

Franke, R., Humphreys, J.M., Hemm, M.R*., et al.* (2002). The Arabidopsis REF8 gene encodes the 3-hydroxylase of phenylpropanoid metabolism. *Plant J.* 30**,** 33-45.

Herrmann, K.M. & Weaver, L.M. (1999). The shikimate pathway. *Annu. Rev. Plant Phys.* 50**,** 473-503.

Hoffmann, L., Maury, S., Martz, F., Geoffroy, P., & Legrand, M. (2003). Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. *J. Biol. Chem.* 278**,** 95-103.

Humphreys, J.M. & Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5**,** 224-229.

Kopka, J., Schauer, N., Krueger, S.*, et al.* (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21**,** 1635-1638.

Kumar, S. (2004). Genetic parameter estimates for wood stiffness, strength, internal checking, and resin bleeding for radiata pine. *Can. J. Forest Res.* 34**,** 2601-2610.

Maddern-Harris, J. (1991). Structure of wood and bark. in Kininmonth, J.A. & Whitehouse, L.J. (Eds), Properties and used of New Zealand radiata pine. Ministry of Forestry, Forest Research Institute, Rotorua, pp. 2-1 - 2-16.

Putoczki, T.L., Nair, H., Butterfield, B., & Jackson, S.L. (2007). Intra-ring checking in Pinus radiata D. Don: the occurrence of cell wall fracture, cell collapse, and lignin distribution. *Trees- Struct. Funct.* 21**,** 221-229.

Robinson, A.R., Gheneim, R., Kozak, R.A., Ellis, D.D., & Mansfield, S.D. (2005). The potential of metabolite profiling as a selection tool for genotype discrimination in Populus. *J. Exp. Bot.* 56**,** 2807-2819.

Roessner, U., Luedemann, A., Brust, D.*, et al.* (2001a). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13**,** 11-29.

Roessner, U., Willmitzer, L., & Fernie, A.R. (2001b). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127**,** 749-764.

Samuels, A.L., Rensing, K.H., Douglas, C.J., Mansfield, S.D., Dharmawardhana, D.P., & Ellis, B.E. (2002). Cellular machinery of wood production: Differentiation of secondary xylem in Pinus contorta var. latifolia. *Planta (Berlin)* 216**,** 72-82.

Savidge, R. (1989). Coniferin, a biochemical indicator of commitment to tracheid differentiation in conifers. *Can. J. Bot.* 67**,** 2663-2668.

Schoch, G., Goepfert, S., Morant, M.*, et al.* (2001). CYP98A3 from Arabidopsis thaliana is a 3'-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. *J. Biol. Chem.* 276**,** 36566-36574.

Williams, D.H. (1981). Internal checking in New-Zealand-grown radiata pine after high-temperature drying. *N.Z.J.For.Sci.* 11**,** 60-64.

CHAPTER 4

The potential of metabolite profiling as a selection tool for genotype discrimination in *Populus*

A version of this chapter has been published and the original publication is available at www.oxfordjournals.org. Robinson, A.R.; Gheneim, R.; Kozak, R.A.; Ellis, D.D., and Mansfield, S.D. 2005. The potential of metabolite profiling as a selection tool for genotype discrimination in *Populus*. J. Exp. Bot. 56**,** 2807-2819.

4.1 Introduction

Improvements in plant breeding for the forest industry are reliant on the development of new tools that allow the early selection of trees based on inherent wood quality traits, in addition to more classical attributes such as growth rate and overall biomass yield (volume) (Campbell and Sederoff, 1996). The demand for this approach to breeding has arisen because in many cases the suitability of wood for specific end uses is heavily influenced by the inherent physical and chemical attributes that it exhibits. This affects the value of wood in the marketplace, as well as the efficiency and economic viability of secondary processes that use wood as a feedstock.

The aromatic biopolymer, lignin, is a principal structural component in woody tissue, and contributes significantly to vascular integrity and wood strength (Donaldson, 2001). Lignin is formed as one of the major products of the phenylpropanoid pathway, and the mechanisms of its biosynthesis have been the focus of intense research (Dixon *et al.*, 2001; Humphreys and Chapple, 2002; Li *et al.*, 2003). Particular attention has been directed towards the identification of relevant biosynthetic enzymes and corresponding genetic material, as well as understanding the regulation of gene expression (transcription, translation, and enzyme-substrate interactions), and its role in developmental and tissue-specific biosynthesis (Anterola and Lewis, 2002; Rogers and Campbell, 2004). In terms of industry, the abundance and variable nature of lignin influences wood durability, the suitability of wood for manufacturing, and has implications for the use of wood as a feedstock for the production of secondary products such as high-grade paper (Huntley *et al.*, 2003).

In the course of secondary xylem biosynthesis, resources are passed through biochemical pathways in order to generate monomeric units, which are subsequently assembled into the constituent polymers (e.g. lignin, cellulose and hemicellulose). This process involves spatially and temporally controlled enzymatic activity that causes flux through multi-reaction pathways; a component of this may be the pooling of some of the chemical intermediates produced. The nature and inherent variability of the constituents of wood manifests phenotypes, and in some way must be related to the biosynthetic material from which these polymers are constructed and their assembly process. The specificities of both flux and pooling of biosynthetic materials are presumably representative of the biosynthetic pathway to which they contribute. Given this, patterns

in the relative abundance of small molecules (metabolites) participating in cellular metabolism could be effective indicators of phenotypes related to wood quality traits. 'Metabolomics' or 'metabolic profiling', the measurement and comparison of metabolic traits, is increasingly being employed as a powerful approach to characterise living organisms (Fiehn *et al.*, 2001), and may also prove useful in the selection of trees in the context of tree improvement programmes.

With the advent of routine high-throughput bench-top chromatography-mass spectrometry, the ability to resolve and identify the metabolites in crude tissue extracts has improved dramatically. The utility of these techniques has been effectively demonstrated in the context of metabolite profiling for plant biology (Fiehn *et al.*, 2000b; Fiehn and Weckwerth, 2003; Frenzel *et al.*, 2002; Roessner *et al.*, 2001b; Tolstikov and Fiehn, 2002). Metabolic profiling, however, has yet to be developed and applied widely in plant breeding, although such use is inevitable as it is a powerful tool to characterize plant phenotypes.

Herein, we evaluate the ability of metabolite profiling to distinguish between the metabolomes of genotypically differentiated lines of the same hybrid tree, expressing different phenotypes that relate to industrially relevant wood chemistry attributes. Due to its unique position in active tree-related functional genomics programs (Brunner *et al.*, 2004), hybrid poplar was chosen as the tree species, and lignin biosynthesis and its associated impact on cell wall formation as the system to demonstrate the use of metabolic profiling to differentiate desirable phenotypes in trees. Transformation of *Populus tremula* × *alba* with a C4H::F5H genetic construct (comprised of the xylem-specific cinnamate 4-hydroxylase (C4H) promoter coupled to the ferulate 5-hydroxylase (F5H) gene (both from Arabidopsis), has been shown to significantly increase the ratio of syringyl (S) to guaiacyl (G) monomers in the lignin of this hybrid (Franke *et al.*, 2000). Increases in the S:G ratio are associated with improved chemical (kraft) pulping efficiency, and as such, have environmental and economic implications for pulp and paper manufacture (Huntley *et al.*, 2003). The results in this study clearly demonstrate the ability of metabolite profiling to differentiate between trees differing in industrially relevant wood quality traits due to this single gene modification.

4.2 Materials and methods

4.2.1 Plant materials and sampling

Hybrid poplar P717 (*Populus tremula* × *alba*) was selected as the control. In addition, two genetically modified lines that exhibit marked changes in wood chemistry and quality attributes were adopted as treatments. These represent separate transformation events involving the same construct, which consists of the xylem-specific cinnamate 4-hydroxylase (C4H) promoter coupled to the ferulate 5-hydroxylase (F5H) gene (both from *Arabidopsis*). The C4H::F5H construct has been shown to significantly increase the ratio of syringyl (S) to guaiacyl (G) monomers in poplar lignin, although the severity of the observed phenotype is transformation event specific (Huntley *et al.*, 2003). The unmodified wild-type has 65.6% mol syringyl content, whereas F5H-82 and F5H-64, have 82.5% mol and 93.4% mol syringyl content, respectively (Huntley *et al.*, 2003). It should be noted that the modified lines, referred to as F5H-82 and F5H-64 in this work and that of Huntley *et al*. (2003) correspond to those referred to as “B” and “I”, respectively, by Franke *et al*. (2000).

At their origin, the control and modified lines were regenerated concurrently and in an equivalent manner, from leaf blade-derived callus after the tissue had undergone Agrobacterium-based transformation. Lines were subsequently maintained as sterile shoot cultures. To generate plant material for this study, shoot cultures were clonally propagated on semi-solid Woody Plant Medium (WPM) (McCown and Lloyd, 1981) (6.3Appendix D.3), supplemented with 0.01 mM α-naphthalene acetic acid (NAA), under a 16 h / 8 h light/dark regime. Fluorescent light was supplied at a photon flux density of 50 µmol/m^2/s. For the generation of test trees, wild-type and F5H-64 plantlets were transferred to soil-based medium upon rooting, and then grown in randomised plots in a greenhouse under a natural light regime. Developing xylem was sampled in August 2003 mid-way through the third growing season, during daylight hours and in full sunlight. Tissue from the cambial zone was obtained from each tree by first peeling a rectangular section of bark/phloem/outer cambium from approximately 15 cm above the ground on the stem, and then scraping the developing xylem with a fresh razor blade. Care was taken to avoid sampling from nodes. The collected material was quickly isolated and transferred to a cryovial, snap-frozen in liquid nitrogen and stored at -80°C.

4.2.2 Suspension cultures

All three lines were propagated as cell suspensions in sterile liquid culture using WPM supplemented with 10 µM 2,4-dichlorophenoxyacetic acid (2,4-D). Cultures were initiated using 1-2 mm internode sections (30-50×) and 10mL of medium in sterile 50 mL Erlenmeyer flasks. Nodal tissue, which contains meristematic cells, was avoided for culture initiation. Each flask was sealed with a foam bung and foil cap, and placed on an orbital shaker at 135 rpm. The light/dark regime was as described for plantlet culture, above. Half of the spent medium was replaced every seven days until the tissue began to proliferate (2-5 weeks). Following proliferation, 10 mL fresh medium was added to the culture to give a total culture volume of 20 mL. When sub-culturing at subsequent weekly intervals, suspensions were first diluted or concentrated so that after a settling period of 30 min, tissue occupied half of the culture volume. A 5 mL aliquot of this (~2.5 mL packed cell volume) was then transferred to a new flask containing 16 mL fresh medium. Stability, based on uniform growth and morphology, was achieved for all cultures within 2-3 months. For metabolite profiling of stable lines, tissue samples were isolated from the growing medium, quickly transferred to cryovials, snap-frozen in liquid nitrogen and stored at -80°C. To obtain daily measurements for the growth rate experiment, cultures were allowed to settle in sterile graduated cylinders for 30 min, after which time cell volume data were recorded and cultures were returned to their flasks.

4.2.3 Nucleic acid preparation and semi-quantitative RT-PCR

Total RNA was extracted from suspension culture tissue using the method of Kolosova *et al*. (2004) (6.3Appendix D.4). Invitrogen SuperScript II reverse transcriptase was used to synthesise first-strand cDNA, which was then used as template in a semi-quantitative PCR with the following primers, yielding a 71 base-pair fragment. Forward primer 5′-CGTTGTCTCTCTTTTCATCTTC-3′, reverse primer 5′-CGTGGACCGGGAGGATATG-3′. PCR products were visualised on an agarose gel using ethidium bromide staining.

4.2.4 Metabolite sample preparation

Frozen tissue was ground to a fine powder using a dental amalgam mixer, employing a liquid N_2-chilled copper/plastic capsule containing three steel ball bearings and the

sample was shaken violently for 15 s. Samples were kept frozen at all times and, once ground, were returned to -80°C.

Metabolites were extracted from tissue samples and prepared for GC/MS using a scaled-down and re-optimised version of a two-phase methanol/chloroform method developed for metabolite extraction from the leaves of *Arabidopsis* (Fiehn *et al.*, 2000b). Approximately 20 mg frozen, ground developing xylem was weighed into a pre-chilled 2 mL lock-cap centrifuge tube (for suspension cultures 50 mg tissue was used). CH_3OH (600 µL) was added immediately and the sample was vortexed for 10 s to halt biological activity and minimise degradation. H_2O (40 µL), 10 µL polar internal standard (10 mg/mL ribitol in H_2O) and 10 µL lipophilic internal standard (10 mg/mL nonadecanoic acid methyl ester in $CHCl_3$) were added. Metabolites were extracted from the sample by incubation for 15 min at 70°C with constant agitation, and following a 5 min centrifugation of the sample at 13 000 rpm the supernatant was transferred to a new 2 mL tube. $CHCl_3$ (800 µL) was added to the pellet and vortexed for 10 s to re-suspend. The sample was then incubated for 5 min at 35°C with constant agitation, and the supernatant recovered following a second 5 min centrifugation at 13 000 rpm, and combined with the supernatant from the CH_3OH extraction. Following the addition of 600 µL H_2O to the combined supernatant and 10s vortexing, the mixture was centrifuged at 4000 rpm for 15 min to separate the methanol/water (upper) and methanol/chloroform (lower) phases. In theory, metabolites partition themselves between the two phases depending on which they have more affinity for – the upper phase being more polar and the lower more lipophilic. A 1 mL aliquot was taken from the upper phase with care, to avoid contamination from the interphase, and stored at -20°C overnight if not processed immediately. Metabolites contained in the lower phase were not analysed in this study.

Samples were then derivatised for GC/MS. A 900 µL aliquot of the methanol/water phase was dried using a Speedvac (Savant) (3-4 h, low temp). For the protection of carbonyl moieties by methoxylation, the pellet was resuspended in 50 µL methoxyamine hydrochloride solution (20 mg/mL in pyridine) and incubated with constant agitation for 2 h at 60°C. Acidic protons were then trimethylsilylated with 200 µL N-methyl-N-trimethylsilyltrifluoro acetamide (MSTFA) and incubated at 60°C with constant agitation for 30 min. Samples were left to stand at room temperature overnight

to ensure the reaction was complete, and then filtered through compacted tissue paper to remove particulate matter prior to analysis by GC/MS.

Metabolites were extracted from tissue samples (cambial scrapings and tissue cultures) and prepared for HPLC analysis by extracting 200 mg liquid nitrogen-frozen, ground tissue in 1.5 mL methanol: water: acetic acid (48.5: 48.5: 1.5) at 60°C for 4 h. Following incubation, the samples were centrifuged for 10 minutes at 13 000 rpm, and the supernatant recovered. Equal volumes of ethyl ether were added and the sample mixed and allowed to phase separate. The upper fraction was removed and retained. The sample was then extracted a second time with ethyl ether, collected, pooled and dried under vacuum. Samples were resuspended in 200 µL methanol and analysed using reverse phase HPLC.

4.2.5 GC/MS analysis

GC/MS analysis was conducted on a ThermoFinnigan Trace GC-PolarisQ ion trap system fit with an AS2000 auto-sampler and a split/splitless injector. The GC was equipped with a low-bleed Restek Rtx-5MS column (fused silica, 30 m, 0.25 mm ID, stationary phase diphenyl 5% dimethyl 95% polysiloxane). The GC conditions were set as follows: inlet temperature 250°C, helium carrier gas flow at constant 1 mL/min, injector split ratio 10:1, resting oven temp 70°C, and GC/MS transfer line temperature 300°C. Following injection of a 1 µL aliquot of sample, the oven was held at 70°C for 2 min and then ramped to 325°C at a rate of 8°C/min. The temperature was held at 325°C for 6 min before being cooled rapidly to 70°C in preparation for the next run.

For MS analysis in positive electron ionisation (EI) mode, the fore-line was evacuated to approximately 40 mTorr, with helium gas flow into the chamber set at 0.3 mL/min. The source temperature was held at 250°C, with an electron ionisation potential of 70 eV. The detector signal was recorded from 3.35 min after injection until 35.5 min, and ions were scanned across the range of 50-650 mu (mass units) with a total scan time of 0.58 s.

4.2.6 HPLC analysis

Phenolic metabolite composition was determined by reverse phase high performance liquid chromatography (HPLC) on a Summit chromatograph (Dionex, Sunnyvale, CA). Separation was achieved on a Symmetry C18 250 mm × 2.0 mm reverse phase column (Waters), and detected by a photodiode array detector. Samples were filtered through

compacted tissue paper prior to injection (50 µL). The column was eluted with a linear gradient of 5% 95:5 water:acetic acid (v/v) to 100%, 25% acetonitrile (v/v) in 95:5 water:acetic acid (v/v) over 70 min at a flow rate of 1.0 mL/min.

4.2.7 Data processing and statistical analysis

ThermoFinnigan 'Xcalibur' software was used for both GC/MS data collection and peak identification and measurement. The grouping of peaks that represented the same compound in multiple chromatograms was automated using the in-house, purpose-built 'PeakMatch' software. Data reduction by principal components analysis (PCA) was carried out using the Statistical Package for the Social Sciences (SPSS) v12.0 (SAS Institute, Inc., Cary, N.C.). All other intermediate data manipulation was carried out using Microsoft Excel 2000.

4.3 Results and discussion

4.3.1 Suspension cultures

Established suspension cultures generated from wild-type and C4H::F5H modified lines (F5H-82 and F5H-64) grew at similar rates, and showed characteristic lag, linear and static phases of growth over a 9 d period (Figure 4.1). As such, samples taken at day 7 for metabolite profiling were from cultures in the transition from linear growth to the static phase. Expression of the *Arabidopsis* F5H transgene in suspension cultures was confirmed by semi-quantitative RT-PCR (image not shown). There was no detectable expression of the *Arabidopsis* F5H transgene in the non-transformed wild-type control, as expected. However, even under the highly controlled conditions of suspension culture, which did not promote organ-specific differentiation, the modified genotypes continued to express the transgene and maintain phenotypes that differed from one another as well as from the wild-type control. The cultures also exhibited distinct morphologies, with wild-type cells being white in colour, F5H-82 greenish, and F5H-64 displaying a distinct brown colour (Figure 4.2). Furthermore, the wild-type cultures were visually finer cultures with smaller cell aggregates, whereas the transgenic cultures tended to be composed of larger cellular aggregates. Colour changes have been observed in the wood of trees from modified poplar lines in which the lignin content or the S:G ratio has been increased (Pilate *et al.*, 2002), and it is possible that the colour changes observed in both wood and suspension-cultured tissue reflect similar

biochemical phenomena. In the case of C4H::F5H, it is likely that the colour is due to the product(s) of a pathway fed by an abundance of an over-supplied syringyl lignin biosynthetic pathway. Despite the continued expression of the transgene in suspension cultures, ultraviolet microscopy revealed no evidence of secondary wall development (images not shown). A possible explanation for the continued activity of the secondary development-specific C4H promoter, in the absence of both secondary development and lignin polymer biosynthesis, is that phenylpropanoid biosynthesis is frequently induced during times of environmental stress; this is likely the case in these liquid cultures.

4.3.2 Metabolite data acquisition and compiling

To elucidate the metabolites present in both actively dividing cambial and suspension-cultured tissue, total ion chromatograms (TIC) of each sample, wild-type and transgenic, were obtained by GC/MS analysis of TMS-derivatives from crude tissue extracts. Analysis of the cambial zone included samples from 15 wild-type and 10 F5H-64 individual tree clones. The analysis of suspension cultures included samples from 20 distinct cultures of each of the wild-type, F5H-82 and F5H-64 lines (60 cultures in total), which were sampled during the transition from linear to static culture growth, 7 d after subculture. For all recorded peaks, total ion counts remained within the linear detection range of the instrument (approximately $1.0e^4$ - $3.0e^8$ counts/s).

In preliminary calculations, each peak in a chromatogram was expressed relative to the area of the ribitol internal standard peak. In addition, peak areas were normalised across all chromatograms (of developing xylem or suspension culture datasets) by adjusting for the exact amount of tissue (mg fresh weight) used in each sample extraction.

In order to circumvent the wobble in retention time for any given compound, a single-pass algorithm ("PeakMatch") was designed to group peaks from multiple chromatograms that have similar retention times based on a user-assigned threshold. It has been well recognised that one of the limitations of metabolomics has been the difficulty in automating the process of grouping peaks that represent the same compound in multiple chromatograms (Fiehn, 2001; 2002; Fiehn *et al.*, 2001; Fiehn and Weckwerth, 2003). However, automation is a necessity when analysing large numbers of replicates displaying hundreds of peaks typical in GC/MS total ion chromatograms

from plant metabolite extracts. To avoid a total-chromatogram-alignment-by-data-point approach such as that used in correlation optimised warping (COW) (Nielsen *et al.*, 1998), and to identify peaks and use peak area to measure compound quantity without warping, alternate software that can match peaks while accommodating the variability in retention time must be employed. In this study, PeakMatch served as a highly effective tool for rapidly compiling large datasets and accomplishing the needed comparisons of the same compound in different samples.

After being compiled in PeakMatch, but prior to statistical analysis, datasets were cleaned of all superfluous peaks not directly related to the sample. These included the internal ribitol standard, solvent impurities, and any peaks from the reagents used in the derivatisation process (linear siloxane chains and other silyl compounds). The retention times of such peaks were identified from the TIC chromatograms of pyridine solvent blanks, and sample blanks in which the extraction and derivatisation were carried out in the absence of any sample tissue. In addition, all but the most prominent peaks eluting after 30 min were excluded from the analysis, as beyond this time the signal to noise ratio declined drastically due to the heavy convolution in the high-mass tail end typical of GC/MS analyses.

To maintain uniformity across the dataset, the sensitivity of peak finding must remain fixed across all chromatograms, although a particular setting will be more or less appropriate for different chromatograms. As a consequence, minor peaks are often detected inconsistently, despite being visible in the chromatogram. To reduce the noise introduced by this erroneous non-detection of minor peaks, peaks sets were thinned in two ways. All peaks detected in <10% of samples from each plant line, and all peaks whose average normalised area for each plant line were less than a specific value (~$1.0E^{-4}$ for developing xylem and ~$5.0E^{-5}$ for suspension culture) were not considered. With completion of all adjustments, the final xylem and suspension culture datasets contained 143 and 182 peaks, respectively.

4.3.3 Principal components analysis

Principal components analysis was conducted separately on developing xylem and suspension culture peak sets. For the xylem dataset, 22 principal components were required to account for >99% of the variance between the 143 peaks across all 25 samples (total 3575 data points) (Figure 4.3). This represents roughly an 85%

reduction in variables. Similarly, for the suspension dataset, 48 principal components were required to account for >99% of the variance between 60 samples across all 182 peaks (total 10920 data points) (Figure 4.3). This represents approximately a 74% reduction in variables. The considerable reduction in variables achieved by PCA suggests the existence of strong relationships between the variables within datasets.

Plotting the factor scores of individual samples from selected principal components, as coordinates on the axes of two- or three-dimensional scatter plots, can generate a graphical representation of the relationship between samples in a PCA. The separation of clusters of samples in such a plot illustrates the existence of differences between distinct metabolic systems (Chen *et al.*, 2003; Fiehn, 2003; Fiehn *et al.*, 2000a; Morris *et al.*, 2004; Roessner *et al.*, 2001a; Roessner *et al.*, 2001b). Standard plots are limited to three dimensions, and the components plotted should be those that best represent the dataset. This implies that the components plotted are those that account for the most variance (*i.e.* the first, second and third components); however, specific latter components have also been shown to be effective in revealing differences between sample groups in some situations (Fiehn *et al.*, 2000a). In such cases, it is often more useful to plot factor scores from these discriminating components.

In this study three, two-dimensional scatter plots were generated for each dataset using component pair combinations from the first three principal components (Figures 4.4 and 4.5). Together, these three principal components account for approximately 46% and 52% of the variance in the xylem and suspension culture datasets, respectively (Table 4.1). The developing xylem plots (Figure 4.4) clearly illustrate that both principal components 2 and 3 (PC-2 and PC-3) distinguished between the wild-type and F5H-64 samples, with PC-2 being more effective. In contrast, PC-1 made no such distinction. It follows that the best visualisation of separation between the two lines is achieved when PC-2 and PC-3 are combined (Figure 4.4c). In this case, loose clustering and complete separation of the wild-type and F5H-64 samples are observed, with these phenomena derived primarily from PC-2, but accentuated by PC-3. Furthermore, clustering of wild-type samples in this plot is visibly a tighter grouping than that of F5H-64 samples. In comparison, the suspension culture plots (Figure 4.5) show that in this PCA, PC-1 distinguished between wild-type, F5H-82 and F5H-64, while PC-2 distinguished F5H-82 from the others. Here, it was

PC-3 that failed to effectively distinguish the lines. Therefore, in this case the best visualisation of separation between the three lines is achieved when PC-1 and PC-2 are combined (Figure 4.5a). This plot illustrates a tight clustering of the three lines, with visible improvement from F5H-82 to F5H-64 and then to wild-type (barring the outlier). Furthermore, all three lines separate cleanly and equally from one another, with the F5H-82 cluster separating from the others in PC-2 such that there is a very clear overall separation.

It is evident from the scatter plots in Figures 4.4 and 4.5 that the PCA detected differences between the metabolisms of the three phenotypically distinct lines, resulting from single gene insertion events. Visual evidence of this can be seen in selected two-dimensional plots (Figures 4.4c and 4.5a), where samples from each line cluster together, and separately from the samples of other lines. This observation supports the theory that differences in wood chemistry can indeed be associated with differences in observable metabolic traits; however, what PCA achieves, and what the correct interpretation of clustering and separation in PCA scatter plots should be, is not entirely simple and warrants discussion. Clustering does not necessarily indicate that those samples in a cluster contain, in this case, a similar abundance of the various metabolites detected. Likewise, neither does the separation of clusters necessarily indicate absolute differences. Rather, clustering of samples in PCA indicates a similarity in the behaviour of variables in relation to one another. Samples that clustered together in this study did so because they each contained a similar set of metabolites whose abundances were correlated in the same way. An accurate interpretation, therefore, affords the results from PCA greater relevance in the context of comparing biochemical systems. The power of this approach lies in that it is based not on isolated comparison of the abundance of individual metabolites in different systems, but instead accounts for the dynamic nature of metabolism, and provides insight into metabolic relationships.

A comparison of developing xylem (Figure 4.4c) and suspension culture (Figure 4.5a) plots reveals differences between the PCA clustering patterns of samples taken from the two sources. It is apparent that clustering and separation is more defined for suspension culture samples than it is for xylem samples. This may be due to differences in the degree of environmental variability experienced by the tissues derived

from the two sources. Actively growing trees will have experienced long-term and recurring differences in temperature, relative humidity, light, water availability, space, and insect herbivory, despite greenhouse climate control. Environmental factors such as these can cause variation in the growth, morphology and, presumably, metabolism of trees of the same genotype. In contrast, sterile tissue cultures grown under strictly controlled laboratory conditions most probably experience less long-term culture-to-culture environmental variability and, consequently, exhibit reduced morphological and biochemical variation. As such, replicate samples of the same genotype show less variability in suspension cultures than they do as greenhouse-grown trees, as illustrated by a comparison of the ”tightness” of clustering in PCA.

A trend observed across both scatter plots is that the wild-type samples tend to cluster more tightly than the modified samples. This suggests increased metabolic and phenotypic variability in the modified genotypes, compared to the non-transformed, wild-type control.

4.3.4 Elucidating individual metabolites

Having established that metabolite profiling coupled with principal components analysis could be employed to distinguish the different lines, the natural progression was to characterise the metabolic traits underlying the clustering and separation phenomena. For this, the component matrix of the PCA was screened for variables (metabolites) with high loadings in the specific principal components that produced clustering and separation in the scatter plots. The greater the loading, the more the variable is a pure measure of the component (Tabachnick and Fidell, 2001), and the more influence it has on the generation of the principal component; therefore, high-loading variables are responsible for generating clusters and separation in principal components where these phenomena occur. It has been suggested that loadings in excess of 0.71 are ‘excellent’, 0.63 ‘very good’, 0.55 ‘good’, 0.45 ‘fair’, and 0.32 ‘poor’ (Comrey and Lee, 1992).

In this study metabolites with at least ‘fair’ loadings were extracted from the component matrix for the first three principal components of developing xylem (Table 4.3) and suspension culture (Table 4.4) datasets. National Institute of Standards and Technology (NIST) MS-Search software equipped with the NIST mass spectra, as well as the Max Planck Institute Trimethylsilane (TMS) (http://www.mpimp-

Golm.mpg.de/mms-library/index-e.html) and our own (Mansfield laboratory) TMS mass spectral libraries was used to assist with the identification of these metabolites. Compounds with high-scoring matches (based on mass spectrum and retention time) were assigned identities and classified as 'amino acid', 'phenolic', 'carbohydrate' or 'other' (including sterols, phosphates, components of the citric acid cycle and adjunct pathways) molecules.

In the PCA for suspension cultures, PC-1 and PC-2 clustered and separated all three lines. In PC-1 (Table 4.4a), 65% of high-loading metabolites were carbohydrates (including monomers, dimers and their phosphorylated or acidic derivatives), which, for the most part, had loading values better than 'good' (as defined by Comrey and Lee, 1992). Additionally, there was evidence of the inorganic phosphate pool, with a few examples of amino acids, glutamate (primary donor of the α-amino group to most amino acids), a participant in the citric acid cycle (malic acid) and a by-product of shikimic acid biosynthesis (quinic acid). Some phenolic compounds were observed, but for the most part barely loaded above the cut-off. With these results, it is appropriate to suggest that in PC-2, the clustering and separation of all three lines with minimal overlap was heavily related to differences in carbohydrate metabolism. A similar analysis of high-loading metabolites in PC-2 (Table 4.4b) revealed components of the citric acid cycle (succinic acid, fumaric acid), components of the triose-phosphate pathway (glyceric and pyruvic acids), shikimic acid (precursor of many phenolic amino acids and secondary metabolites), myo-inositol phosphate (amongst other things, inositol participates in signalling pathways, hormone storage and transport, and the biosynthesis of cell walls and stress-related compounds), and a selection of early- and late-eluting carbohydrates (monomers, dimers). Although the loadings of carbohydrates are typically higher than those of other molecule types in this principal component, the appearance of a series of closely related core metabolites suggests that this aspect of metabolism had a significant influence on the clustering and separation observed in PC-2.

The principal components PC-2 and PC-3 of the developing xylem dataset effectively clustered and separated samples of the wild-type and F5H-64 lines, although PC-2 alone separated the lines with minimal overlap. Examples from all molecule categories were observed, although as with PC-1 of the suspension culture dataset, carbohydrates predominate in the list of high-loading metabolites in xylem PC-2 (Table

4.3b). The list of high-loading metabolites in PC-3 (Table 4.3c) is an even more pronounced case of carbohydrate dominance, with 83% of metabolites identified as carbohydrates. The GC breakdown peaks of sucrose (which all represent the same compound) feature strongly, and it is understandable that they load highly together. Interestingly, inositol and glutamate load highly in this principal component, much as they did in suspension culture PC-1 and PC-2; however, no representatives from the core citric acid and triose-phosphate pathways were observed. It again seems appropriate to attribute the small amount of separation observed in xylem PC-3 to differences in carbohydrate metabolism. Figure 4.6 reveals the variety in abundance, as well as the broad range of retention time of identifiable, high-loading compounds present in the differentiating components of the developing xylem dataset, PC-2 and PC-3.

Xylem PC-1 and suspension culture PC-3 are the first in the respective datasets that do not distinguish between lines (Figures 4.4 and 4.5). These components do, however, carry considerable interest with regard to high-loading metabolites. In both of these components, high-loading amino acid-related metabolites were prominent (Tables 4.2, 4.3a and 4.4c). In suspension culture PC-3, 39% of high-loading metabolites were amino acids, all of which were identified. Likewise, 42% of high-loaders in xylem PC-1 were amino acid-related (of these, 69% were identified). This clustering of amino acids into the first principal components that fail to distinguish between lines suggests that amino acid biosynthesis and metabolism maintained a high level of stability, despite genetic transformation with C4H::F5H.

Notably, the aromatic amino acids tyrosine and tryptophan were observed as very high loaders in xylem PC-1. In some plant species, tyrosine can be used as a precursor in hydroxycinnamic acid biosynthesis (Alemanno *et al.*, 2003; Deluca *et al.*, 1988; Whetten and Sederoff, 1995) and as a precursor to pigments and defence compounds such as alkaloids (Facchini, 2001), flavonoids (Koch *et al.*, 1995) and anthocyanins (Dube *et al.*, 1992; Sakuta *et al.*, 1991). Tryptophan is used in some plant species as a precursor to bioactive alkaloids (Facchini, 2001) and defence phytoalexins (Pedras *et al.*, 2003; Zhao and Last, 1996), as well as the phytohormone auxin (indole 3-acetic acid) (Bartel, 1997). As major products of the shikimic acid pathway, and molecules that are synthesised in close proximity to the usual precursor of monolignol

biosynthesis, phenylalanine, the observed behaviour of tyrosine and tryptophan is intriguing. The tight association of tyrosine with a principal component that did not distinguish between the wild-type and the transgenic lines suggests that, in this case, any flux of resources through this branch of metabolism and into monolignol biosynthesis was not affected by the transformation event. This would agree with the wood chemistry of the modified phenotype, in which the total lignin content (as determined by Klason analysis) was comparable to the control (Huntley *et al.*, 2003). Notably, none of the aromatic amino acids were observed as high-loaders in suspension culture PC-3, and their absence may be related to an absence of predation in suspension culture. Interestingly, phenylalanine was not present in either the developing xylem or suspension culture datasets.

A series of amino acids not directly related to phenolic secondary metabolism were identified as high-loaders in the non-differentiating principal components. Three of the four major nitrogen assimilation amino acids (Suarez *et al.*, 2002) were observed: glutamate in suspension culture PC-3, and aspartate and asparagine in xylem PC-1. Also, the aspartate-derived amino acid, threonine, was identified in both xylem PC-1 and suspension culture PC-3. This amino acid is the precursor to isoleucine, a branched chain amino acid (Giovanelli *et al.*, 1988). Valine and leucine, two other branched chain amino acids, were identified in xylem PC-1 and suspension culture PC-3, respectively. Branched chain amino acids are precursors to secondary metabolism, and are involved in the biosynthesis of cyanogenic glycosides, glucosinolates and acyl sugars (Conn, 1988).

4.3.5 Metabolite Channelling

Surprisingly, very few phenolic compounds are found in the lists of high-loading metabolites from the PCA. The GC/MS analysis detected rather few phenolic metabolites, and only one compound, sinapyl alcohol, was identified as an intermediate of the phenylpropanoid pathway for lignin monomer biosynthesis (reviewed by Dixon *et al.*, 2001). Clearly, however, there is an abundance of small phenolic compounds synthesised in living plant tissue as either intermediates in, or endpoints of, metabolic pathways. Hypothetically, the concept of ‘metabolite channelling’ may provide an explanation for these observations.

A metabolic channel exists when metabolic intermediates are covalently bound to, and passed between, sequential active sites of a multi-functional enzyme or a multi-enzyme complex (Hrazdina and Jensen, 1992; Srere, 1987; Srere, 2000). It is postulated that this arrangement typically occurs where chemical intermediates have no other cellular function except in that particular biosynthetic pathway. When a metabolic channel exists, free pools of chemical intermediates are extremely small, if they exist at all. In this way, cellular solvent capacity is spared for the regulation and efficiency of the metabolic sequence, and also for containment of molecules having cytotoxic properties. Metabolic channels are thought to exist in many branches of plant secondary biosynthesis, and there is good evidence to suggest their participation in the complex regulation of resource partitioning from the end of the shikimate pathway into and through numerous divergent pathways, notably those of flavonoid and lignin biosynthesis (Achnine *et al.*, 2004; Anterola *et al.*, 1999; Rasmussen and Dixon, 1999; Winkel-Shirley, 1999). The results presented here, and those of Achnine *et al* (2004) clearly indicate that analogous channelling mechanisms exist in the biosynthesis of phenolic compounds, and specifically in this case in poplar tree species.

Traditional reverse phase HPLC was employed in order to validate the isolation and identification of monolignol precursors (Figure 4.7). HPLC clearly demonstrated and confirmed (GC/MS) that the only lignin pre-cursor that defferentially accumulated (pooled) in the C4H::F5H 64 transgenic line when compared to wild-type plants was sinapyl alcohol. Given the location of F5H in the lignin biosynthetic pathway, 5-hydroxyconiferaldehyde should accumulate in the differentiating cambial zone, should channelling not be occurring. This compound was not identified by either HPLC or GC/MS (verified by retention time and mass spectra from synthesised compound).

Limited detection of phenolic molecules may be related to the choice of analytical tools. Even with sample derivatisation, the molecular weight cut-off of gas chromatography ranges between 800-1000 Da. Once derivatised, many phenolic and other compounds produced in plant tissues are larger than this and may not be resolved by GC/MS. Notably, this includes the glycosylated phenylpropanoid molecules thought to be storage and/or transportation forms of the monomers for lignin polymer assembly (Samuels *et al.*, 2002). Given the functional role of F5H in lignin biosynthesis, located in the latter part of the phenylpropanoid pathway prior to the biosynthesis of

glycosylated phenylpropanoids, there is a possibility that the direct metabolic impact of F5H up-regulation could be visible in the relative abundances of glycosylated monolignols. In order to resolve such large metabolites from crude tissue extracts, further analysis using complementary analytical techniques that have higher mass cut-offs is currently underway. To this end, extension of the research presented will focus on applying LC/MS-based profiling tools to the study of metabolism in this same poplar model system.

Metabolite profiling of crude extracts derived from the cellular 'bulk' phase is confounded by another important limitation. It is not possible to detect, measure or identify 'product' metabolites that establish physical associations with cellular structural components in the course of metabolism, and maintain them during extraction procedures. This point may be of great significance in the study of cell wall and wood biosynthesis by metabolite profiling. Pyrolysis-MS, with its ability to liberate entire tissue samples and analyse the resulting compounds may provide a solution to this, and is another analytical technique that warrants investigation.

4.4 Concluding remarks

Metabolite profiling analysis of compounds exhibiting cellular pooling in the developing xylem and suspension-cultured tissue of hybrid poplar revealed multiple series of metabolites that correlated with one another in terms of relative abundance. The metabolic interaction networks represented by these series were either affected by a lignin-related C4H::F5H genetic modification, or remained consistent despite it. Thus, it was possible to distinguish between wild-type and transgenic lines exhibiting a range of phenotypic severity, on the basis of observable metabolic traits. Of particular interest were the apparent consistency of the amino acid-related pools between wild-type and transgenic lines, and the heavy role of carbohydrates in distinguishing between lines, despite a modification that related specifically to lignin biosynthesis.

Using GC/MS and traditional reverse phase HPLC it was not possible to detect any intermediate metabolites (*i.e.* 5- hydroxyconiferaldehyde) that related directly to the C4H::F5H genetic modification. This suggests that bulk phase pools of such metabolites do not exist *in vivo*, and metabolite channelling occurs during cell wall lignification in developing xylem and suspension cultures.

This research has established an approach to the investigation of global metabolism in a model tree system, poplar. By analysing the relationships that exist between abundances of the small molecules that pool in plant tissue, it has been possible to define certain aspects of the metabolic space that links gene expression and phenotypic character.

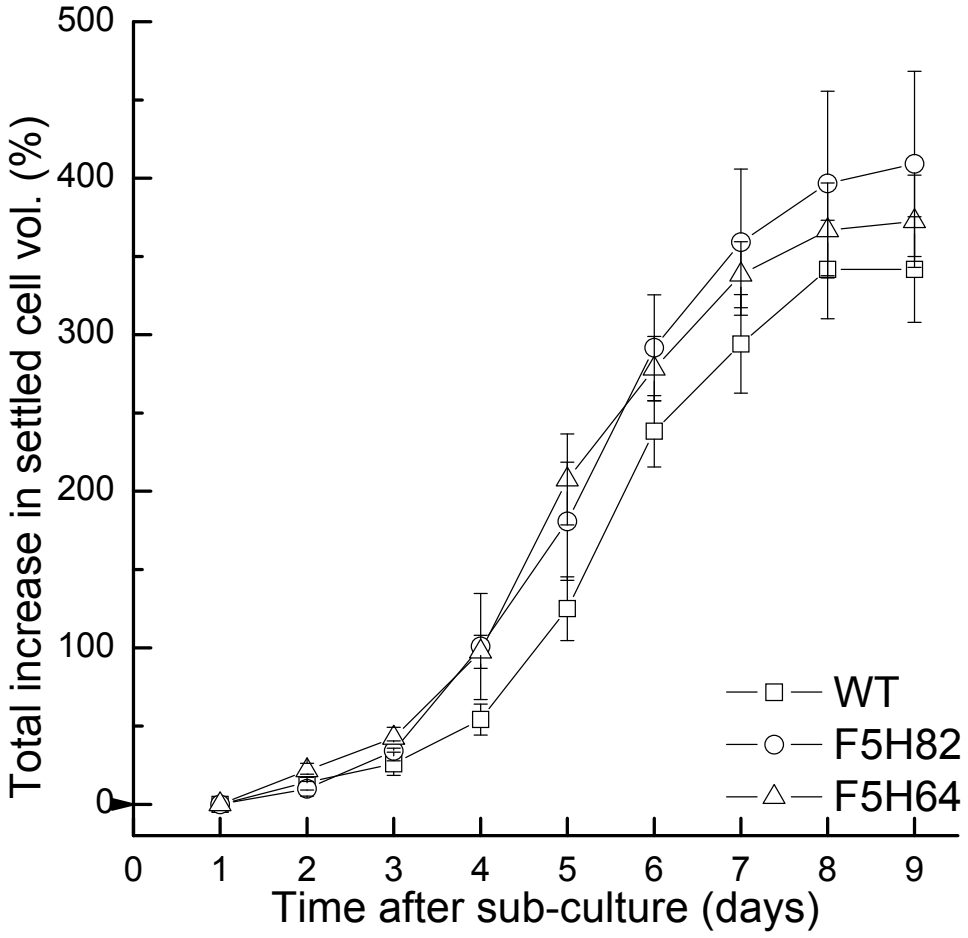


Figure 4.1 Growth characteristics of wild-type and two C4H::F5H transformed *P. tremula* × *alba* suspension cultures based on settled cell volume. Plots represent the mean of twelve replicates, and error bars represent a 95% confidence interval of the mean. Arrow indicates sampling time for metabolite profiling.

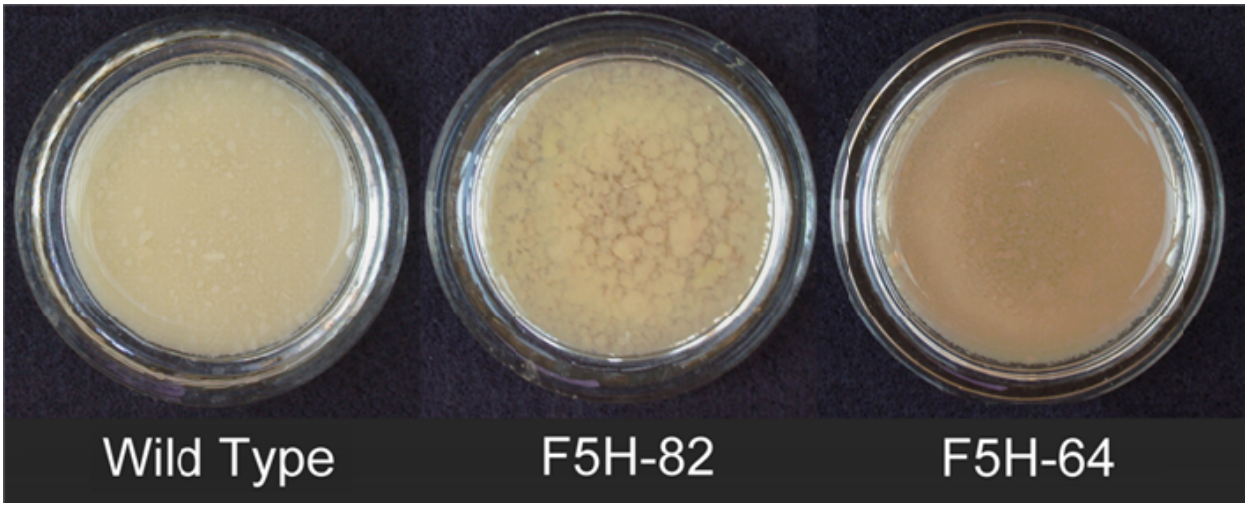


Figure 4.2 Suspension-cultured tissue of wild-type and two C4H::F5H transformed *P. tremula* × *alba* lines. Picture was taken fourteen days after subculture. Watch glass diameter is approximately 6.5 cm.

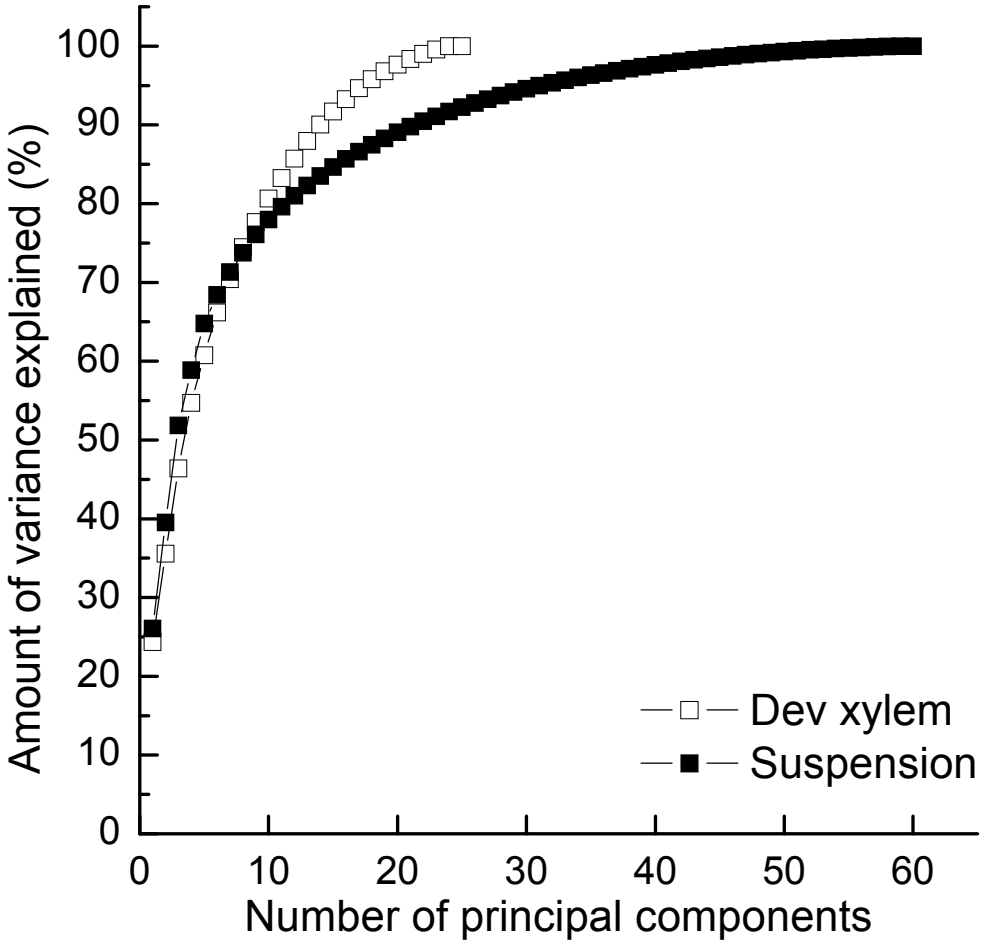


Figure 4.3 Cumulative percentage of dataset variation explained by principal components, for both developing xylem and suspension cultures.

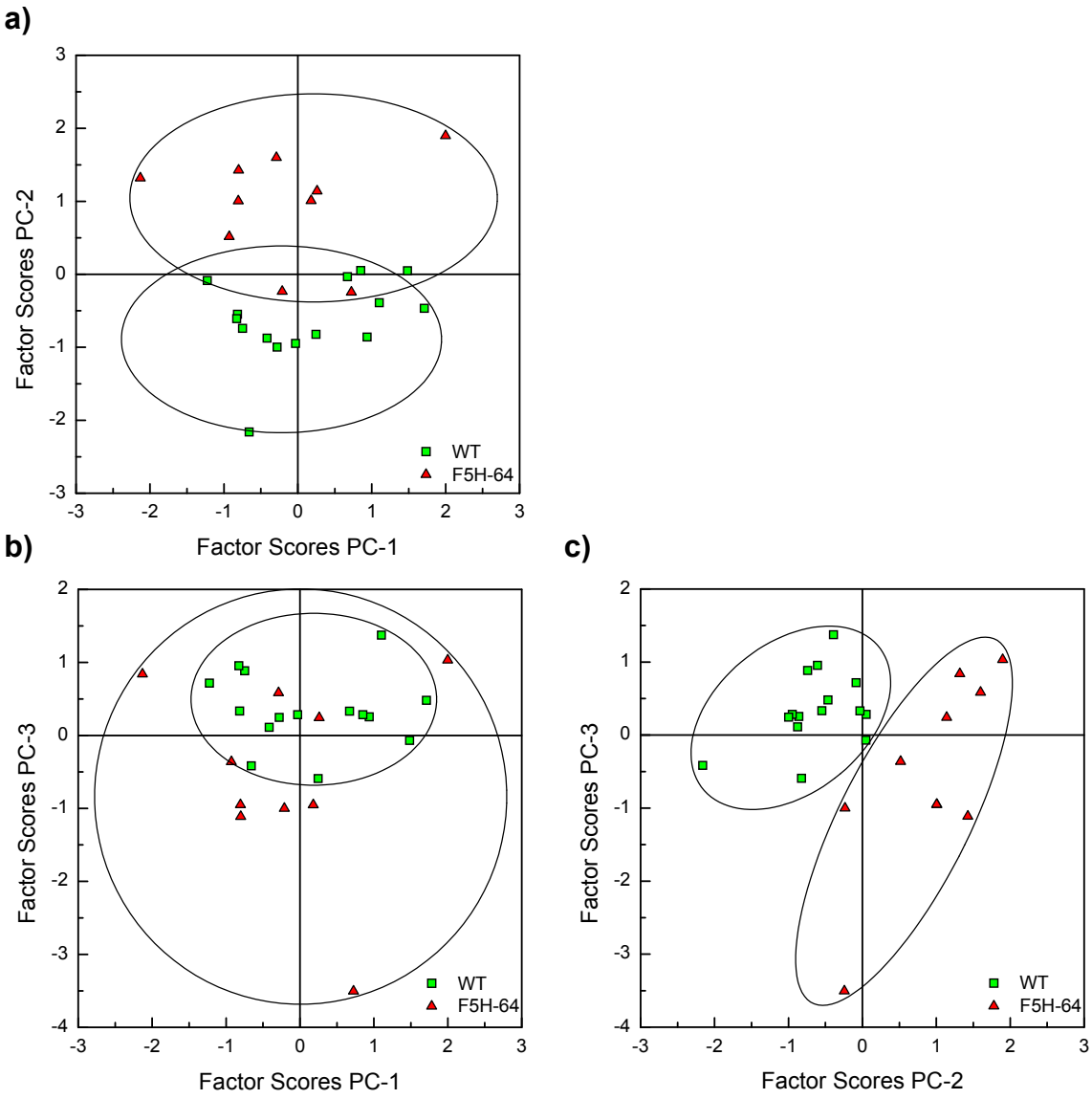


Figure 4.4 Scatter plots of PCA factor scores for wild-type and F5H-64 modified samples from the developing xylem dataset. Axes of two-dimensional plots are derived from a) PC-1 and PC-2, b) PC-1 and PC-3, and c) PC-2 and PC-3. Plotted points represent individual samples, while arbitrary ellipses have been included to assist interpretation and simply border all samples of individual lines. This PCA analysis represents the differentiation of 25 individual trees (15× wild-type and 10× F5H-64).

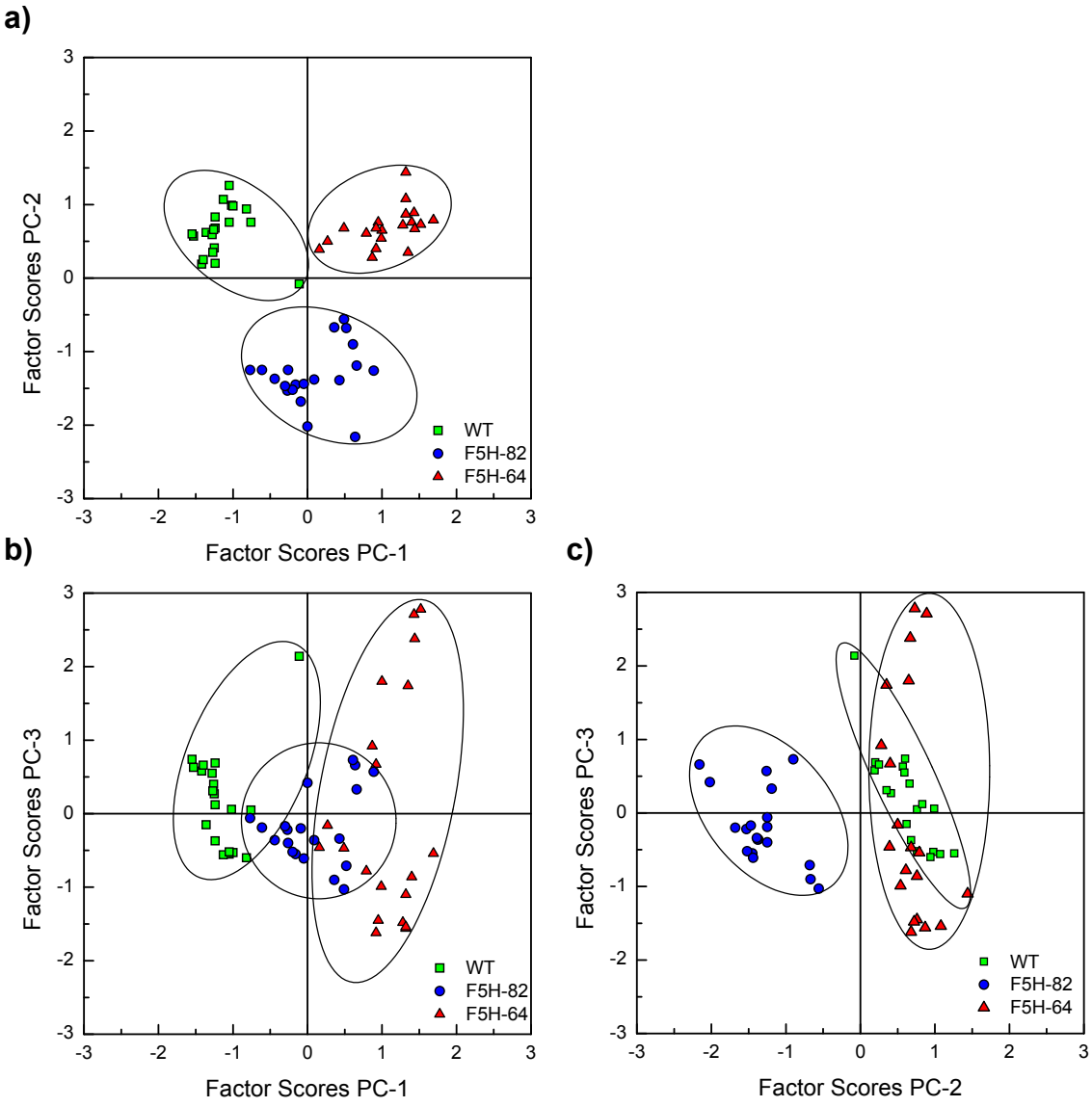


Figure 4.5 Scatter plots of PCA factor scores for wild-type and C4H::F5H transformed *P. tremula* × *alba* samples from the suspension culture dataset. Axes of two-dimensional plots are derived from a) PC-1 and PC-2, b) PC-1 and PC-3, and c) PC-2 and PC-3. Plotted points represent individual samples, while arbitrary ellipses have been included to assist interpretation and simply border all samples of individual lines. This PCA analysis represents the differentiation of 60 individual suspension cultures (20 individual samples per line).

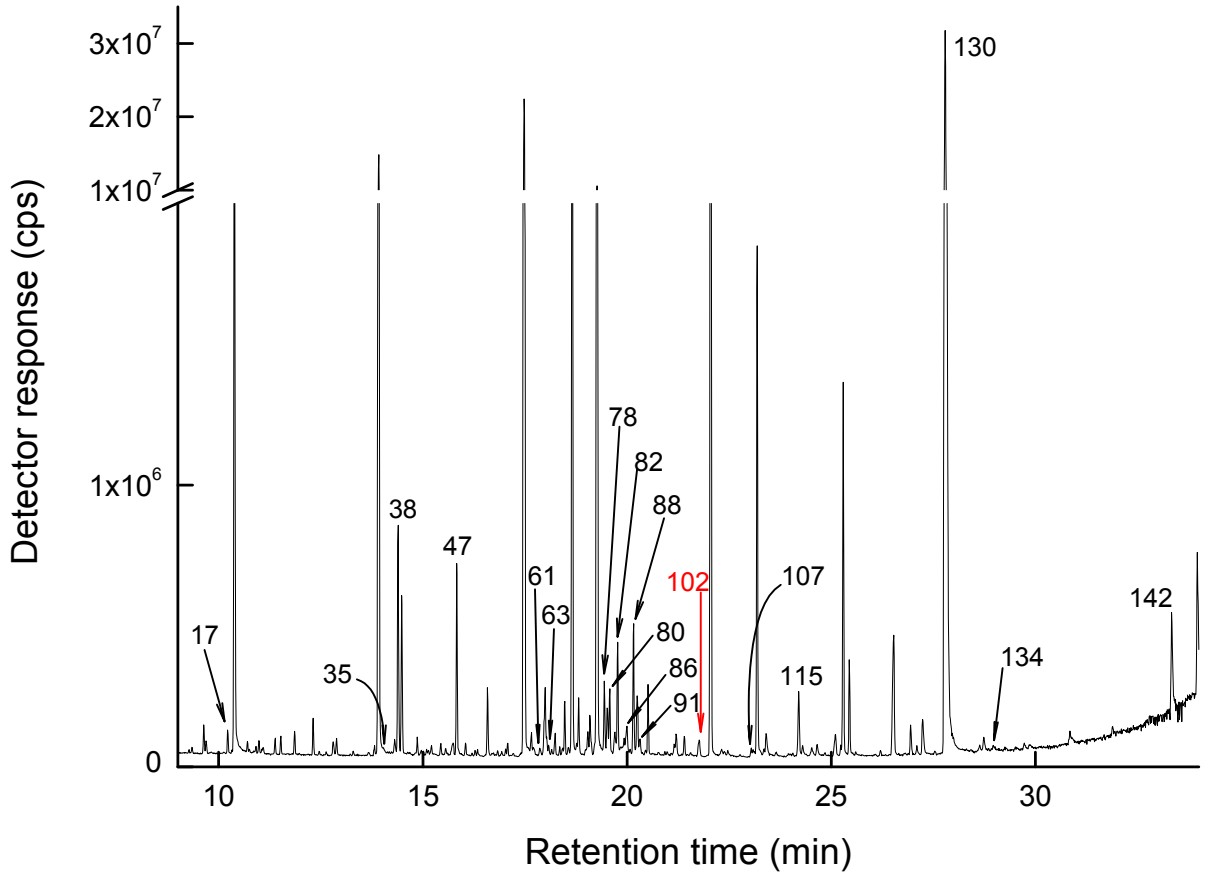


Figure 4.6 Example of a total ion chromatogram (TIC) from a developing xylem sample. Chromatogram has been annotated to indicate identified compounds that loaded highly in PC-2 and PC-3 of the PCA. These components played a significant role in distinguishing between the metabolism of wild-type and F5H-64 suspension culture lines. Refer label numbers to Tables 4.3b and 4.3c for compound identity. The detector response (y-axis) is given in counts/s (cps).

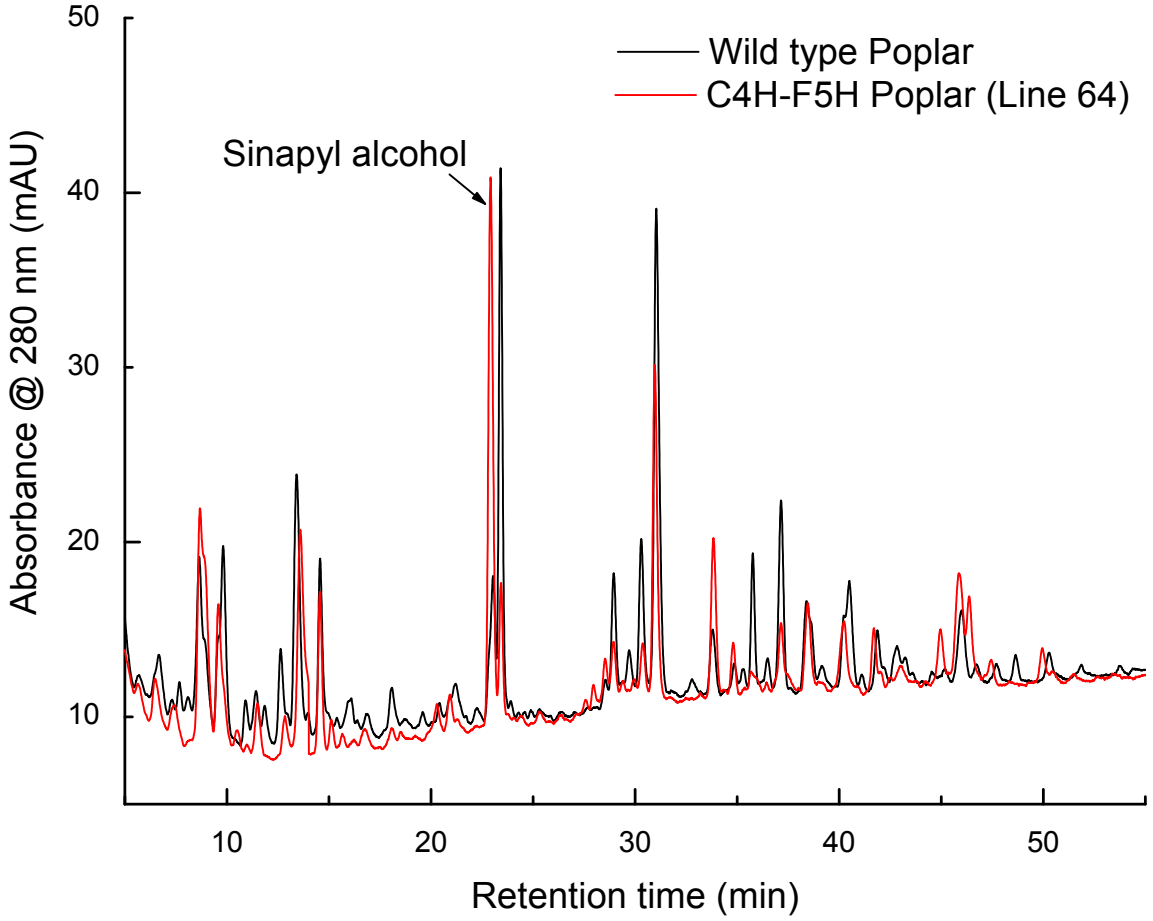


Figure 4.7 Reverse phase HPLC chromatograph of developing xylem sample of wild type and C4H::F5H transgenic plants following acid methanol extraction and detection at 280 nm.

Table 4.1. Percentage of total variance accounted for by combinations of the first three principal components of developing xylem and suspension culture datasets. Combinations revealing the greatest distinction between samples of different lines are in bold type.

Component(s)	Dev xylem	Suspension
1	24.34%	26.07%
2	11.22%	13.46%
3	10.83%	12.33%
1,2	35.56%	**39.53%**
2,3	**22.05%**	25.79%
1,2,3	46.39%	51.86%

Table 4.2. Molecule classification of the metabolites loading highly in PCA component matrices for the first three principal components. Numbers represent the number of molecules from the stated class that load high in specific principal components.

Molecule type	Dev xylem			Suspension		
	PC-1	PC-2	PC-3	PC-1	PC-2	PC-3
Other	8	6	3	12	11	8
Amino Acid	16	3	1	3	1	7
Benzene	1	2	0	4	1	0
Carbohydrate	13	6	19	35	7	3
Total	**38**	**17**	**23**	**54**	**20**	**18**

Table 4.3. Metabolites in the developing xylem dataset that load highly in the PCA component matrix. a) PC-1, b) PC-2, and c) PC-3. Only metabolites loading >0.45 in the component matrix are shown. Metabolites are sorted first by molecule class, and then by sequence of elution in gas chromatography (all peaks extracted from chromatography for PCA were assigned a number based on elution sequence). The loading of each peak is shown, and, where possible, metabolites are identified. Those that could not be identified are labelled as 'unknown', with details in parentheses (molecule type, a number based on the elution sequence, and a letter 'x' indicating developing xylem).

a) Xylem PC-1

Class	Peak#	Loading	Identity
other	11	0.60	acetimidic acid 2TMS
	17	0.52	2-amino ethanol 3TMS
	18	0.75	phosphoric acid 3TMS
	30	0.54	unknown (other#2c); mz: 73 999 | 154 447 | 174 425 | 86 289 | 59 249 | 227 148 | 100 129 | 156 105 | 74 103 | 82 98 |
	38	0.65	4-aminobutyric acid 3TMS
	69	0.70	ornithine 4TMS
	70	0.65	citric acid 4TMS
	141	0.55	unknown (other#7c); mz: 73 999 | 217 803 | 194 772 | 169 524 | 388 499 | 147 333 | 105 313 | 191 287 | 243 279 | 361 240 |
amino	15	0.61	valine 2TMS glycine 3TMS
	21	0.62	serine 3TMS threonine 3TMS
	27	0.54	unknown asparagine 2TMS
	28	0.88	aspartic acid 3TMS
	32	0.74	unknown (amino acid #3c); mz: 73 999 | 218 423 | 261 375 | 162 347 | 147 302 | 100 255 | 113 251 | 141 228 | 215 177 | 74 106 |
	35	0.49	unknown (amino acid #4c); mz: 73 999 | 216 627 | 147 558 | 142 407 | 215 379 | 188 286 | 214 192 | 149 179 | 241 163 | 217 161 |
	37	0.82	asparagine 3TMS
	42	0.89	tyrosine 3TMS
	44	0.91	valine 2TMS
	45	0.65	serine 3TMS
	46	0.75	unknown (amino acid #1c); mz: 174 999 | 73 968 | 86 461 | 156 406 | 59 399 | 79 233 | 100 214 | 175 163 | 74 150 | 147 139 |
	48	0.72	aspartic acid 3TMS
	51	0.80	unknown (amino acid #3c); mz: 73 999 | 218 423 | 261 375 | 162 347 | 147 302 | 100 255 | 113 251 | 141 228 | 215 177 | 74 106 |
	83	0.67	unknown (amino acid #4c); mz: 73 999 | 216 627 | 147 558 | 142 407 | 215 379 | 188 286 | 214 192 | 149 179 | 241 163 | 217 161 |
	87	0.76	asparagine 3TMS
	112	0.74	tyrosine 3TMS
benzene	133	0.79	p-nitrophenyl-glucoside
carb	101	0.74	glucaric acid (or galactaric acid)
	108	0.60	unknown (carb#9c); mz: 73 999 | 204 987 | 205 253 | 129 227 | 189 198 | 131 196 | 217 195 | 191 175 | 75 167 | 169 155 |
	120	0.65	unknown (carb#12c); mz: 204 999 | 73 747 | 81 295 | 147 203 | 205 186 | 217 175 | 189 169 | 171 121 | 191 98 | 206 95 |
	121	0.68	melibiose 8TMS
	122	0.64	unknown (carb#13c); mz: 73 999 | 169 526 | 204 456 | 147 294 | 331 269 | 79 225 | 361 222 | 129 197 | 217 184 | 243 144 |
	123	0.59	myo-Inositol phosphate 7TMS
	127	0.61	sucrose TMS sucrose TMS
	128	0.52	unknown (carb#14c); mz: 73 999 | 147 305 | 219 294 | 274 220 | 75 203 | 129 194 | 143 172 | 285 168 | 535 165 | 358 159 |
	132	0.52	unknown (carb#15c); mz: 73 999 | 361 832 | 169 677 | 147 417 | 243 370 | 217 327 | 271 315 | 129 265 | 362 249 | 193 184 |
	135	0.63	unknown (carb#16c); mz: 73 999 | 169 549 | 355 543 | 147 437 | 217 427 | 271 332 | 243 267 | 129 253 | 283 241 | 356 226 |
	138	0.66	unknown (carb#17c); mz: 73 999 | 361 806 | 169 691 | 147 643 | 443 564 | 217 421 | 129 379 | 243 377 | 271 361 | 362 307 |
	140	0.73	unknown (carb#18c); mz: 204 999 | 73 909 | 361 330 | 217 307 | 271 305 | 243 287 | 129 278 | 147 258 | 205 216 | 191 197 |
	143	0.45	raffinose TMS

b) Xylem PC-2

Class	Peak#	Loading	Identity
other	7	0.47	unknown (other#1c); mz: 73 999 | 191 535 | 130 503 | 75 384 | 143 374 | 77 345 | 175 318 | 79 308 | 147 306 | 69 286 |
	17	0.52	2-amino ethanol 3TMS
	38	0.48	4-aminobutyric acid 3TMS
	85	0.58	unknown (other#5c); mz: 73 999 | 147 841 | 172 310 | 133 260 | 303 232 | 117 208 | 100 161 | 149 156 | 243 142 | 205 129 |
	94	0.49	unknown (other#6c); mz: 389 999 | 347 925 | 147 613 | 73 606 | 299 540 | 463 474 | 189 422 | 259 400 | 348 360 | 390 341 |
	114	0.54	unknown (other#7c); mz: 73 999 | 147 813 | 284 459 | 189 375 | 149 339 | 285 253 | 217 201 | 194 146 | 129 140 | 268 136 |
amino	35	0.49	asparagine 2TMS
	46	0.46	unknown (amino acid #4c); mz: 73 999 | 216 627 | 147 558 | 142 407 | 215 379 | 188 286 | 214 192 | 149 179 | 241 163 | 217 161 |
	59	0.47	unknown (amino acid #6c); mz: 73 999 | 302 289 | 89 191 | 392 138 | 147 114 | 227 106 | 303 89 | 74 87 | 217 68 | 59 65 |
benzene	12	0.52	unknown (benz#1c); mz: 73 999 | 147 668 | 100 395 | 267 351 | 355 150 | 74 141 | 86 137 | 248 136 | 59 134 | 133 132 |
	102	0.64	sinapyl alcohol
carb	63	0.89	2-deoxy d-glucose 4TMS
	64	0.67	unknown (carb#2c); mz: 73 999 | 299 469 | 217 422 | 147 352 | 149 235 | 292 225 | 52 166 | 102 158 | 143 155 | 74 145 |
	67	0.85	unknown (carb#4c); mz: 147 999 | 73 892 | 189 663 | 217 346 | 261 335 | 117 312 | 149 299 | 148 173 | 129 149 | 333 136 |
	91	0.63	galactitol 6TMS (dulcitol, sorbitol are pseudonyms)
	95	0.70	unknown (carb#7c); mz: 217 999 | 73 446 | 147 383 | 218 240 | 201 224 | 52 185 | 117 132 | 219 132 | 189 112 | 291 110 |
	134	0.46	cellobiose TMS

c) Xylem PC-3

Class	Peak	Loading	Identity
other	50	0.51	unknown (other#3c); mz: 69 999 | 245 850 | 147 703 | 73 699 | 83 343 | 55 299 | 189 296 | 217 210 | 97 197 | 149 196 |
	74	0.62	unknown (other#4c); mz: 73 999 | 147 719 | 379 599 | 157 522 | 247 461 | 131 418 | 205 350 | 219 346 | 380 256 | 129 250 |
	142	0.46	sitosterol TMS
amino	47	0.58	glutamic acid 3TMS
carb	60	0.60	unknown (carb#1c); mz: 217 999 | 147 569 | 73 331 | 129 194 | 149 176 | 218 160 | 189 136 | 148 114 | 157 101 | 205 92 |
	61	0.63	galacturonic acid TMS variant
	65	0.53	unknown (carb#3c); mz: 73 999 | 147 760 | 333 680 | 217 631 | 436 516 | 143 416 | 305 379 | 331 339 | 244 311 | 257 295 |
	72	0.58	unknown (carb#5c); mz: 73 999 | 217 305 | 147 294 | 128 115 | 129 104 | 89 103 | 291 83 | 133 77 | 214 75 | 218 73 |
	75	0.68	unknown (carb#6c); mz: 73 999 | 217 318 | 147 280 | 128 148 | 133 121 | 291 114 | 129 111 | 74 82 | 89 76 | 214 73 |
	78	0.46	sucrose TMS
	80	0.51	sucrose TMS
	82	0.50	sucrose TMS
	86	0.59	sucrose TMS
	88	0.60	galactitol 6TMS (sorbitol)
	104	0.61	unknown (carb#8c); mz: 245 999 | 257 955 | 73 750 | 347 335 | 147 318 | 359 274 | 258 266 | 217 263 | 305 223 | 348 216 |
	107	0.47	inositol 6TMS
	108	0.49	unknown (carb#9c); mz: 73 999 | 204 987 | 205 253 | 129 227 | 189 198 | 131 196 | 217 195 | 191 175 | 75 167 | 169 155 |
	113	0.62	unknown (carb#10c); mz: 227 999 | 299 951 | 73 894 | 315 498 | 211 485 | 243 450 | 147 358 | 342 317 | 300 239 | 343 206 |
	115	0.51	sucrose TMS
	119	0.58	unknown (carb#11c); mz: 73 999 | 217 641 | 191 596 | 147 526 | 331 395 | 259 227 | 129 185 | 97 176 | 169 175 | 332 141 |
	122	0.54	unknown (carb#13c); mz: 73 999 | 169 526 | 204 456 | 147 294 | 331 269 | 79 225 | 361 222 | 129 197 | 217 184 | 243 144 |
	130	0.76	sucrose TMS
	140	0.48	unknown (carb#18c); mz: 204 999 | 73 909 | 361 330 | 217 307 | 271 305 | 243 287 | 129 278 | 147 258 | 205 216 | 191 197 |

Table 4.4. Metabolites in the suspension culture datasets that load highly in the PCA component matrix. a) PC-1, b) PC-2 and c) PC-3. Only metabolites loading >0.45 in the component matrix are shown. Metabolites are sorted first by molecule class, and then by sequence of elution in gas chromatography (all peaks extracted from chromatography for PCA were assigned a number based on elution sequence). The loading of each peak is shown, and, where possible, metabolites are identified. Those that could not be identified are labelled as 'unknown', with details in parentheses (molecule type, a number based on the elution sequence, and a letter 's' indicating suspension culture).

a) Suspension

Class	Peak#	Loading	Identity
other	25	0.75	unknown (other#2s); mz: 147 999 | 149 194 | 73 165 | 148 141 | 131 50 | 227 43 | 75 24 | 150 24 | 59 24 | 115 17 |
	27	0.79	propanedioic acid 2TMS
	36	0.70	phosphoric acid 3TMS
	50	0.52	unknown (other#5s); mz: 73 999 | 147 402 | 117 315 | 191 266 | 149 126 | 75 122 | 133 88 | 74 84 | 217 73 | 148 64 |
	56	0.88	2-methylmalic acid 3TMS
	57	0.71	malic acid 3TMS
	71	0.61	3-hydroxy-3-methyl-pentanedioic acid 3TMS
	83	0.48	unknown (other#7s); mz: 272 999 | 82 986 | 182 548 | 73 536 | 55 360 | 273 239 | 154 236 | 147 194 | 346 148 | 256 140 |
	90	0.81	unknown (other#8s); mz: 73 999 | 302 386 | 392 256 | 89 145 | 147 109 | 303 99 | 74 91 | 393 86 | 59 78 | 217 76 |
	107	0.67	quinic acid TMS
	123	0.47	unknown (other#9s); mz: 73 999 | 217 536 | 157 320 | 79 232 | 319 204 | 218 172 | 147 152 | 95 141 | 91 141 | 332 127 |
	135	0.48	unknown (other#10s); mz: 73 999 | 217 338 | 147 259 | 129 253 | 319 236 | 331 162 | 218 145 | 157 128 | 159 103 | 169 96 |
amino	19	0.58	alanine 2TMS
	62	0.57	glutamic acid 2TMS
	66	0.46	unknown (amino acid #1s); mz: 73 999 | 258 894 | 147 398 | 348 230 | 259 214 | 274 195 | 170 127 | 59 118 | 89 102 | 75 93 |
benzene	43	0.47	1-methyl-2-phenyl-ethylamine 2TMS
	64	0.56	unknown (benz#1s); mz: 263 999 | 73 889 | 147 518 | 264 219 | 278 207 | 348 189 | 172 170 | 158 139 | 148 139 | 149 132 |
	173	0.48	epicatechin
	175	0.47	unknown (benz#2s); mz: 73 999 | 368 807 | 355 587 | 559 580 | 560 268 | 621 258 | 369 253 | 265 219 | 356 217 | 648 213 |
carb	74	0.80	xylonic acid lactone 3TMS
	75	0.52	ribonic acid lactone TMS
	80	0.52	fucose TMS
	81	0.69	ribose meox 4TMS
	86	0.83	xylitol 5TMS
	87	0.51	n-acetyl glucosamine MEOX 4TMS
	92	0.56	glucose-1-phosphate oxim TMS
	96	0.52	unknown (carb#1s); mz: 73 999 | 257 797 | 289 632 | 217 510 | 258 201 | 379 176 | 290 169 | 147 160 | 199 154 | 103 131 |
	97	0.90	unknown (carb#2s); mz: 73 999 | 147 594 | 319 305 | 148 207 | 117 159 | 149 150 | 217 139 | 133 136 | 163 125 | 131 115 |
	98	0.83	unknown (carb#3s); mz: 73 999 | 392 412 | 217 298 | 147 189 | 89 174 | 393 137 | 59 92 | 129 86 | 172 77 | 361 77 |
	116	0.73	sorbitol TMS
	120	0.72	glucuronic acid 5TMS
	124	0.63	gluconic acid 6TMS
	126	0.75	gluconic acid lactone 4TMS
	127	0.70	inositol 6TMS
	129	0.74	unknown (carb#4s); mz: 73 999 | 147 559 | 204 374 | 189 311 | 129 190 | 149 184 | 203 183 | 205 144 | 306 137 | 74 136 |
	130	0.73	unknown (carb#5s); mz: 73 999 | 147 569 | 129 518 | 319 378 | 217 285 | 157 222 | 103 130 | 148 128 | 79 124 | 83 120 |
	132	0.64	sucrose TMS
	133	0.70	sucrose TMS
	137	0.53	unknown (carb#6s); mz: 73 999 | 204 597 | 361 338 | 147 271 | 217 145 | 75 126 | 205 124 | 169 124 | 145 121 | 129 117 |
	140	0.62	unknown (carb#7s); mz: 73 999 | 204 616 | 361 287 | 147 220 | 191 123 | 205 118 | 217 115 | 169 112 | 362 101 | 189 100 |
	144	0.58	unknown (carb#8s); mz: 73 999 | 147 251 | 133 176 | 290 98 | 217 95 | 319 93 | 129 77 | 214 65 | 75 63 | 149 61 |

146	0.65	fructose phosphate MEOX 6TMS
147	0.71	glucose-6-phosphate TMS
148	0.78	glucose-6-phosphate MEOX TMS
149	0.46	unknown (carb#9s); mz: 73 999 | 204 490 | 147 330 | 191 262 | 233 232 | 361 225 | 169 154 | 205 144 | 217 133 | 143 126 |
150	0.58	unknown (carb#10s); mz: 73 999 | 147 569 | 129 518 | 319 378 | 217 285 | 157 222 | 103 130 | 148 128 | 79 124 | 83 120 |
155	0.92	unknown (carb#11s); mz: 73 999 | 361 590 | 243 452 | 129 389 | 204 373 | 217 299 | 147 267 | 319 237 | 362 216 | 157 159 |
156	0.73	unknown (carb#12s); mz: 361 999 | 73 907 | 169 398 | 243 370 | 147 277 | 129 265 | 362 252 | 217 209 | 254 193 | 271 192 |
157	0.58	unknown (carb#13s); mz: 73 999 | 437 492 | 243 466 | 361 450 | 333 319 | 147 313 | 129 313 | 207 298 | 362 255 | 218 208 |
160	0.59	sucrose TMS
161	0.59	mannopyranose phosphate 6TMS
168	0.77	turanose 7TMS
170	0.79	unknown (carb#17s); mz: 361 999 | 73 825 | 169 373 | 362 334 | 147 290 | 204 218 | 74 184 | 191 170 | 207 162 | 363 151 |
172	0.57	melibiose MEOX TMS

b) Suspension

Class	Peak#	Loading	Identity
other	14	0.49	pyruvic acid MEOX TMS
	39	0.52	succinic acid 2TMS
	41	0.66	glyceric acid 3TMS
	42	0.47	fumaric acid 2TMS
	125	0.77	palmic acid TMS (contamination)
	131	0.45	stearyl alcohol TMS (contamination)
	141	0.64	steric acid TMS (contamination)
	143	0.72	unknown (other#11s); mz: 73 999 | 284 769 | 147 324 | 272 215 | 217 210 | 285 184 | 194 154 | 374 145 | 74 93 | 149 91 |
	176	0.70	unknown (other#13s); mz: 73 999 | 412 818 | 361 715 | 169 434 | 217 320 | 413 271 | 271 226 | 362 224 | 243 224 | 450 222 |
	177	0.57	unknown (other#14s)73 999 | 361 904 | 169 610 | 271 368 | 217 339 | 253 319 | 191 289 | 147 269 | 487 254 | 362 254 |
	181	0.46	sitosterol TMS
amino	79	0.65	indolepropionate TMS
benzene	99	0.51	shikimic acid 4TMS
carb	77	0.62	ribose MEOX 4TMS
	78	0.60	ribose MEOX 4TMS
	111	0.55	sucrose TMS sucrose TMS
	144	0.54	unknown (carb#8s); mz: 73 999 | 147 251 | 133 176 | 290 98 | 217 95 | 319 93 | 129 77 | 214 65 | 75 63 | 149 61 |
	151	0.53	Myo-Inositol phosphate 7TMS
	161	0.58	mannopyranose phosphate 6TMS
	167	0.68	unknown (carb#16s); mz: 361 999 | 73 696 | 362 274 | 147 259 | 169 242 | 204 154 | 243 122 | 271 116 | 480 115 | 363 103 |

c) Suspension

Class	Peak	Loading	Identity
other	17	0.87	unknown (other#1s); mz: 73 999 | 147 763 | 149 180 | 148 124 | 191 121 | 75 94 | 74 87 | 117 75 | 133 60 | 128 49 |
	28	0.86	unknown (other#3s); mz: 73 999 | 147 763 | 149 180 | 148 124 | 191 121 | 75 94 | 74 87 | 117 75 | 133 60 | 128 49 |
	42	0.59	fumaric acid 2TMS
	45	0.73	unknown (other#4s); mz: 73 999 | 116 301 | 147 301 | 75 278 | 306 207 | 143 180 | 149 158 | 117 104 | 245 90 | 79 79 |
	52	0.57	unknown (other#6s); mz: 73 999 | 147 523 | 110 399 | 228 281 | 75 216 | 217 215 | 77 205 | 134 136 | 149 121 | 148 89 |
	63	0.77	4-aminobutyric acid 3TMS
	83	0.80	unknown (other#7s); mz: 272 999 | 82 986 | 182 548 | 73 536 | 55 360 | 273 239 | 154 236 | 147 194 | 346 148 | 256 140 |
	165	0.49	unknown (other#2s); mz: 399 999 | 203 358 | 400 252 | 95 137 | 81 124 | 327 114 | 73 103 | 97 103 | 83 93 | 267 91 |
amino	19	0.64	alanine 2TMS
	30	0.82	valine 2TMS
	35	0.64	leucine 3TMS
	38	0.80	glycine 3TMS
	44	0.92	serine 3TMS
	47	0.46	
	62	0.68	glutamic acid 2TMS
carb	163	0.76	unknown (carb#14s); mz: 361 999 | 73 777 | 362 357 | 169 301 | 204 258 | 75 188 | 271 177 | 129 167 | 147 149 | 363 127 |
	166	0.61	unknown (carb#15s); mz: 361 999 | 73 723 | 169 492 | 217 405 | 271 297 | 204 284 | 243 249 | 93 244 | 319 237 | 300 226 |
	172	0.71	melibiose MEOX TMS (or cellobiose)

4.5 References

Achnine, L., Blancaflor, E.B., Rasmussen, S., & Dixon, R.A. (2004). Colocalization of L-phenylalanine ammonia-lyase and cinnamate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. *Plant Cell* 16**,** 3098-3109.

Alemanno, L., Ramos, T., Gargadenec, A., Andary, C., & Ferriere, N. (2003). Localization and identification of phenolic compounds in *Theobroma cacao* L. somatic embryogenesis. *Ann. Bot. (London)* 92**,** 613-623.

Anterola, A.M., van Rensburg, H., van Heerden, P.S., Davin, L.B., & Lewis, N.G. (1999). Multi-site modulation of flux during monolignol formation in loblolly pine (Pinus taeda). *Biochem. Biophys. Res. Commun.* 261**,** 652-657.

Anterola, A.M. & Lewis, N.G. (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* 61**,** 221-294.

Bartel, B. (1997). Auxin biosynthesis. *Annu. Rev. Plant Phys.* 48**,** 49-64.

Brunner, A.M., Busov, V.B., & Strauss, S.H. (2004). Poplar genome sequence: Functional genomics in an ecologically dominant plant species. *Trends Plant Sci.* 9**,** 49-56.

Campbell, M.M. & Sederoff, R.R. (1996). Variation in lignin content and composition - Mechanism of control and implications for the genetic improvement of plants. *Plant Physiol.* 110**,** 3-13.

Chen, F., Duran, A.L., Blount, J.W., Sumner, L.W., & Dixon, R.A. (2003). Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. *Phytochemistry* 64**,** 1013-1021.

Comrey, A.L. & Lee, H.B. (1992). A first course in factor analysis. Lawrence Erlbaum Associates, Hillsdale.

Conn, E.E. (1988). Biosynthetic relationship among cyanogenic glycosides, glucosinolates, and nitro-compounds. *ACS Symp. Ser.* 380**,** 143-154.

Deluca, V., Fernandez, J.A., Campbell, D., & Kurz, W.G.W. (1988). Developmental regulation of enzymes of indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiol.* 86**,** 447-450.

Dixon, R.A., Chen, F., Guo, D., & Parvathi, K. (2001). The biosynthesis of monolignols: A "metabolic grid", or independent pathways to guaiacyl and syringyl units? *Phytochemistry* 57**,** 1069-1084.

Donaldson, L.A. (2001). Lignification and lignin topochemistry: An ultrastructural view. *Phytochemistry* 57**,** 859-873.

Dube, A., Bharti, S., & Laloraya, M.M. (1992). Inhibition of anthocyanin synthesis by cobaltous ions in the 1st internode of *Sorghum bicolor* L Moench. *J. Exp. Bot.* 43**,** 1379-1382.

Facchini, P.J. (2001). Alkaloid biosynthesis in plants: biochemistry, cell biology, molecular regulation, and metabolic engineering applications. *Annu. Rev. Plant Phys.* 52**,** 29-66.

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., & Willmitzer, L. (2000a). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18**,** 1157-1161.

Fiehn, O., Kopka, J., Trethewey, R.N., & Willmitzer, L. (2000b). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72**,** 3573-3580.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genom.* 2**,** 155-168.

Fiehn, O., Kloska, S., & Altmann, T. (2001). Integrated studies on plant biology using multiparallel techniques. *Curr. Opin. Biotechnol.* 12**,** 82-86.

Fiehn, O. (2002). Metabolomics: The link between genotypes and phenotypes. *Plant Mol. Biol.* 48**,** 155-171.

Fiehn, O. (2003). Metabolic networks of Cucurbita maxima phloem. *Phytochemistry* 62**,** 875-886.

Fiehn, O. & Weckwerth, W. (2003). Deciphering metabolic networks. *Eur. J. Biochem.* 270**,** 579-588.

Franke, R., McMichael, C.M., Meyer, K., Shirley, A.M., Cusumano, J.C., & Chapple, C. (2000). Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *Plant J.* 22**,** 223-234.

Frenzel, T., Miller, A., & Engel, K.-H. (2002). Metabolite profiling: a fractionation method for analysis of major and minor compounds in rice grains. *Cereal Chem.* 79**,** 215-221.

Giovanelli, J., Mudd, S.H., & Datko, A.H. (1988). In vivo regulation of threonine and isoleucine biosynthesis in *Lemna paucicostata* Hegelm-6746. *Plant Physiol.* 86**,** 369-377.

Hrazdina, G. & Jensen, R.A. (1992). Spatial organisation of enzymes in plant metabolic pathways. *Annu. Rev. Plant Phys.* 43**,** 241-267.

Humphreys, J.M. & Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5**,** 224-229.

Huntley, S.K., Ellis, D., Gilbert, M., Chapple, C., & Mansfield, S.D. (2003). Significant increases in pulping efficiency in C4H-F5H-transformed poplars: Improved chemical savings and reduced environmental toxins. *J. Agric. Food Chem.* 51**,** 6178-6183.

Koch, B.M., Sibbesen, O., Halkier, B.A., Svendsen, I., & Moller, B.L. (1995). The primary sequence of cytochrome P450tyr, the multifunctional N-hydroxylase catalyzing the conversion of L-tyrosine to P-hydroxyphenylacetaldehyde oxime in the biosynthesis of the cyanogenic glucoside dhurrin in *Sorghum bicolor* (L) Moench. *Arch. Biochem. Biophys.* 323**,** 177-186.

Kolosova, N., Miller, B., Ralph, S.*, et al.* (2004). Isolation of high-quality RNA from gymnosperm and angiosperm trees. *BioTechniques* 36**,** 821-824.

Li, L., Zhou, Y., Cheng, X.*, et al.* (2003). Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proc. Natl. Acad. Sci. U. S. A.* 100**,** 4939-4944.

McCown, B.H. & Lloyd, G. (1981). Woody plant medium (WPM) - a mineral nutrient formulation for microculture of woody plant-species. *HortScience* 16**,** 453.

Morris, C.R., Scott, J.T., Chang, H.-M., Sederoff, R.R., O'Malley, D., & Kadla, J.F. (2004). Metabolic profiling: A new tool in the study of wood formation. *J. Agric. Food Chem.* 52**,** 1427-1434.

Nielsen, N.P.V., Carstensen, J.M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A.* 805**,** 17-35.

Pedras, M.S.C., Jha, M., & Ahiahonu, P.W.K. (2003). The synthesis and biosynthesis of phytoalexins produced by cruciferous plants. *Curr. Org. Chem.* 7**,** 1635-1647.

Pilate, G., Guiney, E., Holt, K.*, et al.* (2002). Field and pulping performances of transgenic trees with altered lignification. *Nat. Biotechnol.* [print] 20**,** 607-612.

Rasmussen, S. & Dixon, R.A. (1999). Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. *Plant Cell* 11**,** 1537-1551.

Roessner, U., Luedemann, A., Brust, D.*, et al.* (2001a). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13**,** 11-29.

Roessner, U., Willmitzer, L., & Fernie, A.R. (2001b). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127**,** 749-764.

Rogers, L.A. & Campbell, M.M. (2004). The genetic control of lignin deposition during plant growth and development. *New Phytol.* 164**,** 17-30.

Sakuta, M., Hirano, H., & Komamine, A. (1991). Stimulation by 2,4-dichlorophenoxyacetic acid of betacyanin accumulation in suspension-cultures of Phytolacca americana. *Physiol. Plantarum.* 83**,** 154-158.

Samuels, A.L., Rensing, K.H., Douglas, C.J., Mansfield, S.D., Dharmawardhana, D.P., & Ellis, B.E. (2002). Cellular machinery of wood production: Differentiation of secondary xylem in Pinus contorta var. latifolia. *Planta (Berlin)* 216**,** 72-82.

Srere, P.A. (1987). Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.* 56**,** 89-124.

Srere, P.A. (2000). Macromolecular interactions: tracing the roots. *Trends Biochem. Sci.* 25**,** 150-153.

Suarez, M.F., Avila, C., Gallardo, F.*, et al.* (2002). Molecular and enzymatic analysis of ammonium assimilation in woody plants. *J. Exp. Bot.* 53**,** 891-904.

Tabachnick, B.G. & Fidell, L.S. (2001). Using multivariate statistics. Allyn & Bacon, Boston.

Tolstikov, V.V. & Fiehn, O. (2002). Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* 301**,** 298-307.

Whetten, R. & Sederoff, R. (1995). Lignin biosynthesis. *Plant Cell* 7**,** 1001-1013.

Winkel-Shirley, B. (1999). Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiol. Plantarum.* 107**,** 142-149.

Zhao, J.M. & Last, R.L. (1996). Coordinate regulation of the tryptophan biosynthetic pathway and indolic phytoalexin accumulation in Arabidopsis. *Plant Cell* 8**,** 2235-2244.

CHAPTER 5

Assessing the between-background stability of metabolic effects arising from lignin-related transgenic modifications, in two *Populus* hybrids

A version of this chapter will be submitted for publication. Robinson, A.R., R. Dauwe and Mansfield S.D. Assessing the between-background stability of metabolic effects arising from lignin-related transgenic modifications, in two *Populus* hybrids.

5.1 Introduction

The field of plant metabolomics is currently undergoing a rapid expansion, yet fundamental aspects of the global interconnection between genetic, metabolic and phenotypic traits remain poorly defined. Of key interest in the metabolomics-based study of plant phenotype, and development of screening tools for trait selection, is the stability of broad metabolism/trait relationships within and across various genetic backgrounds. The degree of consistency in such patterns at the whole tissue or organism level, across closely related, as well as disparate plant species, will be a determining factor in the applicability of metabolite pattern data beyond the plant systems in which they are initially defined.

Reports of the non-targeted metabolomic analysis of phenotypic traits are now more frequent in the literature, including landmark analyses that have played important roles in establishing the technology and conceptual framework of the field (Andersson-Gunneras *et al.*, 2006; Le Gall *et al.*, 2003; Meyer *et al.*, 2007; Morris *et al.*, 2004; Robinson *et al.*, 2005; Robinson *et al.*, 2007; Roessner *et al.*, 2001a; Roessner *et al.*, 2001b; Rohde *et al.*, 2004). For progress to occur in this area, it is essential that the scope of such analyses be expanded. To date, metabolome/trait relationships have most frequently been characterised in individual species, or in individual families or clonal lines, thus making the specific plant system a fixed element in the analyses. It would be desirable to increase the dimensionality of the analysis by making the plant system component a variable in its own right. For example, collective metabolomic analysis of a single, specific phenotypic trait gradient across a series of genetic backgrounds (*i.e.*, cultivars, hybrids, species, ecotypes, *etc*.) could assist in identifying and defining broadly applicable relationships.

Wood is a widely used and complex material fundamental to woody plant physiology; therefore, in tree species, many of the most pertinent traits are physico-chemical wood properties. The composition of secondary cell walls in xylem tissue plays a central role in the character of woody tissue, and in its ultimate utility. It is for this reason that resources have been, and continue to be, applied to traditional breeding and transgenic modification efforts, with the intention of effecting desirable wood trait outcomes. The amorphous polymer, lignin, is a primary, integral component of plant cell walls (Donaldson, 2001), and research into its structure/function and biosynthesis is

ongoing, as is the modification of its properties. The monolignol branch of the phenylpropanoid pathway is responsible for generating the monomeric constituents of lignin, and involves the sequential hydroxylation and methylation of phenylalanine-derived cinnamic acid, as well as conversion of acid functional groups into alcohol via an aldehyde intermediate. These reactions take place under the control of a series of well-characterised enzymes (Dixon *et al.*, 2001; Hahlbrock and Scheel, 1989; Humphreys and Chapple, 2002, and references therein). This aspect of plant secondary metabolism has been a popular target for transgene-induced disruption and modification of lignin and cell wall properties in poplar, including the up-regulation of ferulate 5-hydroxylase (F5H) (Franke *et al.*, 2000; Huntley *et al.*, 2003; Robinson *et al.*, 2005), and down-regulation of cinnamyl 3′-hydroxylase (C3′H) (Coleman *et al.*, 2008a; Coleman *et al.*, 2008b).

In the present study, the consistency of the metabolic and phenotypic effects of transgenic constructs targeting lignin biosynthesis, when expressed in similar, yet distinct genetic backgrounds, was assessed. The composition analysis of wood and non-targeted metabolomic analysis of developing xylem from P717 (*Populus tremula* × *alba*) and P39 (*Populus grandidentata* × *alba*) poplar hybrids, separately transformed with each of the C4H::F5H and C3′H-RNAi constructs, allowed a comparison between modified physical and metabolic phenotypes generated by the expression of these constructs in different genetic backgrounds.

5.2 Materials and methods

5.2.1 Plant material

The hybrid poplar genetic backgrounds employed were P717 (*Populus tremula* × *alba*) and P39 (*Populus grandidentata* × *alba*). Additionally, each hybrid was separately modified with the C4H::F5H and C3′H-RNAi genetic constructs via *Agrobacterium*-mediated transformation. The preceding modification and phenotypic analysis of P39 with the C3′H-RNAi construct was conducted by Coleman *et al*. (2008a), while that of P717 with C4H::F5H was conducted by Franke *et al*. (2000). The complementary transformations of P717 with C3′H-RNAi and P39 with C4H::F5H were carried out according to the previously reported protocols (6.3Appendix D.5). Thus, the wild-type backgrounds, as well as several lines of each of P717 C4H::F5H, P39 C4H::F5H, P717

C3′H-RNAi, and P39 C3′H-RNAi, were available for comparative analysis in this study. It should be noted that the P717 C4H::F5H modified lines, referred to as “21”, “26”, “37”, “41”, “64”, “65”, “82” and “85” in this and other work (Robinson *et al.*, 2005) correspond to those referred to as “a” - “h”, respectively, by Franke *et al*. (2000).

Plantlets were first grown from apical explants for 4 weeks in sterile tissue culture, on WPM medium (McCown and Lloyd, 1981) (6.3Appendix D.3) supplemented with 0.01 μM α-naphthalene acetic acid (NAA). These were then transferred to soil-based medium in 1 gallon pots, and grown on flood tables in a greenhouse under natural summer light and ambient temperature conditions. Plants were arranged in random order to minimise positional effects. Watering was initially once a day, but after 8 weeks was increased to twice daily to accommodate the increased biomass load.

After 16 weeks of growth in the greenhouse, the height and stem diameter 5 cm from the root collar was measured for all trees. Then, in a destructive harvest, the bark/phloem was removed from the stem to allow samples of developing xylem to be collected from a region approximately two thirds down the stem, as determined by plastochron index (leaf #1 was taken as the first leaf down from the apex with a mid-vein length greater than five centimetres). These were immediately snap-frozen in liquid N_2 and stored at -80°C.

5.2.2 Metabolomic analysis

5.2.2.1 Metabolite extraction

Frozen developing xylem samples were ground to a fine powder in capsules containing several steel ball bearings, by vigorous agitation for 15 s in a dental amalgam mixer. For each sample, approximately 0.5 mL frozen, ground, developing xylem tissue was placed in a pre-weighed 2 ml microcentrifuge tube, and extracted in 1300 μL solvent mix (3% distilled, deionised water in methanol, with the internal standards ribitol (GC/MS) and ortho-anisic acid (LC/MS) added to 0.25 mg/mL and 0.164 mg/mL, respectively) for 15 min at 70°C, with orbital shaking at 1400 rpm. Following centrifugation for 10 min at 14 000 rpm, 800 μL (for LC/MS) and 200 μL (for GC/MS) aliquots of debris-free supernatant were transferred to fresh 2 mL tubes. The pellets and remaining liquid were dried overnight at 50°C, and the tube/pellet re-weighed, allowing the determination of the dry weight of tissue included in the extraction, via comparison with the previously recorded tube weight (approximately 50 mg).

For GC/MS, 130 µL chloroform and 270 µL distilled, deionised water were combined with the 200 µL aliquot of sample extract. This mixture was vortexed gently, followed by centrifugation for 5 min at 14 000 rpm to separate the methanol/water (upper) and methanol/chloroform (lower) phases. An aliquot (320 µL) of the upper phase, which preferentially partitions the more polar metabolites, was transferred to a fresh tube and dried overnight at 30°C in a Vacufuge (Eppendorf). To derivatise the sample for gas chromatography, the dried pellet was resuspended by vortexing in 50 µL pyridine containing 20 mg/mL methoxyamine HCL (to protect carbonyl moieties by methoxylation), and then incubated at 37°C for 2 h with orbital shaking at 1100 rpm. After a brief centrifugation to settle condensation, 10 µL n-alkane standards mixture (C12, C15, C19, C22, C28, C32, and C36 - used to determine retention time indices in GC analysis) and 70 µL MSTFA were added, and followed by further incubation at 37°C for 30 min, also with shaking. Samples were then filtered through compacted tissue paper to remove particulate matter, and allowed to sit at room temperature for at least 2 h to ensure complete derivatisation prior to GC/MS analysis.

For LC/MS, the 800 µL aliquot of sample extract was first dried overnight at 30°C in a Vacufuge (Eppendorf). The pellet was then resuspended by gentle vortexing in a combination of 500 µL distilled, deionised water and 500 µL cyclohexane, and then centrifuged at 14 000 rpm to separate the lower (water) and upper (cyclohexane) phases. An aliquot (400 µL) of the lower, aqueous phase, which partitions and enriches the more polar metabolites (especially phenolics), was transferred to a fresh tube, dried to 150 µL in a Vacufuge (Eppendorf) at 30°C, and then filtered through compacted tissue paper to remove particulate matter prior to LC/MS analysis.

5.2.2.2 Metabolite extract analysis

GC/MS analysis was conducted on a ThermoFinnigan Trace GC-PolarisQ ion trap system fit with an AS2000 auto-sampler and a split/splitless injector (Thermo Electron Co., Waltham, MA, USA). The GC was equipped with a low-bleed Restek Rtx-5MS column (fused silica, 30 m, 0.25 mm ID, stationary phase diphenyl 5% dimethyl 95% polysiloxane). The GC conditions were set as follows: inlet temperature 250°C, helium carrier gas flow at constant 1 mL/min, injector split ratio 10:1, resting oven temp 70°C, and GC/MS transfer line temperature 300°C. Following injection of a 1 µL aliquot of sample, the oven was held at 70°C for 2 min and then ramped to 325°C at a rate of

8°C/min. The temperature was held at 325°C for an additional 6 min before being cooled rapidly to 70°C in preparation for the next run. Mass spectrometry was conducted in positive electron ionisation (EI) mode, the fore-line was evacuated to approximately 40 mTorr, with helium gas flow into the chamber set at 0.3 mL/min. The source temperature was held at 230°C, with an electron ionisation potential of 70 eV. The detector signal was recorded from 3.35 min after injection until 35.5 min, and ions were scanned across the range of 50-650 mass units (mu) with a total scan time of 0.58 s.

For LC/MS analysis, a 100 µL aliquot of the concentrated aqueous phase sample was injected onto a C18 Luna column (150 × 2.1 mm, 3 µm) (Phenomenex, Torrance, CA), using a Waters 2695 Separations module (Waters, Milford, MA, USA). Separation was performed with a mobile phase linearly changing from 83% solvent A (H_2O:acetonitrile (ACN):formic acid (FA), (100:1:0.1, v/v/v), pH 2.5) to 77% solvent B (ACN:H_2O:FA, (100:1:0.1, v/v/v), pH 2.5) over 21 min, at a flow rate of 0.3 mL/min and a column temperature of 40°C. Detection was conducted using negative ionization on a Micromass Quattro Micro API triple quadrupole mass spectrometer with an APCI source (Micromass, Inc., Manchester, UK). The instrument was operated with the following conditions: source temperature, 130°C; APCI probe temperature, 500°C; corona current, 5.0 µA; cone voltage, 25 V; extractor voltage, 5 V; radio frequency lens, 0.0V. Nitrogen from a nitrogen generator (Domnick Hunter, Ltd., Tyne and Wear, United Kingdom) was used as both the cone gas (50 L/h)and the desolvation gas (200 L/h). Quadrupole-1 parameters were as follows: low mass (LM) resolution, 14; high mass (HM) resolution, 14; ion energy, 0.5 V. Quadrupole-2 parameters were as follows: LM resolution, 14; HM resolution, 14; ion energy, 3.0. Collision cell entrance and exit potential were set at 50 V. Multipliers were set at 650 V. Scan time was 1 s and interscan delay 0.02 s. Data were acquired in continuous mode. Data acquisition and instrument control were performed using Masslynx 4.0 software.

5.2.2.3 Data compiling

Peak finding, peak integration, and retention time correction for GC/MS and LC/MS were performed with the R package XCMS (Smith *et al.*, 2006). The XCMS output of integrated peaks was tested for robust integration based on the assumption that each metabolite detected by MS is represented by at least two highly correlated m/z signals.

Only m/z peaks that showed high intensity correlation (PCC >0.95) and highly similar retention time (difference in median retention time (RT) after XCMS RT correction <0.03 s) with at least one other m/z peak, were retained. Based on these criteria, groups of m/z peaks believed to originate from the same metabolite were formed and the m/z signal with the highest intensity of such a group was selected as representative signal of the corresponding metabolite. The accuracy of XCMS was verified visually with the deconvolution algorithm embedded in NIST AMDIS.

5.2.2.4 Metabolite identification

National Institute of Standards and Technology (NIST) MS-Search software equipped with the NIST mass spectra, as well as the Max Planck Institute Trimethylsilane (TMS) (http://www.mpimp-Golm.mpg.de/mms-library/index-e.html), Gölm Metabolome Database (http://csbdb.mpimp-Golm.mpg.de/csbdb/gmd/gmd.html) (Kopka *et al.*, 2005) and our own (Mansfield UBC laboratory) TMS derivatised mass spectral libraries (containing 513 known compounds) were collectively used to identify metabolites of interest, as highlighted by the statistical analyses of GC/MS metabolite profiles. Identification of metabolites in LC/MS chromatograms was based on retention time and mass spectral (particularly molecular ion MW) matches with chemical standards analysed on site.

5.2.3 Determination of lignin composition by thioacidolysis

For each sample, 10 mg ground, extract-free, oven-dried wood flour was weighed into a glass 5 ml vial with teflon-lined screw-cap (Wheaton). One mL of freshly made reaction mixture (10% boron trifluoride etherate and 2.5% ethanethiol, in recently distilled dioxane (v/v)) was added to each vial and blanketed with nitrogen gas prior to sealing. Vials were then collectively placed in a (100°C) dry heating block for 4 h, with periodic (hourly) manual agitation. The reaction was halted by placing the reactions at -20°C for 5 min. To each vial were added 0.2 mL internal standard mixture (5 mg/mL tetracosane in methylene chloride), and enough 0.4 M sodium bicarbonate to bring reaction pH to between 3 and 4 (~0.3 mL, as determined by pH indicator paper). To extract the reaction products from the aqueous mixture, 2 mL distilled, deionised water and 1 mL methylene chloride were added to each vial, which was then recapped, vortexed, and allowed to settle and phase separate: upper (aqueous) and lower (non-aqueous, and containing lignin-breakdown products) phases. An aliquot (1.5 mL) of the non-aqueous

phase was taken by autopipette, and simultaneously cleared of residual water and filtered by passing through a Pasteur pipette packed with a compacted tissue paper plug and an inch of granular anhydrous sodium sulphate, and transferred directly into a 2 mL polypropylene microfuge tube. Samples were then collectively evaporated to dryness in a Vacufuge (Eppendorf) (approximately 1.5 hr at 45°C), and resuspended in 1 mL methylene chloride. Samples were derivatised by combining a 20 µL aliquot of the resuspended sample with 20 µL pyridine and 100 µL N,O,Bis(trimethylsilyl) acetamide (Sigma). After incubation for at least 2 h at 25°C, a 1 µL aliquot of this reaction was analysed by gas chromatography (GC). For complete details refer to Appendix C.

Gas chromatography was conducted on a Hewlett Packard 5890 series II instrument, fitted with an autosampler, splitless injector, flame ionising detector (FID), and 30 m RTX5ms 0.25 mm ID capillary column. One microlitre injections were processed using helium as a carrier gas at 1 mL/min. Inlet and detector temperatures were set to 250°C, with the oven profile consisting of: initial temperature 130°C, hold 3 min, ramp temperature 3°C/min for 40 min to give a final temperature of 250°C, hold 5 min, and then cooled to 130°C. Peak identification for *p*-hydroxyphenyl-, guaiacyl- and syringyl-derived monolignol moieties was consistent with Rolando *et al*. (1992).

5.2.4 Estimation of Klason lignin via NIR-based modeling

The values for wood total lignin content reported in the results are estimations calculated by a predictive model. This model was based on the combination of wood near infra-red reflectance (NIR) data, and total lignin contents determined by a modified Klason method (Huntley *et al.*, 2003), for a large, unrelated set of 623 hybrid poplar individuals. Via this model, the measurements normally taken by Klason analysis could be estimated by recording NIR spectra and submitting them to the predictive model, thus circumventing the time and resources required by actual Klason analysis.

The light reflectance of wood samples across the near infra-red spectrum was measured with a Quality Spec Pro near infra-red (NIR) spectrophotometer, equipped with a round, 1.5 cm diameter sample window (Analytical Spectral Devices Inc). The wavelength scanning range was from 350 nm – 2500 nm, with 2 nm interval, interpolated to 1 nm.

Prediction modeling was conducted using the Partial Least Squares Regression (PLSR) package provided in The Unscrambler v9.1 software (Camo Technologies, Woodbridge, New Jersey), employing full cross-validation as a modeling option. Prior to PLSR, NIR reflectance data were transformed into the Savitzky-Golay first derivative, with the averaging/smoothing process spanning 25 wavelengths either side of each data point, and an order of two for the polynomial approximation process. The model generated had the following fit for a comparison between actual and predicted values: slope 0.9120; y-intercept 2.0619; correlation coefficient 0.9541; RMSEP (root mean standard error of prediction) 0.8491; SEP (standard error of prediction) 0.8498. The accuracy of this model under cross-validation suggests that estimation of total lignin content by this means carried with it less than a five percent reduction in accuracy, compared to actual determination by wet chemistry.

5.2.5 Statistical analysis of metabolite profiles and quantitative wood traits

Statistical reduction of the metabolite and physico-chemical quantitative wood trait datasets was carried out using a combination of SAS v9.1 software (SAS Institute, Inc., Cary, N.C.) procedures and functions of the R statistics platform (R Foundation for Statistical Computing, Vienna, Austria). All metabolite peak areas were expressed as a proportion of the internal standard compound, and normalised against the extracted, dry weight of each tissue sample used in the solvent extraction.

Where required, Student's t-test was employed to identify metabolites that showed statistically significant differences between selected tree line pairs. The α value for significance in these tests was set at 0.01 (99% confidence).

Models fitting quantitative wood traits in terms of metabolite profiles, capable of prediction, were generated using the R *glm* and *step* functions. The initial model in each stepwise search was a generalized linear model (glm) with error distribution assumed "Poisson", to permit the use of non-normally distributed data. To allow for direct comparison between metabolites that exhibited unequal range distributions, the data of each metabolite were centered and scaled using the R function *scale*. The range of models examined in the stepwise search consisted of the centered and scaled intensity values for all metabolites. The R *step* function selected important metabolite predictors of the "target" trait by a stepwise procedure which minimized the Bayesian information criterion (BIC). The relative influence of each predictor metabolite in the

model was expressed as the coefficient (in this manuscript referred to as the 'estimate' so as to avoid confusion with coefficients related to other statistical measures) of each metabolite in the model equation, which was of the form: log(mu) = (estimate 1) * (selected metabolite 1) + ...+ (estimate n) * (selected metabolite n); where 'mu' is the expected value of response in the target trait. Cross validation of the stepwise models to assess predictive accuracy was conducted using a re-fitted model generated leaving one of the samples out, and then using that model to predict the target trait response value for the excluded sample, based on the profile of its metabolite profile (R *update* and *predict* functions). This process was repeated for each sample involved, and the overall predictive accuracy expressed as the Spearman correlation between the complete sets of measured and predicted values.

Principal components analysis (PCA) was carried out on metabolite profile data using the R *prcomp* function, which enabled the extraction of sample factor scores and metabolite loading scores for each component. Sample scores in the first four components were plotted on the axes of scatter plots to generate a graphical representation of the sample-to-sample variation captured by the analysis, while the metabolite loadings in these factors were taken to indicate the relative importance of each metabolite to any trends observed in those plots.

5.3 Results

5.3.1 Data summary

The analysis of the four background genotype/transgenic construct combinations was conducted via two approaches. The first of these investigated the metabolic trends of each combination, including several independent transgenic events for each construct and a range of transgene-induced severity in phenotypic traits. Sample-wise, this involved sets of clonal replicates of all of the transformed lines generated for each genotype/construct combination, which is referred to here as the "All Lines" set. In the second approach, the shifts in metabolism caused by the same construct in different backgrounds were compared and contrasted at the metabolome and individual metabolite levels. To achieve sufficient statistical degrees of freedom, this analysis required a single representative line from each of the genotype/construct combinations to be grown with ample replication (*i.e.* ~50 individuals of each line). This is referred to

as the “Select Lines” set. Both scenarios included P717 and P39 wild-type backgrounds as references in the analyses. The sample structure of these experiments, along with summaries of phenotypic data for lignin content and composition, is presented in Table 5.1. For the All Lines data set, phenotypic severity for C4H::F5H transformants was primarily graded in terms of the molar ratio of syringyl lignin monomer composition, while for C3′H-RNAi lines, secondary xylem cell wall total lignin content (as estimated by wood NIR-based modeling of Klason analysis data) was the defining attribute. For P39 C4H::F5H, the syringyl monomer content of the most severe phenotype was 12.61% greater than for wild-type, while for P717 C4H::F5H, this difference was 24.92%. Similarly, for P39 C3′H-RNAi, the total lignin content of the most severe phenotype was 12.15% less than for the wild-type, while for P717 C3′H-RNAi this difference was 10.12% (based on total cell wall composition by weight). For the Select Lines set, transformants harbouring the same construct in different genetic backgrounds were selected based on both a high level of phenotypic severity, and a similar level of severity between the backgrounds. Of note is the greater line count and phenotypic spread observed for transgenics having P717 as the genetic background. This bias arose simply from differences in the rate of line recovery from the transformation process.

The metabolite compositions of all samples of developing xylem were analysed by both GC/MS and LC/MS. Once compiled, the GC/MS profiles consisted of 221 distinct metabolite peaks across all samples, 93 of which could be identified with certainty and a further 44 whose molecular class could be assigned (Appendix B.1). Equivalent LC/MS profiles consisted of 52 metabolites, of which 9 were tentatively identified (Appendix B.2).

5.3.2 All Lines dataset analysis

Principal component analyses based on GC/MS and LC/MS data were conducted separately for all lines of each genetic background/construct combination. Factor score plot arrays for these analyses are presented in Figure 5.1 for the C4H::F5H construct, and Figure 5.2 for the C3′H-RNAi construct. In these plots, individual sample markers are coloured as per gradients based on phenotypic severity. It is immediately apparent that a metabolic trend of sample distribution based on syringyl monomer content is not present for C4H::F5H in the P39 background, in either GC/MS- or LC/MS-derived

profiles (Figure 5.1a,b); however, some evidence of a weak trend is seen for GC/MS profiles in the P717 background on account of principal components (PC) 3 and 4 (Figure 5.1c), and a more pronounced, yet still fairly indistinct trend for LC/MS profiles, primarily on account of PC-3, but augmented by PC-1 (Figure 5.1d). For the C3′H-RNAi construct, trends based on total lignin content are notably stronger. In P39, samples from the lone transformed line exhibiting a strong phenotype are clustered towards the extremes of the most significant component, PC-1 (Figure 5.2a,b). In the LC/MS profiles this effect is augmented by PC-2 and PC-3, while in GC/MS profiles PC-2 appears to distinguish an alternative sample subset based on some factor other than total lignin content. C3′H-RNAi in the P717 background shows the strongest trends of all genetic background/construct combinations. For GC/MS profiles, a separation of samples in PC-1 is observed, which is very clearly associated with the gradient in total lignin content. Principal components 3, 4 and to some extent 2 appear to play roles in focusing sample spread as total lignin content decreases. In the LC/MS profiles a strong, but not quite so emphatic trend also exists, and is derived predominantly from PC-2, with minor contribution from PC-1. Across the board, loading scores for individual metabolites were low (data not shown), with only the occasional metabolite loading higher than 0.32, the level at which variance in a metabolite variable accounts for ten percent of the variance in the specific principal component. It was not possible to single out a subset of metabolites with an over-arching influence on the patterns observed in the PCAs, even when strong patterns factor score were observed. This result implies that a large proportion of the metabolites analysed have only a small relationship with the modified phenotype, and that it is their cumulative effect that facilitated the emergence of the trends observed.

A stepwise modeling procedure, based on the Bayesian Information Criterion (BIC), was used to generate linear equations that modeled specific phenotypic traits in the C4H::F5H and C3′H-RNAi transformed lines, based on GC/MS and LC/MS metabolite profiles. The performance of these models in predicting the target trait was assessed via complete cross-validation, which involved using the model expressed in terms of all samples but one to predict the target trait in that individual, repeated for all individuals. In this situation, higher accuracy suggests a more clearly defined linear relationship between collective metabolite abundances and the target trait. The

outcome of such modeling in the collective P39/P717 C4H::F5H transformant lines is presented for syringyl monomer content and lignin S:G ratio (Figure 5.3). When modeling either trait, models based on the GC/MS metabolite data were very accurate, and notably more so than their LC/MS-based counterparts. For the collective P39/P717 C3′H-RNAi transformants, linear modeling outcomes are presented for *p*-hydroxyphenyl monomer content and total cell wall lignin content (Figure 5.4). The observations in this case are similar to those for the models of C4H::F5H, with GC/MS-based models out-performing their LC/MS-based counterparts once again, and the model for *p*-hydroxyphenyl monomer content generated from LC/MS profiles being particularly weak. Additionally, increased variance in predictions of total lignin content should be expected, given that in this case the trait data employed to build the linear model was itself a predicted estimate, which is subject to error of its own.

A deeper investigation of the modeling behaviour of the primary phenotypic severity traits for C4H::F5H (syringyl monomer content) and C3′H-RNAi (estimated total lignin content) lines is presented in Table 5.2. From these data, it is apparent that for both constructs, rather different numbers of metabolites played significant roles in the ensuing models, depending on whether genetic backgrounds were dealt with collectively or each separately. Indeed, there appears to be a loose relationship between the number of samples processed and the number of significant metabolites, although the same could be said for the degree of variance embodied by those sample sets. Also, a higher proportion of submitted metabolites were significant for LC/MS- than GC/MS-based models, and of these, the common proportion between models separately derived from lines in the P39 and P717 backgrounds, was also greater. Overall, the crossover between GC/MS-based models for the two backgrounds was surprisingly low.

5.3.3 Select Lines dataset analysis

A second round of PCA was conducted, this time on the Select Lines sample set. Again, sample distribution by principal component was presented in the form of sample factor score plots. The first comparison was between the two genetic backgrounds, P39 and P717 (Figure 5.5). For both GC/MS and LC/MS metabolite profiles, the distinction between the two hybrid genotypes was clear, although there was a small degree of overlap between sample clusters. For GC/MS analysis the dominant

distinguishing components were PC-1 and PC-4, whereas for LC/MS PC-2 was strongest with contributions from PC-1 and PC-3; however, specific component identities aside, the separation patterns were very similar between the GC/MS and LC/MS metabolite profiles. The next comparison was between the C4H::F5H modified lines from the two genetic backgrounds, including the backgrounds (control trees) themselves (Figure 5.6). For GC/MS profiles, although the separation between P39 and P717 wild-type trees was largely retained in combinations of PC-1, PC-3 and to some extent PC-4, neither of the modified lines appeared to deviate from their backgrounds. While this was also generally true of the LC/MS profiles, in this case, PC-2 cleanly distinguished between backgrounds, and both modified lines showed some signs of deviation away from their backgrounds in PC-3. This was most evident in the plot combining these two components. Furthermore, both PC-1 and PC-4 appeared to make some distinction between the P39 background and its C4H::F5H modified line, but not in the case of their P717 counterparts. The situation for the similar analysis of C3′H-RNAi (Figure 5.7) was different. In this case, the factor score plots from GC/MS profiles presented a convergence of the two background genotypes upon modification with this construct. Although not very well defined, PC-1 tightly clusters the P717 C4H::F5H line at the extreme end of a P717 wild-type “tail”, while PC-4 does the same for the P39 background and its modified line. Interestingly, combinations of PC-1 with either PC-4 or PC-2 superimpose the samples of the two modified lines, giving the impression of metabolic similarity. A similar pattern, but with clearer distinction based on genetic background, is observed for the LC/MS profiles. In this case, PC-2 cleanly separates samples by genetic background. Principal component 3 separates the P39 modified line, and to some extent the P717 modified line, from their backgrounds, while PC-1 similarly separates the P717 modified line, and to some extent the P39 modified line, from the backgrounds. Consequently, the combined plot of PC-1 and PC-3 generates a collective and complete separation of the modified lines from their associated wild-type backgrounds, and the combination of PC-2 with either of these gives tangible distinction between the two backgrounds and between backgrounds and modified lines. Indeed, a three dimensional arrangement of PC-1, PC-2 and PC-3 would yield clustering and complete 3D separation of all four elements included in this particular analysis. As was the case with the All Lines sample set, the component

loadings for individual metabolites in the PCA of the Select Lines set were very low across the board (data not shown). No single metabolite explained ten percent or more of the variance in any significant principal component in the analysis of GC/MS profiles, neither in the analysis of wild-type backgrounds or in that of either genetic construct. This was similar for the analysis of LC/MS profiles, with but few unidentified metabolites very occasionally loading right at the ten percent variance cut-off.

The lack of definition in the loading aspect of PCA necessitated further investigation in order to better define the metabolic distinctions between the P39 and P717 hybrid poplar backgrounds, the modified lines and their backgrounds, and the commonalities in lines of different genetic backgrounds modified with the same genetic construct. To this end, a summary of the results of a t-test-based analysis is presented in Table 5.3. Over one third of metabolites resolved by GC/MS, and nearly two thirds of those resolved by LC/MS showed significant differences between the two hybrid poplar backgrounds. Just under a quarter of GC/MS metabolites and half of LC/MS metabolites were significantly different in the P39 C4H::F5H modified line, compared to the P39 wild-type. In the case of the P717 C4H::F5H lines, a much smaller proportion, less than three percent of GC/MS- and one fifth of LC/MS-derived metabolites, were different. Consequently, very few differential metabolites were common between the two lines modified with this construct and, of these, only one could be identified. The situation was more balanced in the comparison of C3′H-RNAi modified lines with their backgrounds and each other. One fifth of GC/MS- and nearly half of LC/MS-derived metabolites were differential between the P39 C3′H-RNAi line and its background, and for P717 C3′H-RNAi the numbers were close to half, and almost three quarters, respectively. In this case, a fair portion of the collective differential metabolites (approximately one third for both GC/MS and LC/MS) were common to both modified lines and, of these, a fair number could be identified. Detailed listings of the differential metabolites from the three comparisons outlined are presented in various forms in Tables 5.4, 5.5, and 5.6, as well as in Appendices B.3 and B.4. The implications of these data will be addressed in the discussion.

5.4 Discussion

5.4.1 Metabolomics of phenotypic ranges

The principal component analysis of metabolite profiles from the All Lines sample set was intended to reveal global metabolic trends related to transgene-induced phenotypic severity. In the case of the C4H::F5H construct, which influences the balance between syringyl and guaiacyl lignin in favour of the syringyl moieties (Huntley *et al.*, 2003), the metabolic patterns observed were neither as consistent, nor as clearly defined as might have been expected in light of previous analyses (Robinson *et al.*, 2005). The complete absence of phenotype-related gradients in the P39 C4H::F5H modified lines was in contrast to the evident, if somewhat indefinite, gradients observed in the P717 counterparts. If the view is taken that there should at least have been some distinction made between wild-type and modified P39, then one possible explanation might be that the analysis had been limited by phenotypic spread. Even for the P717-based lines, for which analysis involved twice the number of samples with twice the phenotypic (syringyl monomer content) displacement, the gradient was far from clean. Therefore, it is possible that the variance in metabolite profiles from the P39 lines was not enough for such a trend to emerge in PCA. In the P717 C4H::F5H lines the greater definition in the metabolic gradient of LC/MS, compared to GC/MS metabolite profiles, may have originated from statistical and/or analytical factors. From the calculative perspective, the statistical analysis of GC/MS data involved many more metabolite variables than for LC/MS data, and if the metabolic effects of the modification were limited in scope, then the large number of unaffected metabolites could raise the “noise” level and confound the analysis; however, the class partitioning of metabolites between GC and LC analyses may have played a more important role. The C4H::F5H construct operates downstream in secondary metabolism, and many of its detectable downstream effects likely involve larger, more complex metabolites that are more amenable to resolution by liquid, rather than gas chromatography. In any case, the observation of metabolic gradients related to the transgene-induced phenotype in P717 validates this concept at the metabolomic scale.

The results for the PCAs of C3′H-RNAi modified lines further support the idea of broadscale, transgene-induced metabolic gradients. Although a smooth metabolic

gradient was very unlikely for the P39 lines, given the particularly disjointed structure of the phenotypic spread, the evident “peripheral” clustering of samples from the one line exhibiting a strong phenotype, for both GC/MS and LC/MS profiles, was a positive outcome. In the case of the P717 C3′H-RNAi modified lines, a situation involving many lines and a well graduated phenotypic spread, the evidence for a strong association between the phenotypic and metabolic gradients was clearly evident. That the gradients are so clear in this case is likely indicative of the nature of the genetic modification. The severe down-regulation of the native C3′H gene by RNAi-suppression has extensive implications for cell wall structure and function, cellular metabolism and whole-plant form and physiology, which goes beyond the obvious influences on lignin biosynthesis. The marked reduction in lignin biosynthesis and radical changes in lignin composition undoubtedly influence the biosynthesis of other cell wall polymers as well as fundamental balances in primary metabolism as they pertain to developing xylem sink tissue. It is therefore likely that the clean phenotype-linked gradients represent extensive shifts in global metabolism.

The initial impression given by the results is that the stepwise linear modeling of phenotypic traits, based on metabolite profiles, was an effective means of extracting the elements of metabolism that were associated with the target trait. This was more evident for models built upon GC/MS data. There was a clear demonstration that the phenotypic influence of the two constructs could be accurately modeled from metabolic data collectively pooled from both genetic backgrounds. This, in itself, confirmed at least some degree of consistency in the effects of these constructs in the two backgrounds. However, deeper investigation of the metabolite structure of such models, when based on modified lines from the individual backgrounds, revealed that the proportion of significant metabolites that were common between backgrounds was not large, for either construct. The number of metabolites included in a model is heavily weighted in favour of the P717 modified lines. This suggests that, as with the PCA-based comparisons between modified P39 and P717 lines of the All Lines sample set, the lack of comparable phenotypic ranges, and, presumably, metabolic variance, made it difficult for the linear models to be used to directly compare the influence of each genetic background on the metabolic effects of the transgenic constructs.

5.4.2 Direct genetic background comparison

To properly compare the metabolic effect of the C4H::F5H and C3′H-RNAi constructs, operating in the two different genetic backgrounds, it was necessary to assess P39 and P717 transformants within the same statistical analyses. To this end, the Select Lines sample set was comprised of the two wild-type backgrounds and a single P39 and P717 transformant line for each of the transgenic constructs. These lines were selected for their strong, yet fairly well-matched phenotypic severities. Specifically, for the C4H::F5H construct, the P39 and P717 transformants selected had average lignin syringyl monomer contents of 84.26% and 84.78%, respectively. For the C3′H-RNAi construct, the P39 and P717 transformant lines selected had average estimated total lignin contents of 15.99% and 17.05%, respectively.

The first point of note was the solid distinction that PCA made between the metabolite profiles of the two wild-type hybrid poplar genotypes, and this was particularly clear for LC/MS profiles. Although the PCA was not clear in terms of which metabolites were responsible for this distinction, the findings of a stringent t-test analysis that large proportions of both GC/MS and LC/MS metabolites were differential substantiated the distinction. Within this list of metabolites, the confirmed identities spanned major metabolite classes, and included a participant in the tricarboxylic acid cycle (malic acid), a host of small organic and amino acids, primary carbohydrates such as fructose, glucose phosphate and inositol, *etc*. This division between the two hybrids at the metabolic level is an important aspect of this study, because it provides an appropriate foundation for properly comparing the influence of background on the relationships between metabolic and specific phenotypic traits.

The genetic background comparisons for the constructs revealed similar patterns for C4H::F5H and C3′H-RNAi, although in C3′H-RNAi lines this was better defined. In the PCA, the separation of lines based firstly on genetic background, whether or not they were transformed, was consistent with the notion that the background has a fundamental influence on the metabolic profile effected by transgenic constructs, or associated with particular wood-related phenotypes. Furthermore, co-ordinated (*i.e.* occurring together in the same direction in common components) separation of P39 and P717 transgenics away from their backgrounds would suggest that the nature of the construct has its own characteristic influences on metabolism, regardless of

background. Such behaviour was most evident for transformants harbouring the C3′H-RNAi construct, particularly in LC/MS profiles. The different degrees to which this was observed for the two constructs, and between GC/MS and LC/MS profiles may be indicative of construct mode of action, target metabolism, and metabolite partitioning in chromatography, as was also noted for the PCA analysis of the All Lines sample set. It seems likely, however, that the selection of lines exhibiting fairly mild phenotypic severity, in order for a matched pair to be studied, has been an important limiting factor in this analysis of C4H::F5H. In particular, the P717 C4H::F5H line selected showed minimal metabolic differentiation from its background in terms of individual metabolite comparisons, especially in GC/MS profiles.

The simplification of the sample structure, from including All Lines to including only the Select Lines, did not improve the loading scores of metabolites in individual principal components of PCA. As such, the best insights into the finer metabolic properties of transformants were gained from t-test based analyses.

For lines transformed with the C4H::F5H construct there was one metabolite that was unexpectedly absent from those that were differential between wild-type and modified lines. In a previous analysis of C4H::F5H modified P717, pools of metabolites believed to be intermediates in the phenylpropanoid and lignin specific pathways were not detected, except for an exceptionally small pool of sinapyl alcohol that was somewhat larger in the modified line with the most extreme phenotype (line P717 C4H:F5H-64) (Robinson *et al.*, 2005). In the current analysis there was evidence that this metabolite was present in the GC/MS chromatograms of at least some of the P717 C4H::F5H-64 samples from the All Lines sample set (data not shown), but the small peak could not be discriminated from the large inositol peak that co-eluted with it. In the Select Lines sample set, the two C4H::F5H lines included had been selected for their matching phenotypic severity, and the lack of a matching P39 C4H::F5H line meant that it was not possible to include the extreme P717 C4H::F5H-64 line. Given this, and the fact that a difference was previously barely detectable in the extreme phenotype (Robinson *et al.*, 2005), it seems unlikely that a significant difference in the abundance of sinapyl alcohol would have been observed between wild-type and the select C4H::F5H lines even if this metabolite had been resolved.

The continued general absence of the proposed intermediates in monolignol biosynthesis, from metabolite profiles of wild-type and C4H::F5H or C3′H-RNAi modified lines, lends support to the existence of metabolic channels in this pathway. As previously suggested in closely related work (Achnine *et al.*, 2004; Anterola *et al.*, 1999; Rasmussen and Dixon, 1999; Robinson *et al.*, 2005; Winkel-Shirley, 1999), it appears that many intermediates in this pathway may be covalently bound to, and passed between sequential active sites of multi-enzyme complexes. This sort of arrangement is proposed as an appropriate mechanism for sparing cellular solvent capacity, maximising the efficiency of metabolic pathways, and reducing the liberation and pooling of metabolites with cytotoxic potential (e.g., unconjugated phenolics) (Hrazdina and Jensen, 1992; Srere, 1987; Srere, 2000).

The most noteworthy feature of the lists of metabolites common to P39 and P717 transformants is that the proportional change in abundance of all of these specific metabolites, relative to their respective backgrounds, are similar regardless of background. Where a construct induced an increase or decrease in one background, the same was true in the other. This, of course, was most evident in the analysis of C3′H-RNAi lines, in which many differential metabolites were common to both backgrounds. In this list, positively identified metabolites included representatives of the TCA cycle (succinic and malic acid pools decreased approximately 70% and 50%, respectively), other small acids (a host of metabolites including ribonic, gluconic, glucaric and galactaric acids all decreased considerably), and carbohydrate source molecules and precursors of glycan cell wall polymers (the glucose 6-phosphate pool decreased by around 50%). Such changes are likely indicative of a whole-plant reduction in fitness and metabolic activity, itself suggested by the severely altered wood and growth traits in these lines (Coleman *et al.*, 2008a). Furthermore, several larger metabolites in GC/MS and LC/MS profiles were seen to increase dramatically in the C3′H-RNAi modified lines, and although these metabolites were not identified in this analysis, they may be the same as some of those phenolics and phenolic glycosides seen to behave similarly in a previous HPLC-based metabolite profile analysis of P39 C3′H-RNAi lines (Coleman *et al.*, 2008a).

5.5 Concluding remarks

This research attempted to characterise the interrelationships between hybrid poplar background genotype, transgenic modification, metabolite profiles and wood-related phenotypic traits. Its findings have demonstrated that transgene-induced phenotypic gradients in physico-chemical wood traits can be associated with similar gradients in the global metabolism of secondary xylem biosynthesis. This result implies that the same may be true for phenotypic gradients arising through natural genetic variation, intensive breeding, or environmental factors. It is also apparent that while distinct, at a global level the wood-forming metabolisms of different poplar hybrids can, to some extent, respond similarly to the influences of genetic manipulation of lignin-related genes. This further implies that with the correct approach, it may be possible to associate the emergence of specific wood traits from different genetic backgrounds – be they transgene-induced or otherwise – with stable metabolic signatures.

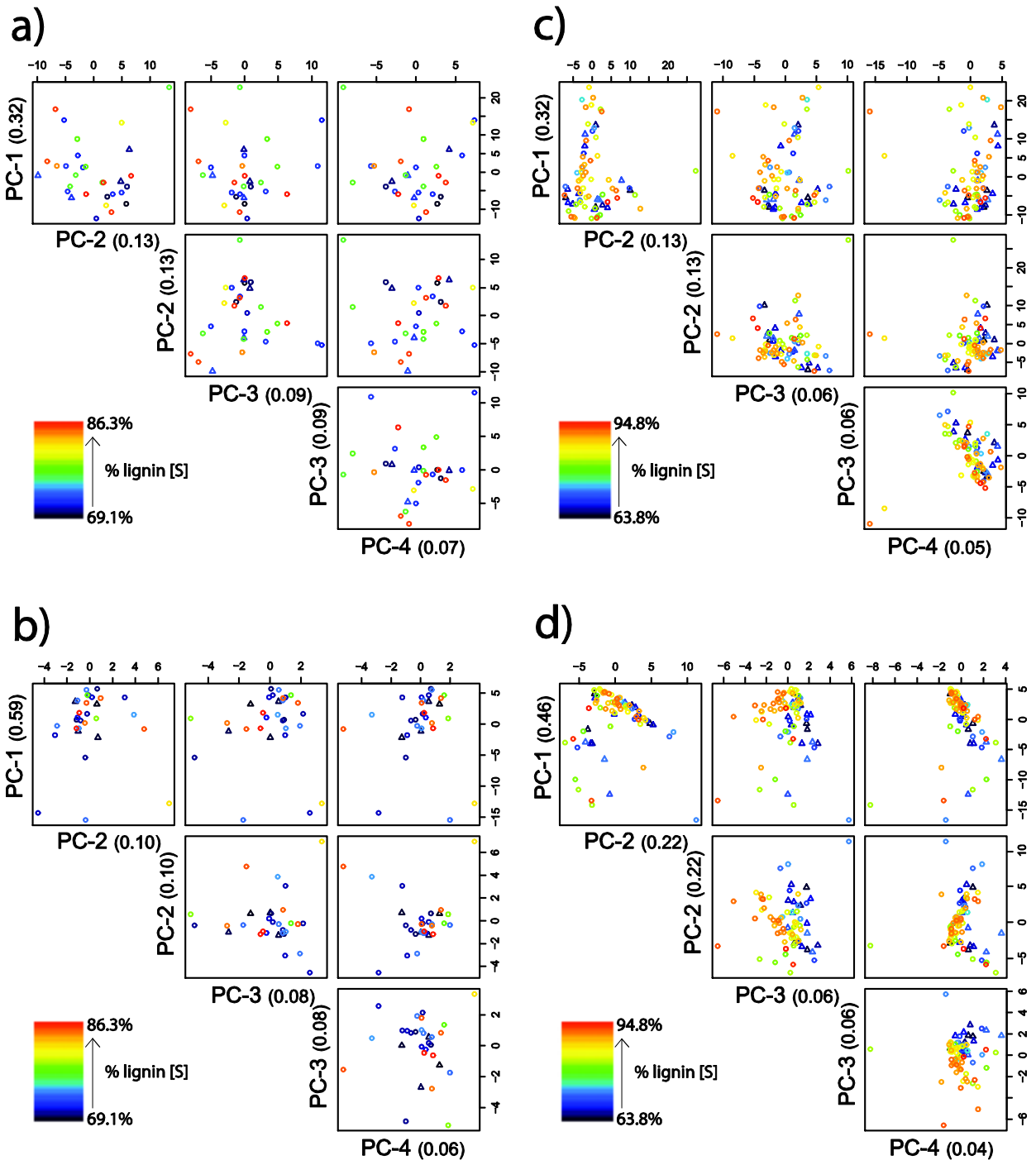


Figure 5.1 Factor score plots from principal components analysis of metabolite profiles from wildtype and multiple lines transformed with the C4H::F5H construct. a) GC/MS profiles from P39 wildtype and modified, b) LC/MS profiles from P39 wildtype and modified, c) GC/MS profiles from P717 wildtype and modified, d) LC/MS profiles from P717 wildtype and modified. Wild-type samples are represented by triangular markers, and genetically modified individuals are represented by circles. Individual markers are coloured according to lignin S monomer content of each sample, with the colour gradient spanning the phenotypic range of the sample set. Numbers in parentheses indicate proportion of dataset variance explained by individual principal components.

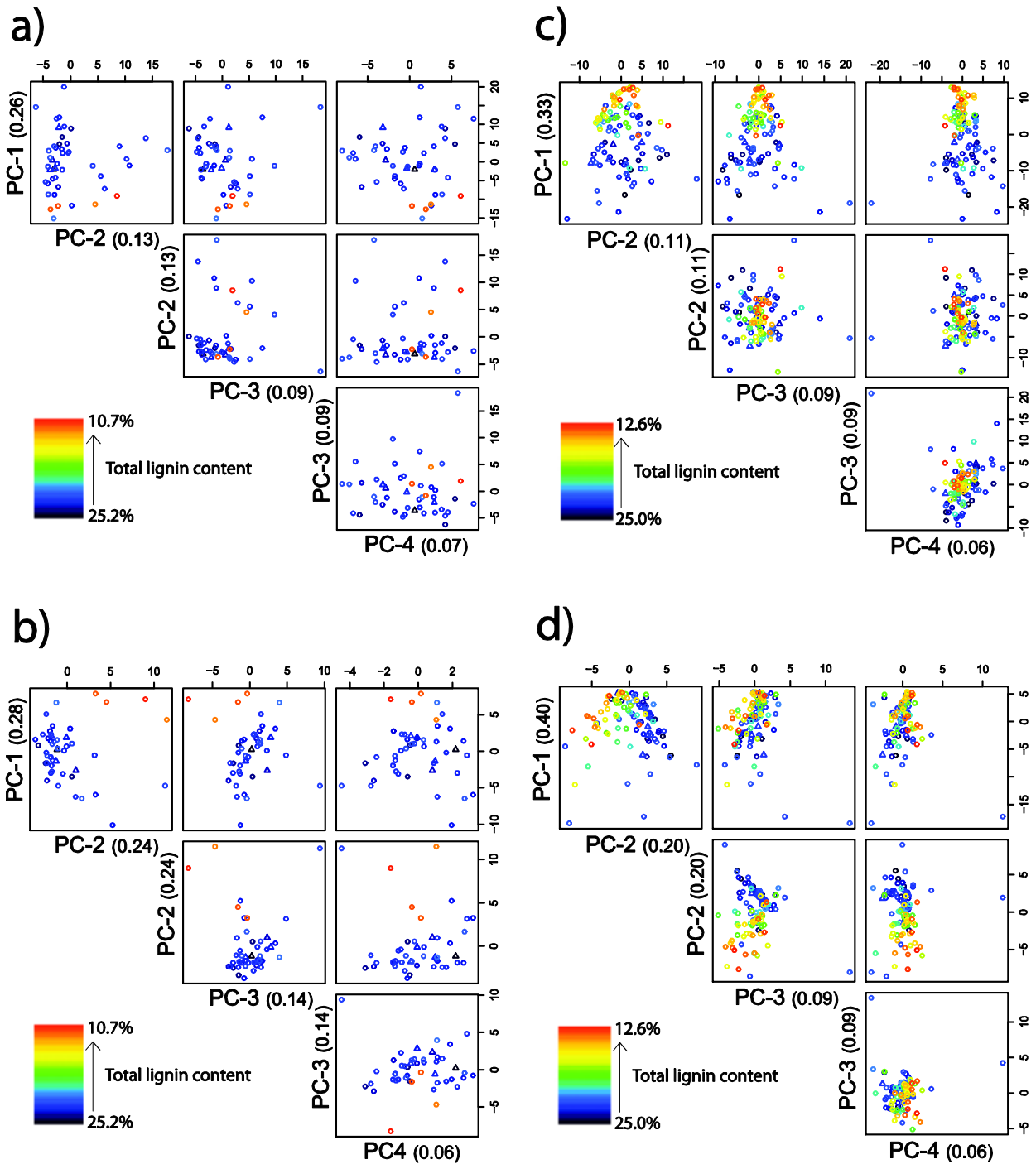


Figure 5.2 Factor score plots from principal components analysis of metabolite profiles from wildtype and multiple lines transformed with the C3′H-RNAi construct. a) GC/MS profiles from P39 wildtype and modified, b) LC/MS profiles from P39 wildtype and modified, c) GC/MS profiles from P717 wildtype and modified, d) LC/MS profiles from P717 wildtype and modified. Wild-type samples are represented by triangular markers, and genetically modified individuals are represented by circles. Individual markers are coloured according to wood total lignin content of each sample, with the colour gradient spanning the phenotypic range of the sample set. Numbers in parentheses indicate proportion of dataset variance explained by individual principal components.

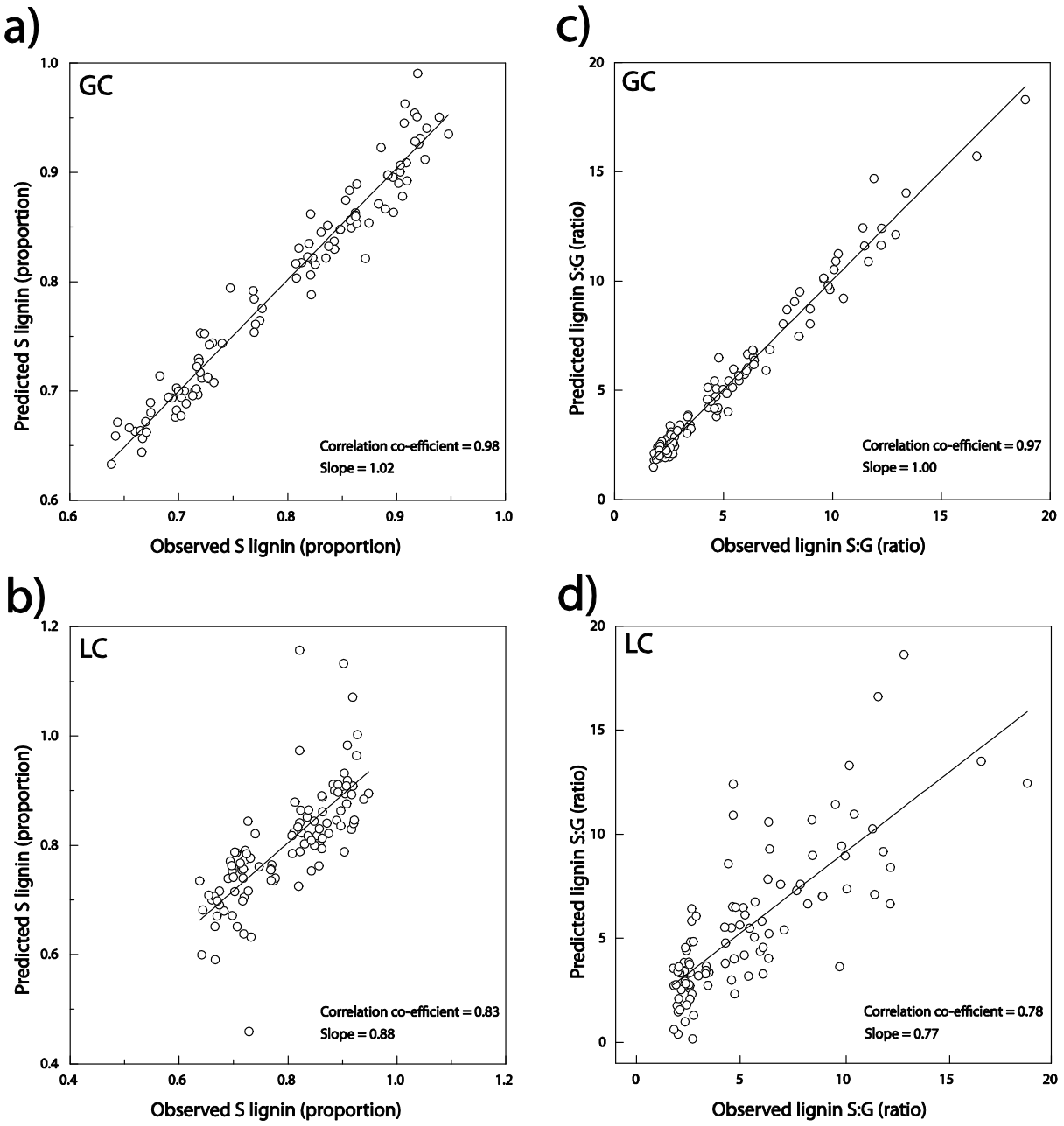


Figure 5.3 Comparison of measured versus predicted quantitative traits in C4H::F5H modified poplar. a) Lignin S monomer proportion modeled with GC/MS metabolite profile data, b) Lignin S monomer proportion modeled with LC/MS data, c) Lignin S:G ratio modeled with GC/MS data, d) Lignin S:G ratio modeled with LC/MS data. Predictions were based on linear models generated from metabolite profiles by a stepwise modeling procedure, under cross-validation. Circular markers represent individual samples. Wild-type and modified P39 and P717 samples were combined in model building and are not distinguished in the plots. Fitted line is line of best fit in a regression not constrained to the origin.

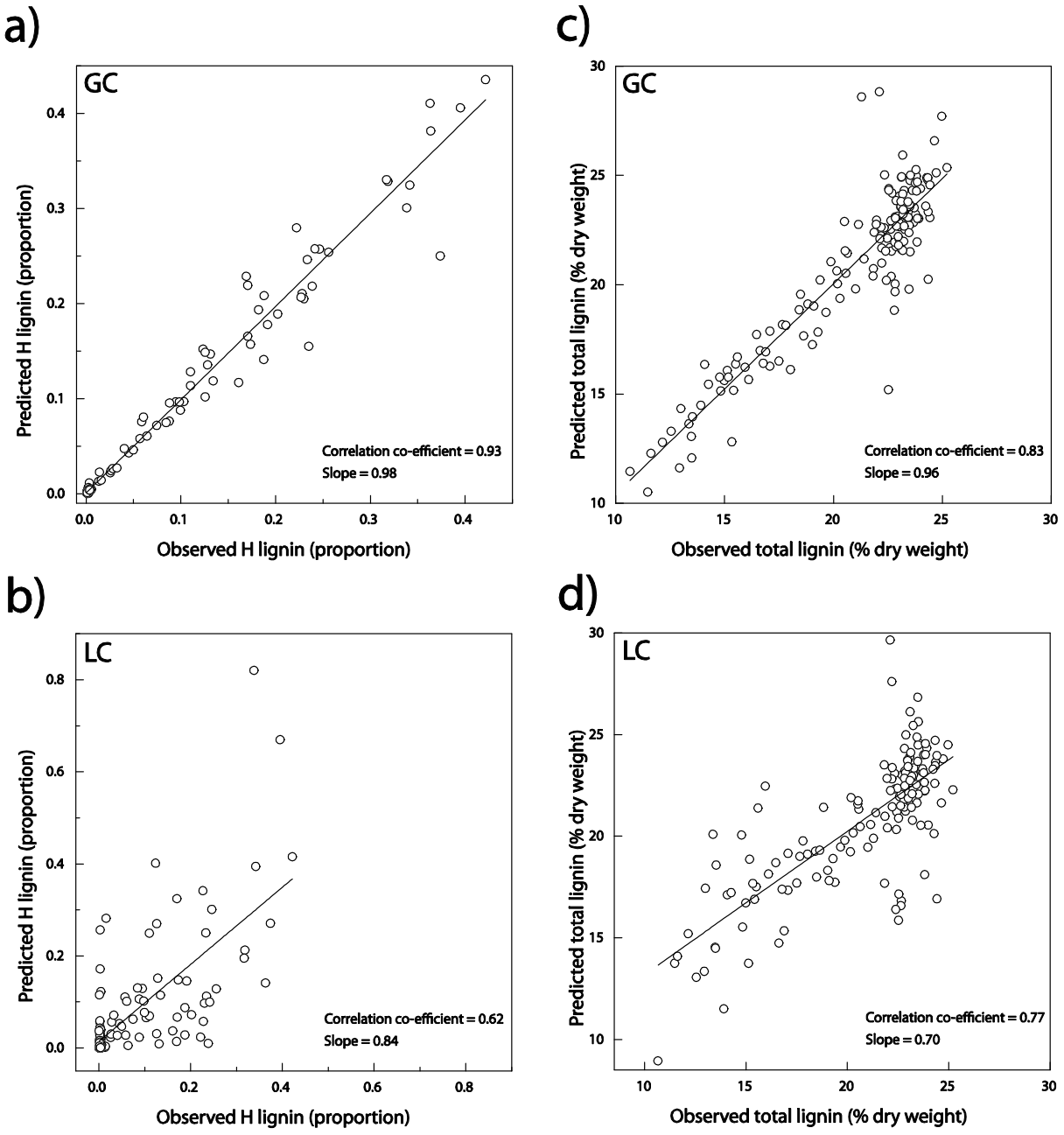


Figure 5.4 Comparison of measured versus predicted quantitative traits in C3′H-RNAi modified poplar. a) Lignin H monomer proportion modeled with GC/MS metabolite profile data, b) Lignin H monomer proportion modeled with LC/MS data, c) Total lignin content modeled with GC/MS data, d) Total lignin content modeled with LC/MS data. Predictions were based on linear models generated from metabolite profiles by a stepwise modeling procedure, under cross-validation. Circular markers represent individual samples. Wild-type and modified P39 and P717 samples were combined in model building and are not distinguished in the plots. Fitted line is line of best fit in a regression not constrained to the origin.

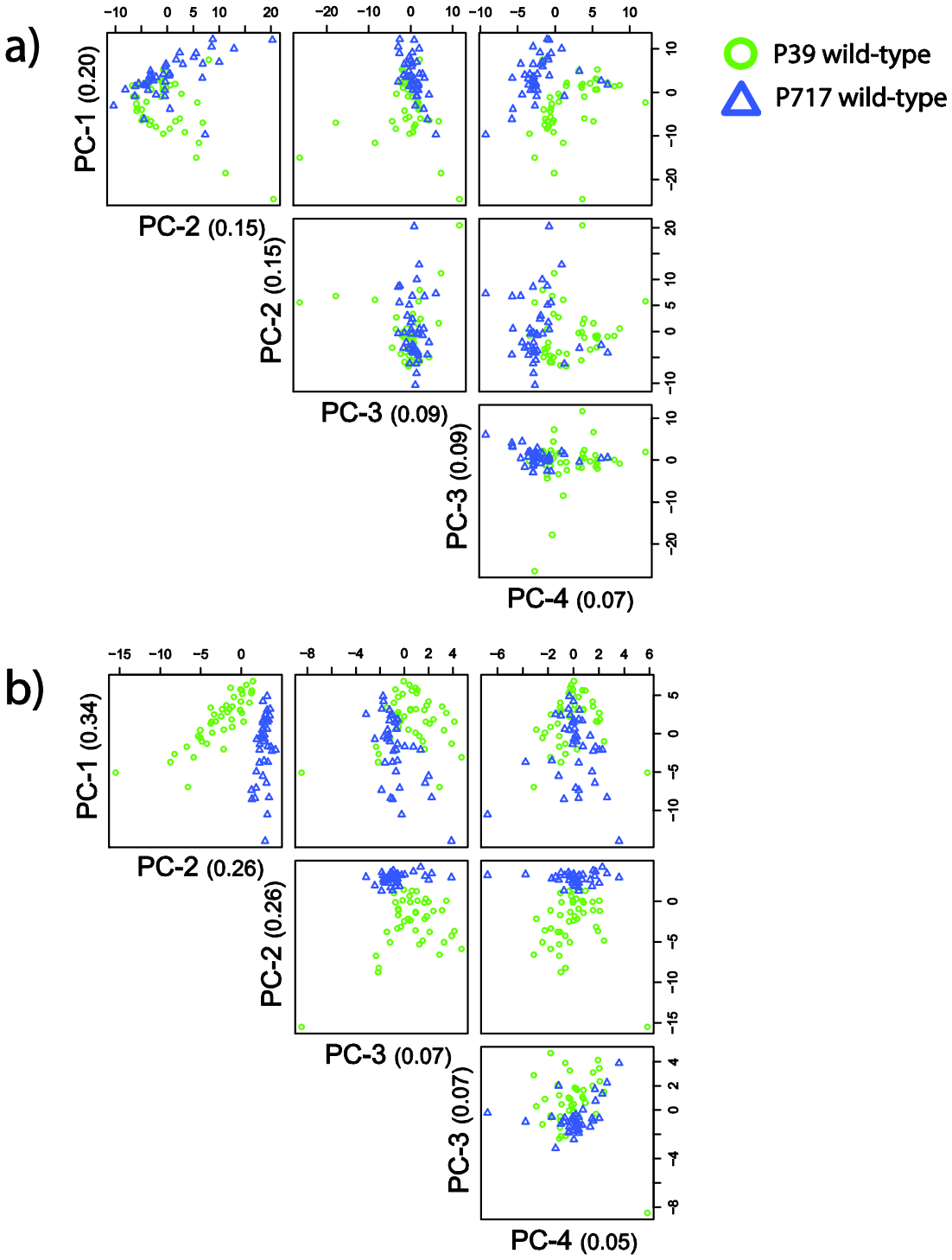


Figure 5.5 Factor score plots from principal components analysis of metabolite profiles from P37 and P717 hybrid poplar wild-types. a) GC/MS metabolite profiles, b) LC/MS metabolite profiles. Marker designation for individual samples as indicated in figure. Numbers in parentheses indicate proportion of dataset variance explained by individual principal components.

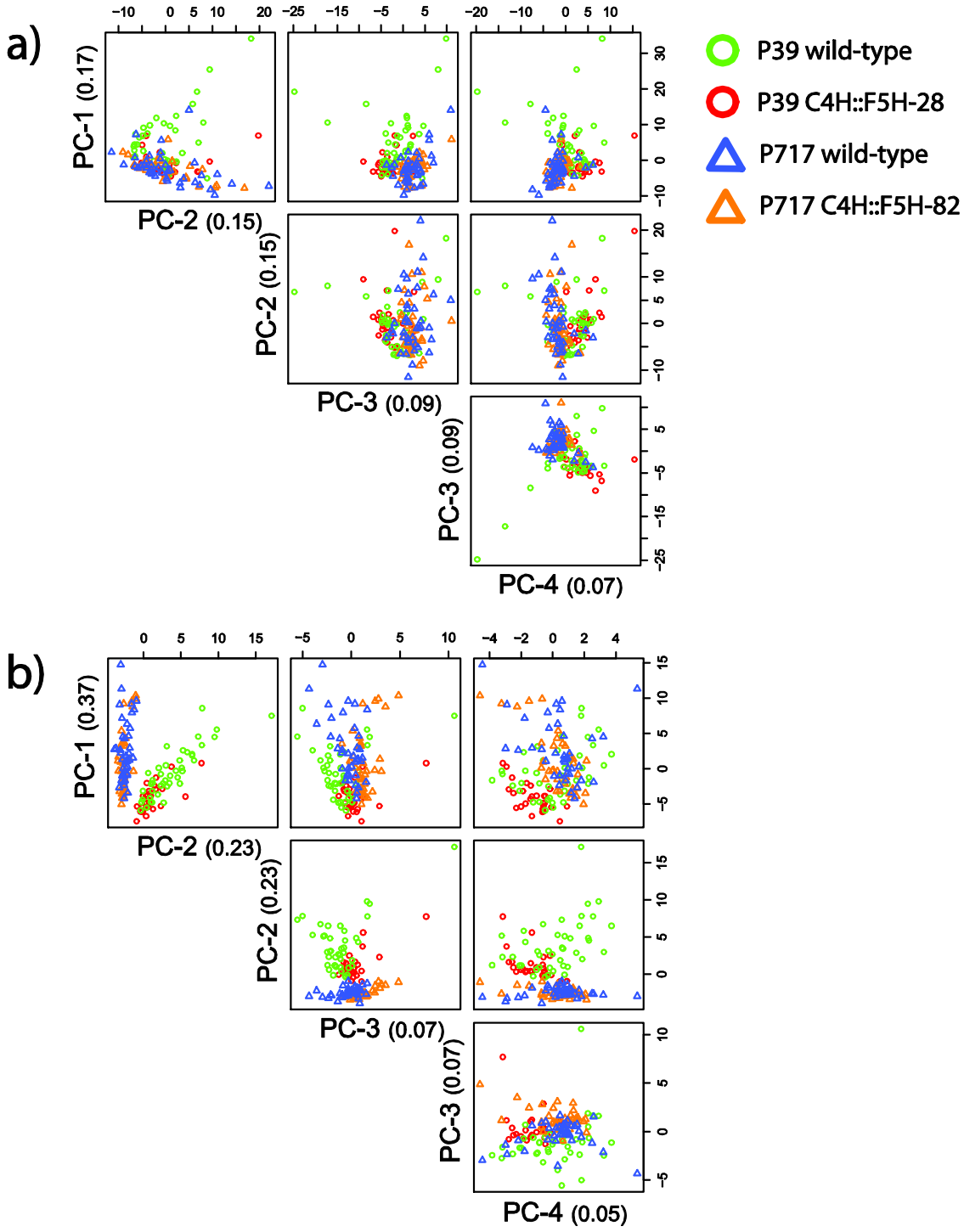


Figure 5.6 Factor score plots from principal components analysis of metabolite profiles from P37 and P717 hybrid poplar wild-types and a C4H::F5H modified line of each hybrid. a) GC/MS metabolite profiles, b) LC/MS metabolite profiles. Marker designation for individual samples as indicated in figure. Numbers in parentheses indicate proportion of dataset variance explained by individual principal components.

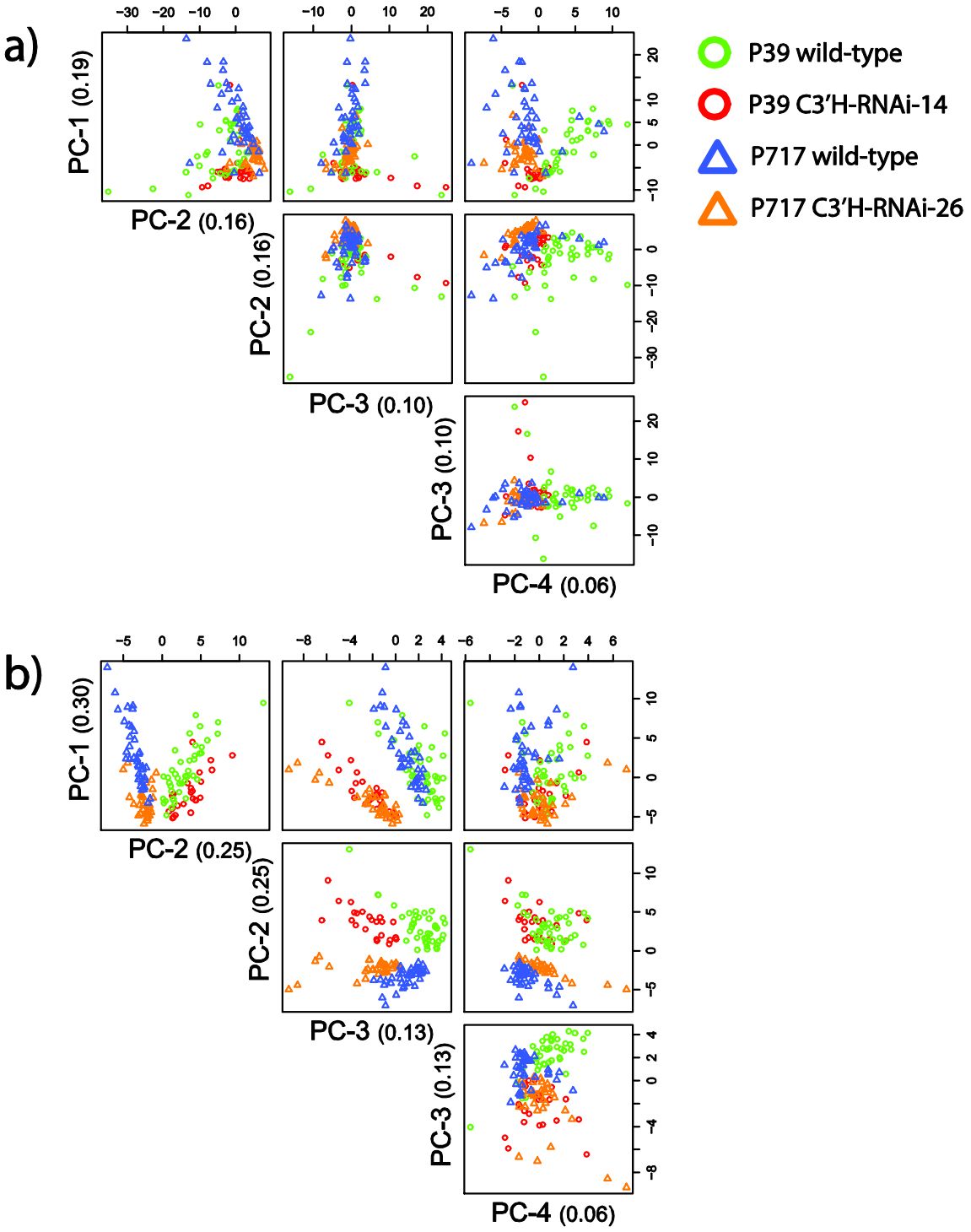


Figure 5.7 Factor score plots from principal components analysis of metabolite profiles from P37 and P717 hybrid poplar wildtypes and a C3′H-RNAi modified line of each hybrid. a) GC/MS metabolite profiles, b) LC/MS metabolite profiles. Marker designation for individual samples as indicated in figure. Numbers in parentheses indicate proportion of dataset variance explained by individual principal components.

Table 5.1. Sample structure of hybrid poplar datasets and measurements of quantitative wood traits, summarised by line. Modified lines sorted according to phenotypic severity, using lignin S monomer content for C4H::F5H and total lignin content for C3′H-RNAi, indicated by bold type. Height and base stem diameter not taken for P717 C4H::F5H lines in the All Lines dataset.

				Mean (Standard Deviation)						
Set	**Background/construct**	**Line ID**	**Sample n**	**height (cm)**	**dia (mm)**	**totallig (%)**	**thioH (%)**	**thioG (%)**	**thioS (%)**	**thioS:G (ratio)**
All lines	P39 wild-type	WT02	4	266.5 (19.9)	13.94 (1.04)	23.27 (0.75)	0.14 (0.05)	28.62 (1.31)	71.24 (1.34)	2.50 (0.16)
	P39 C4H::F5H	02	4	277.0 (30.3)	11.85 (2.40)	22.68 (1.48)	0.15 (0.04)	28.40 (0.66)	**71.45 (0.64)**	2.52 (0.08)
		04	4	298.8 (8.1)	12.49 (0.87)	23.40 (0.21)	0.14 (0.09)	28.15 (0.42)	**71.72 (0.35)**	2.55 (0.05)
		01	5	275.8 (28.2)	12.83 (1.61)	22.86 (1.72)	0.21 (0.11)	23.92 (8.90)	**75.86 (8.82)**	3.78 (2.06)
		03	4	287.3 (16.3)	13.54 (0.82)	22.96 (0.43)	0.17 (0.09)	22.57 (0.30)	**77.27 (0.33)**	3.43 (0.06)
		06	4	265.0 (38.3)	13.41 (4.86)	22.30 (0.77)	0.11 (0.02)	18.41 (5.32)	**81.48 (5.32)**	4.79 (1.68)
		28	5	253.2 (33.3)	12.12 (1.67)	22.17 (1.44)	0.14 (0.06)	16.01 (2.31)	**83.85 (2.28)**	5.34 (0.90)
	P717 wild-type	WT02	16	-	-	21.89 (1.01)	0.26 (0.08)	32.38 (2.62)	67.35 (2.60)	2.10 (0.26)
	P717 C4H::F5H	85	10	-	-	20.85 (1.63)	0.28 (0.06)	27.70 (2.08)	**72.02 (2.09)**	2.62 (0.26)
		82	12	-	-	20.55 (1.60)	0.27 (0.04)	17.23 (1.17)	**82.50 (1.19)**	4.82 (0.40)
		41	6	-	-	20.59 (0.89)	0.26 (0.06)	14.62 (1.54)	**85.12 (1.58)**	5.88 (0.66)
		21	5	-	-	19.78 (0.39)	0.24 (0.06)	14.59 (3.59)	**85.18 (3.55)**	6.17 (1.81)
		37	5	-	-	20.29 (0.51)	0.26 (0.03)	11.03 (1.16)	**88.72 (1.17)**	8.13 (0.98)
		26	5	-	-	20.90 (1.37)	0.32 (0.06)	8.63 (0.73)	**91.06 (0.74)**	10.63 (1.01)
		65	9	-	-	21.01 (1.00)	0.28 (0.05)	8.64 (1.10)	**91.08 (1.13)**	10.71 (1.46)
		64	6	-	-	22.74 (1.82)	0.34 (0.10)	7.39 (1.95)	**92.27 (1.97)**	13.32 (3.85)
	P39 wild-type	WT01	5	317.6 (26.2)	14.56 (1.31)	23.74 (0.88)	0.32 (0.12)	29.34 (0.46)	70.34 (0.4)	2.40 (0.05)
	P39 C3'H-RNAi	022	5	301.4 (38.3)	12.54 (3.34)	**23.63 (0.41)**	0.17 (0.10)	29.56 (0.46)	70.27 (0.47)	2.38 (0.05)
		515	4	301.4 (40.8)	13.60 (2.44)	**23.27 (0.27)**	0.24 (0.14)	30.06 (0.81)	69.70 (0.81)	2.32 (0.09)
		510	4	312.5 (37.1)	14.85 (1.17)	**23.09 (1.23)**	0.26 (0.15)	29.34 (0.27)	70.41 (0.29)	2.40 (0.03)
		012	5	289.2 (40.8)	13.58 (3.12)	**23.09 (0.86)**	0.18 (0.06)	30.33 (0.15)	69.49 (0.12)	2.29 (0.02)
		053	5	316.2 (43.8)	13.58 (1.50)	**22.97 (0.46)**	0.29 (0.06)	28.63 (0.75)	71.09 (0.72)	2.49 (0.09)
		064	5	314.8 (28.0)	15.52 (1.68)	**22.93 (0.26)**	0.32 (0.12)	28.80 (0.58)	70.88 (0.48)	2.46 (0.07)
		044	5	302.4 (27.6)	13.98 (1.49)	**22.76 (0.54)**	0.19 (0.09)	30.22 (0.51)	69.58 (0.45)	2.30 (0.05)
		610	4	298.0 (19.1)	14.13 (2.41)	**22.65 (0.91)**	0.27 (0.11)	29.82 (0.80)	69.91 (0.72)	2.35 (0.08)
		014	4	68.0 (26.5)	5.03 (1.60)	**11.48 (0.62)**	20.01 (2.87)	17.87 (2.21)	62.11 (1.16)	3.51 (0.43)
	P717 wild-type	WT01	4	250.5 (1.9)	11.20 (2.29)	23.14 (0.47)	0.26 (0.06)	33.63 (1.26)	66.12 (1.25)	1.97 (0.11)
	P717 C3'H-RNAi	46	5	275.2 (24.1)	12.41 (1.44)	**23.98 (0.74)**	0.29 (0.09)	33.04 (1.15)	66.67 (1.22)	2.02 (0.11)
		25	5	251.6 (35.8)	11.56 (1.51)	**23.84 (1.00)**	0.31 (0.06)	36.09 (4.36)	63.61 (4.40)	1.79 (0.30)
		10	2	274.0 (1.4)	14.66 (1.23)	**23.59 (0.31)**	1.00 (0.59)	32.45 (0.35)	66.56 (0.24)	2.05 (0.01)
		28	5	197.4 (40.7)	8.38 (2.91)	**23.26 (0.78)**	0.27 (0.04)	10.31 (0.65)	89.42 (0.65)	8.70 (0.59)
		23	4	229.5 (10.8)	9.51 (1.24)	**23.24 (0.80)**	0.26 (0.05)	11.41 (1.11)	88.34 (1.12)	7.81 (0.83)
		09	4	239.8 (14.3)	8.98 (1.38)	**23.22 (0.38)**	0.29 (0.06)	9.35 (0.52)	90.36 (0.50)	9.69 (0.59)
		50	5	257.8 (15.4)	10.72 (1.65)	**23.20 (0.77)**	0.32 (0.10)	33.09 (1.52)	66.59 (1.56)	2.02 (0.14)
		17	5	255.6 (10.4)	11.24 (1.24)	**22.93 (0.51)**	0.32 (0.10)	34.61 (1.28)	65.07 (1.28)	1.88 (0.11)
		40	5	254.4 (7.9)	10.63 (1.15)	**22.91 (0.82)**	0.24 (0.04)	10.51 (1.27)	89.25 (1.26)	8.59 (1.03)
		13	4	232.5 (16.1)	9.97 (1.18)	**22.63 (1.58)**	0.28 (0.07)	10.53 (0.51)	89.19 (0.55)	8.49 (0.48)
		07	3	251.0 (4.4)	12.27 (2.59)	**22.55 (1.00)**	2.18 (0.73)	37.97 (6.82)	59.85 (7.19)	1.63 (0.43)
		03	5	222.6 (18.8)	9.45 (2.27)	**22.03 (0.93)**	0.29 (0.03)	10.45 (0.87)	89.27 (0.86)	8.59 (0.74)
		14	3	236.7 (7.5)	10.88 (0.55)	**21.64 (0.92)**	2.98 (1.45)	32.72 (0.94)	64.30 (0.52)	1.97 (0.04)
		15	5	224.0 (9.3)	12.26 (1.74)	**20.83 (0.75)**	4.91 (1.28)	31.88 (1.44)	63.21 (1.88)	1.99 (0.14)
		32	3	203.0 (15.5)	9.95 (2.13)	**20.06 (0.88)**	7.76 (1.48)	27.61 (1.73)	64.63 (1.19)	2.35 (0.18)
		11	4	231.3 (5.9)	10.87 (2.43)	**18.54 (0.83)**	9.57 (1.55)	26.51 (0.86)	63.92 (1.78)	2.42 (0.12)
		12	5	220.8 (10.7)	11.22 (2.44)	**18.35 (1.80)**	11.08 (4.27)	27.50 (2.85)	61.42 (1.51)	2.25 (0.19)
		33	3	211.7 (9.8)	10.85 (1.93)	**18.29 (0.43)**	10.79 (1.55)	29.18 (0.86)	60.03 (0.93)	2.06 (0.05)
		01	4	256.8 (15.8)	10.43 (2.07)	**17.41 (0.97)**	14.72 (2.21)	24.67 (0.94)	60.62 (1.82)	2.46 (0.11)
		26	5	172.2 (53.5)	7.94 (1.51)	**16.72 (1.43)**	18.25 (5.26)	24.56 (4.07)	57.20 (1.98)	2.38 (0.39)
		34	5	207.0 (37.2)	9.82 (1.77)	**15.54 (0.55)**	22.86 (6.69)	20.39 (2.26)	56.75 (4.74)	2.79 (0.18)
		49	4	228.8 (20.9)	11.93 (1.67)	**14.69 (0.68)**	22.64 (2.68)	20.13 (1.20)	57.24 (1.59)	2.85 (0.11)
		35	3	179.7 (34.2)	9.41 (0.72)	**14.08 (1.09)**	28.47 (6.87)	19.94 (1.73)	51.59 (5.15)	2.59 (0.05)
		04	4	144.0 (32.3)	8.73 (2.63)	**13.73 (1.01)**	38.86 (2.56)	17.48 (1.03)	43.66 (1.80)	2.50 (0.11)
		43	3	189.3 (34.1)	7.92 (1.24)	**13.14 (0.30)**	32.47 (1.19)	17.78 (0.79)	49.74 (0.39)	2.80 (0.10)
Select line	P39 wild-type	WT03	42	140.3 (32.7)	7.50 (1.77)	23.35 (1.32)	0.15 (0.05)	28.63 (3.87)	71.22 (3.89)	2.56 (0.52)
	P717 wild-type	WT03	40	181.2 (20.9)	9.71 (2.00)	24.79 (1.16)	0.39 (0.07)	28.46 (6.61)	71.15 (6.58)	3.00 (2.24)
	P39 C4H::F5H	28	28	175.1 (20.2)	9.11 (1.59)	22.87 (1.01)	0.16 (0.03)	15.58 (1.55)	**84.26 (1.55)**	5.47 (0.66)
	P717 C4H::F5H	82	36	186.8 (8.7)	9.97 (1.59)	22.98 (0.79)	0.46 (0.05)	14.76 (0.71)	**84.78 (0.70)**	5.76 (0.30)
	P39 C3'H-RNAi	14	24	70.1 (19.7)	4.03 (0.64)	**15.99 (1.11)**	20.00 (3.05)	17.10 (1.90)	62.90 (3.82)	3.73 (0.47)
	P717 C3'H-RNAi	26	34	167.7 (15.9)	9.15 (1.50)	**17.05 (1.17)**	18.49 (3.52)	20.64 (2.55)	60.87 (2.15)	2.99 (0.35)

Table 5.2. Summary and comparison of quantitative trait linear models' structure and performance under cross-validation. a) modeling lignin S monomer proportion in C4H::F5H modified P39 and P717 poplar both together and individually, and b) modeling lignin H monomer proportion in C3′H-RNAi modified P39 and P717 poplar both together and individually. Data are provided for both GC/MS profile- and LC/MS profile-based models in each table. Analysis based on All Lines dataset.

a)

Model: Lignin S proportion	Linear model performance under cross-calidation (GC | LC)				Common metabolites (GC | LC)		
	sample n	peaks in model	corr coeff	slope	P39,P717	P39	P717
P39,P717; C4H::F5H	103 | 103	79 | 33	0.98 | 0.83	1.02 | 0.88	na	8 | 14	18 | 26
P39 C4H::F5H	30 | 30	15 | 19	0.95 | 0.91	0.90 | 0.95	8 | 14	na	2 | 15
P717 C4H::F5H	73 | 73	45 | 36	0.98 | 0.81	1.02 | 0.97	18 | 26	2 | 15	na

b)

Model: Total lignin content	Linear model performance under cross-calidation (GC | LC)				Common metabolites (GC | LC)		
	sample n	peaks in model	corr coeff	slope	P39,P717	P39	P717
P39,P717; C3'H-RNAi	153 | 153	90 | 28	0.83 | 0.71	0.96 | 0.70	na	5 | 7	28 | 19
P39 C3'H-RNAi	46 | 46	12 | 16	0.76 | 0.62	0.98 | 1.01	5 | 7	na	5 | 12
P717 C3'H-RNAi	107 | 107	61 | 31	0.92 | 0.68	0.99 | 0.67	28 | 19	5 | 12	na

Table 5.3. Summary of GC/MS- and LC/MS-detected metabolites showing differential abundances between P39 and P717 hybrid poplar backgrounds, and between C4H::F5H and C3′H-RNAi transformants and these backgrounds. Analysis based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Numbers in parenthesis are the total number of metabolites tested. "Collectively different" indicates total number of unique metabolites identified in the P39- and P717-based comparisons, combined. "Commonly different" indicates the total number of metabolites commonly identified in both the P39- and P717-based comparisons.

Genetic construct	wild-types		C4H::F5H		C3'H-RNAi	
Peak set	GC (221)	LC (52)	GC (221)	LC (52)	GC (221)	LC (52)
Different from P39 wild-type	-	-	50	27	44	22
Different from P717 wild-type	-	-	6	11	95	37
Collectively different	79	31	54	31	104	40
Commonly different	-	-	2	7	35	19
Identified "commons"	32	5	1	0	13	3

Table 5.4. List of identified differential metabolites in the comparison between P39 and P717 hybrid poplar backgrounds, based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Average abundance of metabolites in P717 are expressed relative to P39.

Comparison between P39 and P717 wild-types					
GC/MS metabolites different between backgrounds			LC/MS metabolites different between backgrounds		
Peak#	Identity	Average abundance P717 rel:P39	Peak#	Identity	Average abundance P717 rel:P39
G1_003	Pyruvic acid (1MEOX) (1TMS)	0.52	L2_013	Pinoquercetin; MW316	0.51
G1_005	Glycolic acid (2TMS)	0.68	L2_019	Catechol; MW110	1.81
G1_010	2-Pyrrolidinone (1TMS)	0.04	L2_023	Vitexin; MW432	0.50
G1_018	Urea (2TMS)	0.29	L2_036	3-ferulolquinic acid; MW368	30.40
G1_024	Ethanolamine (3TMS)	1.56	L2_041	Salicortin; MW424	1.64
G1_026	Phosphoric acid (3TMS)	1.81	L2_043	Phenyllactic acid; MW166	21.85
G1_035	Glyceric acid (3TMS)	0.50			
G1_045	3-Hydroxymyristic acid (2TMS)	1.71			
G1_048	2-Hydroxybenzyl alcohol (2TMS)	1.57			
G1_053	Malic acid (3TMS)	0.75			
G1_055	L-Asparagine (2TMS)	4.35			
G1_059	Pyroglutamic acid (2TMS)	0.66			
G1_060	4-Aminobutyric acid (3TMS)	0.65			
G1_062	L-Norvaline (3TMS)	3.64			
G1_079	4-Hydroxybenzoic acid (2TMS)	1.39			
G1_089	D-Ribonic acid lactone (3TMS)	0.27			
G1_109	Ribonic acid (5TMS)	0.59			
G1_123	Quinic acid (5TMS)	0.13			
G1_124	Fructose MEOX (5TMS)	0.29			
G1_125	Sorbose MEOX (5TMS) [BP]	0.26			
G1_126	Fructose MEOX (5TMS) [BP]	0.21			
G1_132	Galactitol (6TMS)	2.27			
G1_140	Glutamine (4TMS) rep?	0.60			
G1_143	Palmitic acid (1TMS)	1.61			
G1_145	Galactaric acid (6TMS)	1.62			
G1_151	Inositol (6TMS)	0.37			
G1_152	3-Deoxy-arabino-hexaric acid (5TMS)	1.94			
G1_169	Glucose-6-phosphate MEOX (6TMS)	0.64			
G1_187	Salicin (?TMS)	1.89			
G1_196	Sucrose (8TMS)	0.70			
G1_204	Trehalose (8TMS)	0.26			
G1_215	Digalactosylglycerol (9TMS)	0.62			

Table 5.5. Complete list of "collective" differential metabolites in the comparisons between P39 C4H::F5H and wild-type background and P717C4H::F5H and wild-type background, based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Average abundance of metabolites in modified lines are expressed relative to respective wild-type background.

Transgenic Construct: C4H::F5H			
GC/MS metabolites different between Mod and WT			
		Avg abundance rel:WT	
Peak#	Identity	P39	P717
G1_003	Pyruvic acid (1MEOX) (1TMS)	1.53	
G1_011	Unidentified G1_011	1.34	
G1_017	Unidentified G1_017	5.29	
G1_026	Phosphoric acid (3TMS)	0.45	
G1_027	Unidentified G1_027	0.20	
G1_031	Succinic acid (2TMS)	1.93	
G1_035	Glyceric acid (3TMS)	1.98	
G1_038	Fumaric acid (2TMS)	1.41	
G1_047	Unidentified G1_047	1.57	
G1_049	Unidentified G1_049	0.36	
G1_053	Malic acid (3TMS)	1.74	
G1_054	Unidentified G1_054	2.91	
G1_055	L-Asparagine (2TMS)	0.17	
G1_063	Unidentified G1_063	1.62	
G1_066	Unidentified G1_066	0.17	
G1_067	Unidentified G1_067; Organic acid	2.93	
G1_070	Unidentified G1_070	0.15	
G1_075	Unidentified G1_075; Amino acid	0.23	
G1_076	Unidentified G1_076; junk	0.25	
G1_077	Unidentified G1_076; junk	0.24	
G1_094	Unidentified G1_094	0.53	
G1_096	Unidentified G1_096	1.45	
G1_104	Unidentified G1_104; Carbohydrate	2.18	
G1_105	Unidentified G1_105	1.65	
G1_106	L-Glycerol-3-phosphate (4TMS)	1.32	
G1_108	Unidentified G1_108	1.71	
G1_109	Ribonic acid (5TMS)	1.72	
G1_114	Unidentified G1_114; Organic acid	1.62	
G1_128	Glucose MEOX (5TMS)	2.24	
G1_130	Glucose MEOX (5TMS) [BP]	2.47	
G1_138	Unidentified G1_138	0.63	
G1_141	Gluconic acid (6TMS)	1.68	
G1_142	Galactonic acid (6TMS)	1.92	
G1_147	Unidentified G1_147	1.91	
G1_151	Inositol (6TMS)	1.83	
G1_153	Unidentified G1_153	0.34	
G1_155	Unidentified G1_155	7.69	
G1_160	Unidentified G1_160	1.85	
G1_167	Unidentified G1_167	1.41	
G1_168	Galactose-6-phosphate MEOX (TMS)	1.76	
G1_169	Glucose-6-phosphate MEOX (6TMS)	1.41	
G1_174	Unidentified G1_174	0.51	
G1_175	Unidentified G1_175; Carbohydrate	0.54	
G1_178	Unidentified G1_178		0.46
G1_183	Unidentified G1_183	1.55	
G1_191	Unidentified G1_191; Carbohydrate		0.63
G1_196	Sucrose (8TMS)	1.31	
G1_198	Unidentified G1_198	1.68	
G1_206	Unidentified G1_206; Phenolic		0.67
G1_207	Unidentified G1_207		0.51
G1_208	Unidentified G1_208	1.42	
G1_209	Unidentified G1_209	1.56	
G1_212	Unidentified G1_212	0.42	0.45
G1_213	Galactinol (9TMS)	1.52	1.68

LC/MS metabolites different between Mod and WT			
		Avg abundance rel:WT	
Peak#	Identity	P39	P717
L2_001	Coumaroyl glucoside; MW326	0.49	0.65
L2_003	Unidentified L2_001; MW348	0.41	0.67
L2_004	Unidentified L2_004	0.36	0.63
L2_005	p-Coumaryl shikimate; MW320		0.75
L2_006	Unidentified L2_006		1.37
L2_007	Unidentified L2_007		1.32
L2_009	Unidentified L2_009	0.60	
L2_010	Unidentified L2_010	0.60	0.47
L2_014	Unidentified L2_014; MW324	0.56	
L2_016	Unidentified L2_016	0.57	0.58
L2_017	Unidentified L2_017	0.61	
L2_018	Unidentified L2_018	0.50	
L2_021	Unidentified L2_021; MW434	0.56	
L2_023	Vitexin; MW432	0.48	
L2_025	Unidentified L2_025	0.33	
L2_026	Unidentified L2_026	0.34	0.59
L2_027	Unidentified L2_027	0.37	
L2_028	Unidentified L2_028; MW518	0.70	
L2_029	Unidentified L2_029; MW550	0.56	
L2_030	Unidentified L2_030; MW442	0.48	
L2_031	Unidentified L2_031	0.60	
L2_033	Unidentified L2_032; MW286	0.54	
L2_037	Unidentified L2_037; MW576	0.43	
L2_038	Unidentified L2_038; MW406	0.52	
L2_040	Unidentified L2_040; MW132	0.55	
L2_042	Unidentified L2_042	0.59	
L2_043	Phenyllactic acid; MW166	0.48	
L2_046	Unidentified L2_046		0.67
L2_048	Unidentified L2_048; MW584	0.63	0.60
L2_049	Unidentified L2_049; MW466	0.54	
L2_050	Unidentified L2_050; MW506	0.39	

Table 5.6. List of "common" differential metabolites in the comparisons between P39 C3′H-RNAi and wild-type background and P717 C3′H-RNAi and wild-type background, based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Average abundance of metabolites in modified lines are expressed relative to respective wild-type background.

Transgenic Construct: C3'H-RNAi			
GC/MS metabolites different between Mod and WT			
		Avg abundance rel:WT	
Peak#	Identity	P39	P717
G1_031	Succinic acid (2TMS)	0.27	0.34
G1_044	Unidentified G1_044	3.40	4.59
G1_047	Unidentified G1_047	0.47	0.64
G1_053	Malic acid (3TMS)	0.54	0.60
G1_054	Unidentified G1_054	0.24	0.30
G1_059	Pyroglutamic acid (2TMS)	0.61	0.67
G1_060	4-Aminobutyric acid (3TMS)	1.44	0.72
G1_063	Unidentified G1_063	0.49	0.56
G1_074	Unidentified G1_074; Organic acid	0.47	0.55
G1_096	Unidentified G1_096	0.39	0.59
G1_102	Unidentified G1_101; Sugar alcohol	0.26	0.35
G1_106	L-Glycerol-3-phosphate (4TMS)	0.64	0.67
G1_109	Ribonic acid (5TMS)	0.41	0.40
G1_110	Unidentified G1_108; Organic acid	0.54	0.46
G1_114	Unidentified G1_114; Organic acid	0.46	0.49
G1_132	Galactitol (6TMS)	0.72	0.57
G1_133	Unidentified G1_133; Organic acid	0.50	0.31
G1_141	Gluconic acid (6TMS)	0.06	0.24
G1_144	Glucaric acid (6TMS)	0.15	0.17
G1_145	Galactaric acid (6TMS)	0.20	0.18
G1_147	Unidentified G1_147	0.37	0.33
G1_148	Unidentified G1_148; Organic acid	0.17	0.21
G1_149	Unidentified G1_149; Organic acid	0.15	0.15
G1_150	Unidentified G1_150; Organic acid	0.37	0.46
G1_152	3-Deoxy-arabino-hexaric acid (5TMS)	0.35	0.38
G1_167	Unidentified G1_167	0.65	0.52
G1_169	Glucose-6-phosphate MEOX (6TMS)	0.61	0.52
G1_170	Glucose-6-phosphate MEOX (6TMS) 2nd Pk	0.52	0.54
G1_174	Unidentified G1_174	0.21	0.23
G1_175	Unidentified G1_175; Carbohydrate	0.19	0.22
G1_183	Unidentified G1_183	0.14	0.16
G1_197	Unidentified G1_197; Glycoside	10.50	4.42
G1_205	Unidentified G1_205; Carbohydrate	26.14	3.15
G1_209	Unidentified G1_209	0.56	0.74
G1_212	Unidentified G1_212	0.39	0.20

LC/MS metabolites different between Mod and WT			
		Avg abundance rel:WT	
Peak#	Identity	P39	P717
L2_001	Coumaroyl glucoside; MW326	24.48	12.40
L2_003	Unidentified L2_001; MW348	0.68	0.32
L2_004	Unidentified L2_004	0.41	0.28
L2_005	p-Coumaryl shikimate; MW320	14.63	11.39
L2_006	Unidentified L2_006	0.25	0.18
L2_007	Unidentified L2_007	0.36	0.30
L2_008	Unidentified L2_008	0.61	0.57
L2_009	Unidentified L2_009	0.25	0.36
L2_014	Unidentified L2_014; MW324	0.35	0.40
L2_018	Unidentified L2_018	0.50	0.31
L2_024	Unidentified L2_024	0.51	0.24
L2_029	Unidentified L2_029; MW550	0.67	0.57
L2_030	Unidentified L2_030; MW442	0.55	0.66
L2_035	Grandidentatin; MW424	20.45	69.00
L2_036	3-ferulolquinic acid; MW368	0.23	0.54
L2_041	Salicortin; MW424	4.32	1.73
L2_043	Phenyllactic acid; MW166	0.65	0.47
L2_047	Unidentified L2_047; MW264	0.37	0.51
L2_048	Unidentified L2_048; MW584	0.03	0.03

5.6 References

Achnine, L., Blancaflor, E.B., Rasmussen, S., & Dixon, R.A. (2004). Colocalization of L-phenylalanine ammonia-lyase and cinnamate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. *Plant Cell* 16**,** 3098-3109.

Andersson-Gunneras, S., Mellerowicz, E.J., Love, J.*, et al.* (2006). Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J.* 45**,** 144-165.

Anterola, A.M., van Rensburg, H., van Heerden, P.S., Davin, L.B., & Lewis, N.G. (1999). Multi-site modulation of flux during monolignol formation in loblolly pine (Pinus taeda). *Biochem. Biophys. Res. Commun.* 261**,** 652-657.

Coleman, H.D., Park, J.Y., Nair, R., Chapple, C., & Mansfield, S.D. (2008a). RNAi-mediated suppression of p-coumaroyl-CoA 3'-hydroxylase in hybrid poplar impacts lignin deposition and soluble secondary metabolism. *Proc. Natl. Acad. Sci. U. S. A.* 105**,** 4501-4506.

Coleman, H.D., Samuels, A.L., Guy, R.D., & Mansfield, S.D. (2008b). Perturbed lignification impacts tree growth in hybrid poplar- A function of sink strength, vascular integrity, and photosynthetic assimilation. *Plant Physiol.* 148**,** 1229-1237.

Dixon, R.A., Chen, F., Guo, D., & Parvathi, K. (2001). The biosynthesis of monolignols: A "metabolic grid", or independent pathways to guaiacyl and syringyl units? *Phytochemistry* 57**,** 1069-1084.

Donaldson, L.A. (2001). Lignification and lignin topochemistry: An ultrastructural view. *Phytochemistry* 57**,** 859-873.

Franke, R., McMichael, C.M., Meyer, K., Shirley, A.M., Cusumano, J.C., & Chapple, C. (2000). Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *Plant J.* 22**,** 223-234.

Hahlbrock, K. & Scheel, D. (1989). Physiology and molecular-biology of phenylpropanoid metabolism. *Annu. Rev. Plant Phys.* 40**,** 347-369.

Hrazdina, G. & Jensen, R.A. (1992). Spatial organisation of enzymes in plant metabolic pathways. *Annu. Rev. Plant Phys.* 43**,** 241-267.

Humphreys, J.M. & Chapple, C. (2002). Rewriting the lignin roadmap. *Curr. Opin. Plant Biol.* 5**,** 224-229.

Huntley, S.K., Ellis, D., Gilbert, M., Chapple, C., & Mansfield, S.D. (2003). Significant increases in pulping efficiency in C4H-F5H-transformed poplars: Improved

chemical savings and reduced environmental toxins. *J. Agric. Food Chem.* 51**,** 6178-6183.

Kopka, J., Schauer, N., Krueger, S.*, et al.* (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21**,** 1635-1638.

Le Gall, G., Colquhoun, I.J., Davis, A.L., Collins, G.J., & Verhoeyen, M.E. (2003). Metabolite profiling of tomato (Lycopersicon esculentum) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *J. Agric. Food Chem.* 51**,** 2447-2456.

McCown, B.H. & Lloyd, G. (1981). Woody plant medium (WPM) - a mineral nutrient formulation for microculture of woody plant-species. *HortScience* 16**,** 453.

Meyer, R.C., Steinfath, M., Lisec, J.*, et al.* (2007). The metabolic signature related to high plant growth rate in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 104**,** 4759-4764.

Morris, C.R., Scott, J.T., Chang, H.-M., Sederoff, R.R., O'Malley, D., & Kadla, J.F. (2004). Metabolic profiling: A new tool in the study of wood formation. *J. Agric. Food Chem.* 52**,** 1427-1434.

Rasmussen, S. & Dixon, R.A. (1999). Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. *Plant Cell* 11**,** 1537-1551.

Robinson, A.R., Gheneim, R., Kozak, R.A., Ellis, D.D., & Mansfield, S.D. (2005). The potential of metabolite profiling as a selection tool for genotype discrimination in Populus. *J. Exp. Bot.* 56**,** 2807-2819.

Robinson, A.R., Ukrainetz, N.K., Kang, K.Y., & Mansfield, S.D. (2007). Metabolite profiling of Douglas-fir (Pseudotsuga menziesii) field trials reveals strong environmental and weak genetic variation. *New Phytol.* 174**,** 762-773.

Roessner, U., Luedemann, A., Brust, D.*, et al.* (2001a). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13**,** 11-29.

Roessner, U., Willmitzer, L., & Fernie, A.R. (2001b). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127**,** 749-764.

Rohde, A., Morreel, K., Ralph, J.*, et al.* (2004). Molecular phenotyping of the pal1 and pal2 mutants of Arabidopsis thaliana reveals far-reaching consequences on and carbohydrate metabolism. *Plant Cell* 16**,** 2749-2771.

Rolando, C., Monties, B., & LaPierre, C. (1992). Thioacidolysis. in Lin, S. & Dence, C. (Eds), Methods in lignin chemistry. Springer-Verlag, Berlin, pp. 334-349.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78**,** 779-787.

Srere, P.A. (1987). Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.* 56**,** 89-124.

Srere, P.A. (2000). Macromolecular interactions: tracing the roots. *Trends Biochem. Sci.* 25**,** 150-153.

Winkel-Shirley, B. (1999). Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiol. Plantarum.* 107**,** 142-149.

CHAPTER 6

Summary and future research

6.1 Thesis summation

A great deal of research concerning the nature of plant biochemistry and metabolism was conducted prior to the advent of modern metabolomics. Although the findings of this classic research comprise the foundation of, and continue to assist in our understanding of plant metabolism, traditional techniques were generally only capable of addressing specific aspects of metabolism in a very focused manner. Over the course of the last decade, the greater goal of plant functional genomics, and more specifically metabolomics, has been to expand the "window" through which metabolism may be viewed. Consequently, the interrelations within and between entire metabolic processes may now be characterised in a collective fashion. To this end, this body of work represents efforts to perform broadscale, non-targeted metabolomics analyses on industrially relevant and model system tree species, with a specific focus on the relationships between metabolite profiles and physico-chemical wood traits.

With regard to tree development and wood quality in an industrial context, both Douglas-fir and radiata pine were targets for metabolomics analyses. In Douglas-fir, metabolomics was assessed for its capacity to discern biological variation among full-sib families in a tree breeding population. The differential accumulation of metabolites in profiles derived from developing xylem was examined through a series of statistical analyses that incorporated family, site, tree growth and quantitative phenotypic wood traits (wood density, microfibril angle, wood chemistry and fibre morphology). Analyses revealed that metabolic and phenotypic traits alike were strongly related to site, while similar associations relating to genetic (family) structure were weak in comparison. Furthermore, correlations between specific phenotypic traits (*i.e.* tree growth, fibre morphology and wood chemistry) and metabolic traits (*i.e.* carbohydrate and lignin biosynthetic metabolites) were identified, demonstrating a coherent relationship between genetics, metabolism, environmental and phenotypic expression in wood-forming tissue of this species.

In juvenile radiata pine, metabolomics was used to investigate the relationship between the metabolism of developing xylem and the propensity for tree families to exhibit an intra-ring internal checking wood defect, which devalues lumber products. Based on either complete metabolite profiles, or reduced profiles consisting only of metabolites whose abundance was strongly correlated with the trait, it was possible to

differentiate between siblings from families having different levels of internal checking severity. Furthermore, it was possible to model the relationship between metabolite profiles and internal checking such that the severity of the defect in individual trees could be predicted accurately on the basis of profile data alone.

Investigation of the relationships between metabolism, genotype and phenotype was also conducted in a controlled, model system setting involving hybrid poplar genotypes transformed with transgenic constructs related to lignin biosynthesis, and which affected growth and physico-chemical wood traits. The initial study demonstrated that the expression of the C4H::F5H transgenic construct in *Populus tremula* × *alba*, which leads to an increased ratio of syringyl to guaiacyl lignin monomers in xylem tissue, also resulted in detectable shifts in metabolite profiles from developing xylem or non-lignifying suspension tissue cultures. Transformants were not only distinguished from the wild-type in lignin-related metabolism, but also, predominantly, in other metabolite classes such as the carbohydrates.

The comprehensive follow-up to this research assessed the consistency of modified physical (*i.e.* wood properties) and developing xylem metabolic phenotypes generated via separate expression of two genetic constructs (C4H::F5H and C3′H-RNAi) in distinct hybrid poplar genetic backgrounds (*Populus tremula* × *alba* and *Populus grandidentata* × *alba*). This work demonstrated that transgene-induced phenotypic gradients in physico-chemical wood traits can be associated with similar gradients in the global metabolism of secondary xylem biosynthesis. Furthermore, it was apparent that while distinct, at a global level the wood-forming metabolisms of different poplar hybrids can, to some extent, respond similarly to the influences of genetic manipulation. These findings have significant, positive implications for the potential development of broadly applicable metabolic markers for wood traits.

In 2002, at the time when this research was begun, plant metabolomics as a field had only recently been conceived and put to effect (Fiehn *et al.*, 2000; Roessner *et al.*, 2000). The intervening years have witnessed the rise of this new branch of functional biology, with a rapidly growing body of literature (Dettmer *et al.*, 2007), broadening applications, and considerable technical advances - particularly in the quality of software tools available for data handling and statistical analysis (Smith *et al.*, 2006; Tautenhahn *et al.*, 2008; Thimm *et al.*, 2004). The progression of this research

concurrently with the early growth of plant metabolomics concepts and technology has meant that the experiments in this body of work frequently employed state-of-the-art approaches, and now comprise a large fraction of the broad-scale, non-targeted metabolomics research conducted on tree species to date. This work has fulfilled the initial postulation that, in several scenarios, phenotypic wood traits would correlate with the non-targeted metabolite profiles of developing xylem. In doing so, it has revealed that a specific wood trait, which arises from the action of heritable genetic or environmental factors, or the effects of gene misregulation, can have a complex metabolic basis involving broad aspects of cellular metabolism; however, resource availability and the conceptual and technical limitations of contemporary metabolomics methodology have constrained these analyses. As is evident from this research and from the literature to date, the derivation of concrete and detailed biological understanding from broad-scale metabolic profile data, as well as the extension of phenotype-distinguishing correlative relationships between profiles and phenotype into practical and robust diagnostic tools, largely remain as challenges for the field of plant metabolomics to tackle in earnest.

6.2 Future research

Throughout this document it has been demonstrated that correlative relationships exist between particular wood traits and specific elements in metabolite profiles of developing xylem, and contended that such relationships could have utility in screening applications concerned with such traits. From the applied perspective, the next phase of this work should therefore involve intensive validation of this claim of utility. In particular, the carefully considered (re)construction of predictive models, based on larger sample sets representing a broad range and even distribution in the severity of the trait of interest, is paramount. The subsequent extension of model testing beyond cross-validation scenarios, to include testing against new and diverse sample sets, and with model refinement on that basis, will also be required. Such extensions will constitute essential steps in the realisation of metabolomics' utility in tree breeding and assessment applications.

From the perspective of furthering the understanding of tree biology, all of the studies described in this thesis could be repeated or extended under refined conditions.

Given the availability or resources, deeper insight could be gained from broader analytical scope and additional means of data presentation. This research has primarily focused on the inter-relationships between metabolite profiles and phenotypic traits, with some consideration of genetics, gene expression and environment; however, because metabolic data is most informative when viewed in conjunction with other measurements, the value of metabolomics analyses may be increased when additional 'omics'-scale analyses are conducted in parallel. As such, increasing the dimensionality of these metabolomics-based studies by performing concurrent genetics (*i.e.* genomic sequence data) or gene expression (*i.e.* micro-array data) analyses could lead to increased insight into the biological system(s) under inspection. Such multi-omics studies have begun to appear in the literature, and are set to become a fixture of tree functional biology (Dauwe *et al.*, 2007; Leple *et al.*, 2007). The insight provided by metabolomics analyses is also limited by the resolution of metabolite detection, and the ability to identify those metabolites resolved. As such, future efforts might consider alternative or additional sample extraction procedures, different classes of analytical instrumentation (such as MALDI or FT-MS techniques), and the expansion and improvement of standard compound libraries. Finally, with the powerful combination of multi-omics analyses coupled with a high level of metabolite resolution and identification, the importance of orderly presentation of the increasingly complex data/results is undeniable. An excellent mode of presentation is metabolic pathway scaffolding, in which genetic, gene expression, and metabolomic data are superimposed on established pathway diagrams. The presentation of comprehensive data in this manner can bring about considerable improvements in data interpretability, for both researchers and readers alike.

6.3 References

Dauwe, R., Morreel, K., Goeminne, G.*, et al.* (2007). Molecular phenotyping of lignin-modified tobacco reveals associated changes in cell-wall metabolism, primary metabolism, stress metabolism and photorespiration. *Plant J.* 52**,** 263-285.

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

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18**,** 1157-1161.

Leple, J.C., Dauwe, R., Morreel, K.*, et al.* (2007). Downregulation of cinnamoyl-coenzyme a reductase in poplar: Multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* 19**,** 3669-3691.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N., & Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23**,** 131-142.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78**,** 779-787.

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9**,** doi:10.1186/1471-2105-1189-1504.

Thimm, O., Blasing, O., Gibon, Y.*, et al.* (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37**,** 914-939.

APPENDIX A

Appendix for Chapter 2

Appendix A.1. Broad-sense heritabilities and identities of all significant metabolites

Metabolite Information[#]			Heritability[+]	Mass Spectra of Unknowns[!]
Peak#	Compound ID	Class	H^2 0.05	m/z of ten largest peaks (abundance relative to base peak)
20	Acetic Acid		0.00	
54	Acetic Acid, bisoxyl		0.00	
56	Unknown		0.00	274(20)228(92)184(67)149(21)147(100)136(21)134(40)110(77)77(27)73(82)
60	Phosphoric acid		0.00	
92	Alanine, B-		0.00	
117	Erythronic acid		0.00	
141	Ribose		0.00	
148	Unknown	Benzene Structure	0.00	369(31)341(36)295(51)281(64)222(30)221(88)209(32)207(100)147(67)73(71)
159	Unknown	Carbohydrate	0.00	333(40)331(27)305(39)292(87)218(30)217(100)189(31)147(83)143(26)73(72)
169	Pinitol		0.00	
173	Quinic acid		0.00	
182	Glucose {BP}		0.00	
202	Unknown	Sugar Acid	0.00	361(25)334(27)333(100)305(35)292(23)243(26)217(25)191(23)147(33)73(55)
232	Unknown	Dimeric Sugar	0.00	437(15)363(18)362(34)361(100)271(18)243(13)217(42)204(14)169(27)147(15)
237	Phenolic	Phenolic	0.00	429(41)355(53)341(19)295(25)282(24)281(100)221(68)207(36)147(47)73(67)
241	Unknown	Phenolic Glycoside	0.00	362(30)361(100)271(30)243(27)235(18)217(37)169(30)147(32)129(26)73(56)
250	Unknown	Unknown	0.00	367(32)361(100)313(20)312(64)271(23)243(26)217(20)169(37)147(24)73(63)
156	Unknown	Carbohydrate	0.00	Gölm Metabolite Database: EITTMS_N12C_ATHR_1770.9_1135EC25_
147	Unknown	Unknown	0.01	306(21)286(47)245(24)244(83)217(21)163(48)147(59)142(56)129(24)73(100)
239	Unknown	Phenolic Dimer	0.02	429(32)356(31)355(80)341(22)282(22)281(100)221(56)207(34)147(44)73(68)
186	Unknown	Sugar Acid	0.03	Gölm Metabolite Database: EITTMS_N12C_ATHR_1871.9_1135EC44_
230	Sucrose		0.03	
82	Unknown	Unknown	0.03	289(12)247(11)217(17)149(33)148(14)147(100)127(54)116(18)75(12)73(68)
139	Unknown	Amino Acid	0.05	279(31)246(66)232(26)218(42)174(40)159(100)149(24)147(92)100(36)73(67)
118	Unknown	Amino Acid	0.05	332(20)242(27)230(29)219(32)218(100)174(74)147(71)100(25)86(29)73(93)
221	Fructose 6P		0.05	
175	Fructose		0.06	
214	Unknown	Carbohydrate	0.07	217(72)207(21)205(25)204(100)191(22)189(27)149(26)147(73)129(22)73(72)
233	Unknown	Dimeric Sugar	0.08	399(71)361(100)243(38)237(49)217(45)203(80)169(84)147(84)129(44)73(86)
218	Unknown	Sugar Phosphate	0.08	285(25)284(89)272(84)228(23)217(51)194(20)149(33)148(21)147(42)73(100)
120	Threonic acid		0.09	
91	Unknown	Amino Acid	0.09	218(61)174(42)160(30)149(25)148(15)147(82)130(15)116(16)73(100)10(26)
22	Unknown	Unknown	0.09	366(4)205(5)204(26)150(3)149(22)148(17)147(100)132(3)131(7)73(13)
187	Unknown	Benzene Structure	0.09	429(13)415(13)341(22)283(18)282(25)281(100)221(22)207(16)147(37)73(59)
81	Unknown	Amino Acid	0.10	302(12)290(13)289(30)288(100)148(25)172(30)148(13)147(34)100(33)73(84)
30	Unknown	Benzene Structure	0.10	357(12)356(18)355(54)323(5)285(7)269(27)268(28)267(100)251(8)73(50)
73	Glyceric acid		0.10	
177	Fructose {BP}		0.11	
104	Malic acid		0.11	
244	Coniferin		0.13	
209	Inositol		0.13	
246	Unknown	Phenolic / Glucoside	0.13	450(37)362(33)361(95)297(50)271(33)243(42)217(100)169(45)147(45)73(76)
160	Ribonic acid		0.14	
85	Unknown	Amino Acid	0.15	248(9)176(9)175(16)174(100)147(15)146(7)100(17)86(26)73(37)59(11)
74	Fumaric acid		0.15	
238	Unknown	Phenolic / Sugar	0.15	423(27)362(32)361(100)297(19)271(25)243(29)217(31)169(46)147(27)73(63)
224	Unknown	Carbohydrate	0.16	435(30)434(44)433(100)362(21)361(28)360(53)318(31)217(27)147(41)73(28)
67	Maleic Acid		0.16	
229	Adenosine		0.17	
222	Glucose 6P		0.18	
183	Unknown	Carbohydrate	0.20	319(22)305(21)221(26)217(70)207(18)205(34)204(100)189(22)147(76)73(56)
115	Unknown	Small Acid	0.20	300(16)274(14)246(27)245(62)226(41)149(25)148(16)147(100)134(19)73(46)
181	Unknown	Carbohydrate	0.20	480(18)273(45)205(76)189(74)149(24)148(28)147(100)117(18)73(38)57(22)
223	Glucose 6P {BP}		0.21	
178	Glucopyranose		0.23	
149	Unknown	Amino Acid	0.24	273(14)244(27)219(16)191(15)149(24)147(100)111(45)82(32)73(85)55(21)
216	Unknown	Benzene structure	0.24	430(42)429(79)356(27)255(61)341(30)281(64)221(49)207(31)147(54)73(100)
192	Unknown	Unknown	0.24	292(29)291(75)221(52)217(57)149(34)147(100)133(36)103(30)75(29)73(65)
165	Unknown	Sugar Acid	0.30	334(37)333(100)305(30)292(41)219(23)217(37)147(48)143(36)117(23)73(99)
138	Arabinose		0.31	
135	Xylose {BP}		0.34	
211	Unknown	Carbohydrate	0.35	320(28)319(96)315(24)217(40)205(28)157(52)149(20)147(100)129(66)73(76)
137	Xylose		0.42	
226	Unknown	Phenolic	0.67	429(35)355(31)341(22)283(19)282(32)281(100)221(43)207(26)147(40)73(55)

[#] Compound identity or class determined through mass-spectral and retention time matches with standard compounds. {BP} Metabolite by-product, as suggested by the Gölm Metabolite Database.

[+] Only metabolites for which it was possible to calculate broad-sense heritabilities are presented (64 of 139).

[!] For unidentified compounds with strong hits onto 'unidentified' compounds in the Gölm Metabolite Database, the GMD reference is given.

Appendix A.2. Comparison of all significant metabolite canonical correlation coefficients with factor analysis factor scores and broad-sense heritabilities.

Metabolite Information#			CCA$	Factor analyses*			Heritability+	Mass Spectra of Unknowns!
Peak#	Compound ID	Class	Metabolite1	F-1	F-2	F-3	H^2 0.05	m/z of ten largest peaks (abundance relative to base peak)
31	Unknown	Unknown	**0.615**	0.75		-0.35		192(18)191(100)190(14)184(66)149(14)148(13)147(82)134(40)77(22)73(66)
82	Unknown	Unknown	**0.611**	0.82			0.03	289(12)247(11)217(17)149(33)148(14)147(100)127(54)116(18)75(12)73(68)
118	Unknown	Amino Acid	**0.548**	0.86			0.05	332(20)242(27)230(29)219(32)218(100)174(74)147(71)100(25)86(29)73(93)
91	Unknown	Amino Acid	**0.545**	0.84			0.09	218(61)174(42)160(30)149(25)148(15)147(82)130(15)116(16)73(100)10(26)
92	Alanine, B-		**0.509**	0.71			0.00	
140	Unknown	Carbohydrate	**0.500**	0.64				334(20)320(45)304(22)230(32)191(21)163(19)149(33)48(20)147(100)73(86)
108	Unknown	Unknown	**0.449**	0.38	0.61			293(100)253(57)252(67)251(59)237(76)221(50)191(86)175(63)147(90)73(97)
13	Unknown	Benzene Structure	**0.434**		0.56	-0.44		296(5)295(13)247(3)225(7)210(5)209(32)208(22)207(100)191(8)73(9)
33	Unknown	Unknown	**0.424**	0.56		-0.33		355(6)186(8)185(54)170(10)167(21)153(15)152(100)134(15)86(64)59(42)
62	Unknown	Small acid	**0.422**	0.52				314(9)301(10)300(20)299(100)284(2)283(7)227(3)225(10)211(4)73(3)
214	Unknown	Carbohydrate	**0.410**	0.68			0.07	217(72)207(21)205(25)204(100)191(22)189(27)149(26)147(73)129(22)73(72)
178	Glucopyranose		**0.397**		0.65		0.23	
96	Unknown	Hydrocarbon Chain	**0.395**	0.60	0.38			314(27)301(22)245(100)193(17)191(64)147(51)116(22)110(18)77(17)73(64)
128	Unknown	Carbohydrate	**0.384**	0.74				231(37)220(48)217(87)203(29)149(33)147(100)133(21)130(24)129(72)73(79)
54	Acetic acid, bis-ox		**0.372**	0.65			0.00	
241	Unknown	Phenolic Glycoside	**0.371**	0.62			0.00	362(30)361(100)271(30)243(27)235(18)217(37)169(30)147(32)129(26)73(56)
138	Arabinose		**0.363**	0.46	0.41		0.31	
200	Unknown	Phenolic	**0.358**	0.44				369(24)295(60)282(23)281(83)222(25)221(100)207(73)149(20)147(44)73(46)
24	Ammonium		**0.353**	0.41	-0.35	-0.43		
120	Threonic acid		**0.349**	0.49			0.09	
182	Glucose {BP}		**0.346**				0.00	
246	Unknown	Phenolic/Glucoside	**0.338**	0.58			0.13	450(37)362(33)361(95)297(50)271(33)243(42)217(100)169(45)147(45)73(76)
135	Xylose {BP}		**0.337**	0.51			0.34	
14	Unknown	Small Acid	**0.334**	0.37	-0.34	-0.55		218(15)217(14)203(9)163(22)149(43)148(19)147(100)133(29)131(9)73(33)
224	Unknown	Carbohydrate	**0.334**	0.50		0.30	0.16	435(30)434(44)433(100)362(21)361(28)360(53)318(31)217(27)147(41)73(28)
111	Pyroglutamic acid		**0.333**	0.54				
20	Acetic Acid		**0.323**	0.59			0.00	
68	Unknown	Unknown	**0.318**			-0.49		256(66)248(47)206(45)186(50)174(100)164(38)120(54)84(38)77(56)73(66)
175	Fructose		**0.310**				0.06	
177	Fructose {BP}		**0.308**				0.11	
75	Unknown	Benzene Structure	**0.305**		0.73			355(34)281(56)269(14)268(19)267(65)223(19)222(28)221(100)147(42)73(40)
235	Maltose		**0.302**	0.43				
179	Glucose		**0.300**					
173	Quinic acid		**-0.304**	-0.31	-0.49		0.00	
150	Rhamnose		**-0.308**			0.31		
115	Unknown	Small Acid	**-0.311**			0.41	0.20	300(16)274(14)246(27)245(62)226(41)149(25)148(16)147(100)134(19)73(46)
81	Unknown	Amino Acid	**-0.316**	-0.37			0.10	302(12)290(13)289(30)288(100)148(25)172(30)148(13)147(34)100(33)73(84)
30	Unknown	Benzene Structure	**-0.331**				0.10	357(12)356(18)355(54)323(5)285(7)269(27)268(28)267(100)251(8)73(50)
190	Unknown	Sugar Alcohol	**-0.359**					273(14)244(27)219(16)191(15)149(24)147(100)111(45)82(32)73(52)55(21)
74	Fumaric acid		**-0.373**			0.35	0.15	
133	Unknown	Benzene Structure	**-0.379**	-0.39		0.45		402(21)401(51)357(22)256(33)355(100)327(21)281(34)267(35)221(36)73(70)
70	Unknown	Benzene Structure	**-0.379**	-0.55				430(11)429(23)343(19)342(18)341(85)327(12)326(15)325(47)147(34)73(100)
215	Unknown	Sugar Phosphate	**-0.405**			0.68		343(56)342(79)341(87)315(90)299(100)243(85)227(74)211(97)75(67)73(81)
85	Unknown	Amino Acid	**-0.444**	-0.41	-0.34		0.15	248(9)176(9)175(16)174(100)147(15)146(7)100(17)86(26)73(37)59(11)
217	Unknown	Carbohydrate	**-0.450**			0.85		Gölm Metabolite Database: EITTMS_N12C_STUR_2277.7_1135EC29_
205	Unknown	Sugar Acid	**-0.451**	-0.50		0.41		220(17)219(26)217(34)205(23)204(100)189(18)157(17)147(34)129(17)73(91)
210	Unknown	Sugar Alcohol	**-0.458**			0.43		435(15)434(30)433(63)344(20)343(67)318(26)204(25)191(100)147(62)73(86)
244	Coniferin		**-0.459**			0.66	0.13	
164	Shikimic acid		**-0.487**	-0.45	0.39	0.56		
73	Glyceric acid		**-0.546**	-0.52			0.10	
169	Pinitol		**-0.659**	-0.70		0.32	0.00	

Compound identity or class determined through mass-spectral and retention time matches with standard compounds. {BP} Metabolite by-product, as suggested by the Gölm Metabolite Database.

$ Only metabolites with canonical correlation coefficients >+/- 0.3 (*i.e.* significant), across all 139 metabolites are presented.

* Of metabolites with significant canonical correlation coefficients, only significant factor scores (>+/- 0.3) in the site-differentiating factors are presented.

+ Of metabolites with significant canonical correlation coefficients, it was only possible to calculate broad-sense heritabilities for some.

! For unidentified compounds with strong hits onto 'unidentified' compounds in the Gölm Metabolite Database, the GMD reference is given.

APPENDIX B

Appendix for Chapter 5

Appendix B.1. Entire list of metabolites identified in GC/MS profiles. RT = retention time, RI = retention index, BP indicates by-product, as suggested by the Gölm metabolite database.

List of all 221 metabolites resolved from GC/MS chromatograms - Page 1 of 4

Peak#	RTime	RI	Identity. Where unknown: ten (if possible) most abundant masses (mz: mass relative-abundance basepeak999|)
G1_001	6.31	1026.3	Unidentified G1_001; mz: 152 999 | 166 257 | 153 128 | 167 78 | 122 65 | 78 44 | 97 30 | 83 25 | 136 22 | 96 11 |
G1_002	6.40	1032.0	Unidentified G1_002; mz: 93 999 | 123 619 | 95 385 | 73 231 | 125 212 | 55 107 | 65 97 | 94 79 | 75 75 | 165 72 |
G1_003	6.60	1043.6	Pyruvic acid (1MEOX) (1TMS)
G1_004	6.77	1054.3	Lactic acid (2TMS)
G1_005	7.04	1070.3	Glycolic acid (2TMS)
G1_006	7.54	1100.4	L-Alanine (2TMS)
G1_007	8.23	1142.4	Oxalic acid (2TMS)
G1_008	8.23	1142.5	Unidentified G1_007 org acid; mz: 147 999 | 133 425 | 149 324 | 148 215 | 190 192 | 131 166 | 160 141 | 59 123 | 205 114 | 162 99 |
G1_009	8.26	1144.4	3-Hydroxypropanoic acid (2TMS)
G1_010	8.29	1146.0	2-Pyrrolidinone (1TMS)
G1_011	8.83	1178.4	Unidentified G1_011; mz: 281 999 | 282 272 | 147 239 | 283 186 | 265 166 | 369 148 | 249 103 | 207 58 | 370 55 | 284 51 |
G1_012	8.86	1180.1	Unidentified G1_012; mz: 209 999 | 193 784 | 210 216 | 211 157 | 194 149 | 97 16 | 65 3 |
G1_013	8.93	1184.6	Monomethylphosphate (2TMS)
G1_014	8.95	1185.6	Unidentified G1_014; mz: 174 999 | 190 647 | 218 605 | 156 573 | 86 446 | 59 298 |
G1_015	9.50	1219.2	Unidentified G1_015; mz: 228 999 | 110 665 | 73 618 | 184 546 | 134 457 | 77 428 | 69 158 | 136 129 | 75 122 | 229 114 |
G1_016	9.57	1223.3	L-Valine (2TMS)
G1_017	9.92	1244.3	Unidentified G1_017; mz: 73 999 | 147 831 | 117 719 | 234 693 | 130 493 | 131 349 | 102 300 | 89 205 | 149 190 | 59 190 |
G1_018	10.03	1251.3	Urea (2TMS)
G1_019	10.08	1254.2	Benzoic acid (1TMS)
G1_020	10.12	1256.9	Unidentified G1_020; mz: 110 999 |
G1_021	10.29	1266.9	L-Serine (2TMS)
G1_022	10.39	1273.1	Unidentified G1_022; mz: 84 999 | 56 378 | 186 245 |
G1_023	10.44	1276.0	Unidentified G1_023; mz: 147 999 | 175 972 | 131 393 | 172 177 | 79 166 | 102 116 | 60 111 | 177 99 | 132 97 | 103 93 |
G1_024	10.44	1276.2	Ethanolamine (3TMS)
G1_025	10.55	1282.6	L-Leucine (2TMS)
G1_026	10.62	1286.8	Phosphoric acid (3TMS)
G1_027	10.63	1287.1	Unidentified G1_027; mz: 133 999 | 208 851 | 191 769 | 192 413 | 148 220 | 77 107 | 178 90 | 386 78 | 230 63 | 72 53 |
G1_028	10.93	1305.7	L-Isoleucine (2TMS)
G1_029	10.97	1306.6	L-Threonine (2TMS)
G1_030	11.15	1318.5	Glycine (3TMS)
G1_031	11.22	1323.0	Succinic acid (2TMS)
G1_032	11.28	1326.6	Unidentified G1_032; mz: 341 999 | 325 652 | 429 471 | 343 242 | 430 210 |
G1_033	11.30	1326.6	Unidentified G1_033; mz: 240 999 | 241 159 |
G1_034	11.33	1329.6	Unidentified G1_034; mz: 73 999 | 254 452 | 239 294 | 151 133 | 255 94 | 74 86 | 166 73 | 136 69 | 256 38 |
G1_035	11.60	1346.0	Glyceric acid (3TMS)
G1_036	11.62	1347.4	Unidentified G1_036; mz: 184 999 | 285 444 | 77 254 | 174 164 | 185 101 |
G1_037	11.71	1352.9	Unidentified G1_037; mz: 200 999 | 147 544 | 154 489 | 112 265 | 243 247 | 228 206 | 201 206 | 172 153 | 59 115 | 255 98 |
G1_038	11.75	1355.2	Fumaric acid (2TMS)
G1_039	12.00	1370.2	Unidentified G1_039; mz: 130 999 | 316 815 | 88 587 | 226 272 |
G1_040	12.04	1372.5	L-Alanine (3TMS)
G1_041	12.09	1375.6	L-Serine (3TMS)
G1_042	12.39	1393.8	Cytosine (2TMS)
G1_043	12.54	1402.9	L-Threonine (3TMS)
G1_044	12.70	1412.1	Unidentified G1_044; mz: 254 999 | 239 981 | 223 247 | 255 230 | 241 228 | 240 217 | 73 101 | 133 94 | 257 93 | 147 58 |
G1_045	13.00	1430.8	3-Hydroxymyristic acid (2TMS)
G1_046	13.05	1434.0	Unidentified G1_044; mz: 73 999 | 147 737 | 160 710 | 116 478 | 130 410 | 75 300 | 234 163 | 161 146 | 74 146 | 117 131 |
G1_047	13.12	1438.0	Unidentified G1_047; mz: 248 999 | 249 268 | 290 197 | 134 58 |
G1_048	13.26	1446.3	2-Hydroxybenzyl alcohol (2TMS)
G1_049	12.54	1463.5	Unidentified G1_049; mz: 73 999 | 117 991 | 232 830 | 233 174 | 244 164 | 147 121 | 116 117 | 118 110 | 259 107 | 74 106 |
G1_050	13.65	1470.0	Unidentified G1_050; mz: 135 999 | 134 323 | 209 238 | 165 172 | 77 134 | 179 128 | 105 111 | 91 55 | 194 54 | 79 42 |
G1_051	13.91	1485.6	Pyruvic acid oxime (3TMS)
G1_052	14.04	1493.8	Unidentified G1_052; mz: 327 999 | 282 540 | 415 341 | 283 326 | 399 165 |
G1_053	14.17	1502.2	Malic acid (3TMS)
G1_054	14.20	1503.5	Unidentified G1_054; mz: 423 999 | 424 361 | 497 290 | 425 200 | 498 128 | 335 123 | 499 62 | 336 49 | 333 46 | 426 44 |
G1_055	14.31	1511.6	L-Asparagine (2TMS)
G1_056	14.37	1515.9	Unidentified G1_054; mz: 230 999 | 142 320 | 304 251 | 231 247 |
G1_057	14.59	1532.3	L-Methionine
G1_058	14.63	1535.2	L-Aspartic acid (3TMS)
G1_059	14.67	1537.7	Pyroglutamic acid (2TMS)
G1_060	14.73	1542.4	4-Aminobutyric acid (3TMS)
G1_061	14.89	1554.0	4-Methyl-5-hydroxy-3-penten-2-one (1TMS)
G1_062	15.07	1565.4	L-Norvaline (3TMS)
G1_063	15.13	1569.7	Unidentified G1_063; mz: 258 999 | 348 278 | 274 256 | 259 254 | 163 149 | 100 104 | 349 99 | 260 85 | 59 77 |
G1_064	15.20	1576.6	Unidentified G1_064; mz: 219 999 | 129 516 | 117 254 | 218 151 | 203 121 |
G1_065	15.32	1585.6	Threonic acid (4TMS)
G1_066	15.37	1588.8	Unidentified G1_066; mz: 261 999 | 162 805 | 243 768 | 100 402 | 113 376 | 141 318 | 215 201 | 207 175 | 116 154 | 91 119 |
G1_067	15.42	1592.4	Unidentified G1_067; mz: 261 999 | 162 805 | 243 768 | 100 402 | 113 376 | 141 318 | 215 201 | 207 175 | 116 154 | 91 119 |
G1_068	15.44	1594.4	Unidentified G1_067; mz: 114 999 | 73 627 | 290 384 | 100 332 | 276 319 | 115 191 | 172 171 | 188 149 | 191 145 | 291 141 |
G1_069	15.53	1600.6	L-Proline (2TMS)
G1_070	15.68	1611.2	Unidentified G1_070; mz: 218 999 | 261 798 | 162 508 | 141 273 | 262 174 | 219 164 | 56 117 | 263 94 | 232 91 | 363 85 |

List of all 221 metabolites resolved from GC/MS chromatograms - Page 2 of 4

Peak#	RTime	RI	Identity. Where unknown: ten (if possible) most abundant masses (mz: mass relative-abundance basepeak999|)
G1_071	15.72	1614.8	Glutamine (4TMS) split peak 1
G1_072	15.73	1615.2	Glutamine (4TMS) split peak 2
G1_073	15.79	1619.9	Unidentified G1_073; mz: 244 999 | 154 556 | 147 410 | 73 403 | 243 384 | 241 376 | 211 318 | 245 216 | 149 122 | 246 92 |
G1_074	15.92	1629.5	Unidentified G1_074; mz: 292 999 | 75 562 | 102 444 | 129 374 | 333 358 | 131 348 | 293 338 | 117 331 | 74 90 | 221 46 |
G1_075	15.96	1632.0	Unidentified G1_075; mz: 215 999 | 188 979 | 190 477 | 214 420 | 148 250 | 213 204 | 331 194 | 303 149 | 304 127 | 207 95 |
G1_076	15.96	1632.2	Unidentified G1_076; mz: 147 999 | 148 171 | 149 156 | 200 112 | 225 44 |
G1_077	15.97	1633.1	Unidentified G1_076; mz: 73 999 |
G1_078	16.07	1640.2	L-Glutamic acid (3TMS)
G1_079	16.12	1643.8	4-Hydroxybenzoic acid (2TMS)
G1_080	16.18	1648.3	L-Phenylalanine (2TMS)
G1_081	16.22	1652.9	Unidentified G1_081; mz: 156 999 | 200 828 | 230 588 | 302 159 | 201 130 | 202 115 |
G1_082	16.28	1655.2	Unidentified G1_082; mz: 216 999 | 188 816 | 73 644 | 231 364 | 172 216 | 218 152 | 189 151 | 213 142 | 330 97 | 190 89 |
G1_083	16.28	1655.5	Unidentified G1_083; mz: 147 999 | 217 915 | 149 249 | 204 220 | 203 209 | 148 169 | 131 120 | 130 119 | 219 118 | 133 105 |
G1_084	16.33	1659.3	Unidentified G1_084; mz: 333 999 | 143 978 | 73 565 | 147 399 | 149 170 | 189 121 |
G1_085	16.36	1664.4	Unidentified G1_085; mz: 200 999 | 315 948 | 147 851 | 73 791 | 216 455 | 142 385 | 112 355 | 172 296 | 316 259 | 149 154 |
G1_086	16.41	1668.5	Unidentified G1_086; mz: 221 999 | 295 243 | 399 102 |
G1_087	16.50	1671.5	Unidentified G1_087; mz: 355 999 | 401 396 | 356 371 | 267 247 | 327 235 | 357 229 | 403 178 | 402 177 | 358 79 | 385 67 |
G1_088	16.60	1676.7	2 4 5-Trihydroxypentanoic acid (4TMS)
G1_089	16.65	1680.0	D-Ribonic acid lactone (3TMS)
G1_090	16.86	1695.4	L-Asparagine (3TMS)
G1_091	16.88	1697.0	Unidentified G1_091; mz: 217 999 | 307 213 | 218 182 | 290 174 | 103 129 | 277 83 |
G1_092	17.05	1611.8	Unidentified G1_092; mz: 193 999 | 271 605 | 194 221 | 272 173 | 195 170 | 211 93 | 273 72 | 286 50 | 165 44 | 255 39 |
G1_093	17.09	1714.6	Unidentified G1_093; mz: 200 999 | 147 437 | 315 386 | 233 365 | 204 245 | 177 227 | 201 212 | 261 131 | 130 121 | 189 116 |
G1_094	17.15	1718.5	Unidentified G1_094; mz: 147 999 | 73 716 | 217 535 | 319 379 | 149 225 | 133 181 | 83 157 | 221 148 | 55 139 | 148 132 |
G1_095	17.21	1722.9	1H-Indole-2 3-dione 1-(tert-butyldimethylsilyl)-5-isopropyl- 3-(O-methyloxime)
G1_096	17.27	1727.2	Unidentified G1_096; mz: 231 999 | 73 770 | 147 646 | 143 402 | 220 267 | 149 166 | 232 137 | 233 125 | 229 110 | 144 84 |
G1_097	17.30	1729.8	Xylitol (5TMS)
G1_098	17.32	1730.9	Unidentified G1_098; mz: 147 999 | 73 937 | 129 749 | 217 702 | 218 232 | 75 117 | 205 115 | 130 105 | 74 104 | 159 79 |
G1_099	17.41	1737.8	2-Aminoadipic acid (3TMS)
G1_100	17.65	1752.7	Rhamnose MEOX (4TMS) [BP]
G1_101	17.79	1756.3	Unidentified G1_101; mz: 221 999 | 401 576 | 355 290 | 475 242 | 489 229 | 403 167 | 563 167 | 223 121 | 476 90 | 430 87 |
G1_102	17.94	1776.2	Unidentified G1_101; mz: 217 999 | 147 957 | 73 324 | 218 180 | 149 173 | 189 151 | 205 136 | 148 126 | 129 116 | 117 100 |
G1_103	17.98	1779.1	Unidentified G1_103; mz: 147 999 | 155 993 | 73 871 | 273 580 | 229 425 | 183 327 | 149 232 | 167 202 | 148 179 | 133 138 |
G1_104	18.10	1787.7	Unidentified G1_104; mz: 73 999 | 217 832 | 147 776 | 129 368 | 143 309 | 157 242 | 102 161 | 149 155 | 221 114 | 148 111 |
G1_105	18.10	1788.3	Unidentified G1_105; mz: 333 999 | 305 199 | 307 149 | 294 137 | 346 75 | 207 46 | 348 45 | 422 40 | 295 37 | 52 19 |
G1_106	18.16	1792.3	L-Glycerol-3-phosphate (4TMS)
G1_107	18.20	1795.5	L-Glutamine (3TMS)
G1_108	18.26	1799.8	Unidentified G1_108; mz: 293 999 | 333 634 | 148 555 | 218 505 | 331 425 | 294 424 | 133 417 | 219 373 | 205 178 | 231 177 |
G1_109	18.27	1800.1	Ribonic acid (5TMS)
G1_110	18.47	1814.8	Unidentified G1_108; mz: 293 999 | 333 634 | 148 555 | 218 505 | 331 425 | 294 424 | 133 417 | 219 373 | 205 178 | 231 177 |
G1_111	18.65	1828.2	Unidentified G1_111; mz: 147 999 | 281 935 | 73 461 | 369 335 | 282 328 | 557 242 | 370 228 | 283 226 | 200 223 | 149 214 |
G1_112	18.71	1832.1	Shikimic acid (4TMS)
G1_113	18.80	1839.1	Unidentified G1_113; mz: 420 999 | 335 740 | 231 655 | 149 561 | 128 526 | 492 490 | 291 469 | 201 449 | 421 422 | 331 268 |
G1_114	18.81	1839.7	Unidentified G1_114; mz: 333 999 | 292 383 | 334 300 | 305 190 | 293 99 | 217 90 | 306 84 | 75 70 | 171 63 | 346 58 |
G1_115	18.90	1846.2	Citric acid (4TMS)
G1_116	19.01	1854.1	Unidentified G1_116; mz: 333 999 | 73 427 | 147 385 | 292 380 | 305 317 | 334 257 | 345 175 | 335 148 | 217 116 | 130 105 |
G1_117	19.05	1857.3	Tagatose methoxyamine [BP] (5TMS)
G1_118	19.15	1864.5	Unidentified G1_118; mz: 156 999 | 318 232 | 73 229 | 147 204 | 157 117 | 230 99 | 346 72 | 128 71 | 319 65 | 302 49 |
G1_119	19.17	1866.1	Unidentified G1_119; mz: 267 999 | 345 686 | 268 295 | 346 293 | 197 288 | 135 199 | 207 192 | 269 191 | 347 189 | 57 125 |
G1_120	19.28	1874.1	Unidentified G1_120; mz: 379 999 | 73 764 | 147 761 | 247 487 | 131 445 | 157 430 | 146 307 | 219 255 | 261 215 | 380 202 |
G1_121	19.34	1878.3	Unidentified G1_121; mz: 129 999 | 147 552 | 319 521 | 306 453 | 191 326 | 190 307 | 320 139 | 305 127 | 175 123 | 207 106 |
G1_122	19.39	1881.9	Unidentified G1_122; mz: 70 999 | 302 593 | 186 437 | 303 191 | 212 160 | 158 153 | 68 84 | 103 83 | 122 76 | 219 69 |
G1_123	19.51	1890.5	Quinic acid (5TMS)
G1_124	19.68	1903.3	Fructose MEOX (5TMS)
G1_125	19.75	1909.8	Sorbose MEOX (5TMS) [BP]
G1_126	19.81	1914.5	Fructose MEOX (5TMS) [BP]
G1_127	19.85	1918.4	Mannose MEOX (5TMS)
G1_128	20.00	1931.5	Glucose MEOX (5TMS)
G1_129	20.08	1938.4	L-Lysine (4TMS)
G1_130	20.24	1951.6	Glucose MEOX (5TMS) [BP]
G1_131	20.32	1959.0	L-Tyrosine (3TMS)
G1_132	20.39	1964.5	Galactitol (6TMS)
G1_133	20.48	1972.5	Unidentified G1_133; mz: 217 999 | 73 447 | 147 404 | 307 404 | 331 260 | 191 232 | 218 218 | 308 131 | 306 131 | 103 90 |
G1_134	20.73	1990.6	Unidentified G1_134; mz: 217 999 | 73 948 | 361 664 | 147 618 | 169 370 | 129 270 | 243 269 | 189 200 | 271 193 | 362 192 |
G1_135	20.75	1995.8	Unidentified G1_135; mz: 389 999 | 183 840 | 147 410 | 189 329 | 390 322 | 257 241 | 149 236 | 267 217 | 188 205 | 299 175 |
G1_136	20.84	2003.0	Gulonic acid (6TMS)
G1_137	20.88	2006.5	Unidentified G1_135; mz: 295 999 | 310 900 | 251 463 | 177 455 | 221 274 | 311 226 | 236 224 | 296 206 | 252 97 | 191 94 |
G1_138	20.93	2011.2	Unidentified G1_138; mz: 239 999 | 415 789 | 143 206 | 209 185 | 204 170 | 157 155 | 417 147 | 241 140 | 240 131 | 83 118 |
G1_139	21.00	2016.8	Unidentified G1_139; mz: 217 999 | 147 349 | 73 254 | 218 196 | 189 104 | 219 97 | 129 76 | 149 75 | 394 70 | 307 60 |
G1_140	21.06	2022.4	Glutamine (4TMS) repeat?

List of all 221 metabolites resolved from GC/MS chromatograms - Page 3 of 4

Peak#	RTime	RI	Identity. Where unknown: ten (if possible) most abundant masses (mz: mass relative-abundance basepeak999|)
G1_141	21.16	2031.0	Gluconic acid (6TMS)
G1_142	21.22	2035.7	Galactonic acid (6TMS)
G1_143	21.41	2051.8	Palmitic acid (1TMS)
G1_144	21.43	2054.1	Glucaric acid (6TMS)
G1_145	21.44	2055.0	Galactaric acid (6TMS)
G1_146	21.63	2067.7	Unidentified G1_146; mz: 204 999 | 73 787 | 147 356 | 319 295 | 217 289 | 205 254 | 189 236 | 157 194 | 129 172 | 220 158 |
G1_147	21.66	2070.6	Unidentified G1_147; mz: 147 999 | 73 570 | 217 392 | 143 332 | 149 329 | 449 222 | 191 182 | 148 168 | 229 121 | 190 102 |
G1_148	21.68	2072.2	Unidentified G1_148; mz: 333 999 | 334 415 | 143 310 | 73 238 | 292 236 | 335 222 | 305 172 | 447 94 | 419 88 | 373 72 |
G1_149	21.79	2084.8	Unidentified G1_149; mz: 333 999 | 73 374 | 334 315 | 147 291 | 143 250 | 292 204 | 305 167 | 447 158 | 335 126 | 189 113 |
G1_150	21.98	2105.5	Unidentified G1_150; mz: 333 999 | 73 835 | 147 469 | 143 396 | 305 311 | 334 284 | 189 208 | 292 196 | 335 160 | 217 133 |
G1_151	22.33	2127.8	Inositol (6TMS)
G1_152	22.59	2149.7	3-Deoxy-arabino-hexaric acid (5TMS)
G1_153	22.58	2152.2	Unidentified G1_153; mz: 352 999 | 147 916 | 157 708 | 148 357 | 217 332 | 205 278 | 353 225 | 320 201 | 158 161 | 117 143 |
G1_154	22.61	2155.6	Unidentified G1_154; mz: 147 999 | 157 615 | 129 466 | 205 464 | 320 408 | 149 379 | 133 278 | 221 237 | 130 224 | 352 195 |
G1_155	22.72	2164.4	Unidentified G1_155; mz: 221 999 | 147 809 | 207 364 | 129 240 | 319 213 | 157 178 | 131 174 | 223 150 | 204 150 | 402 131 |
G1_156	22.79	2170.5	Unidentified G1_156; mz: 310 999 | 295 952 | 251 380 | 177 312 | 221 235 | 311 223 | 296 193 | 236 164 | 252 76 | 297 70 |
G1_157	22.87	2177.8	L-Histidine (?TMS)
G1_158	23.30	2217.0	Octadecadienoic acid (1TMS)
G1_159	23.36	2223.1	Unidentified G1_159; mz: 204 999 | 73 411 | 147 312 | 205 231 | 191 229 | 189 203 | 217 194 | 235 189 | 117 111 | 206 101 |
G1_160	23.42	2229.3	Unidentified G1_160; mz: 357 999 | 315 603 | 299 536 | 445 379 | 373 271 | 358 261 | 503 214 | 446 175 | 359 169 | 316 154 |
G1_161	23.61	2248.9	Octadecanoic acid (1TMS)
G1_162	23.64	2252.3	Unidentified G1_162; mz: 299 999 | 315 341 | 317 326 | 148 250 | 587 246 | 369 190 | 228 160 | 433 156 | 301 153 | 207 149 |
G1_163	23.88	2271.0	Unidentified G1_163; mz: 429 999 | 355 868 | 281 654 | 430 430 | 221 331 | 341 292 | 431 286 | 401 239 |
G1_164	24.20	2308.6	Unidentified G1_164; mz: 73 999 | 217 699 | 371 442 | 189 354 | 157 318 | 211 295 | 642 222 | 314 204 | 641 195 | 462 181 |
G1_165	24.30	2320.5	Unidentified G1_165; mz: 73 999 | 147 967 | 214 407 | 129 270 | 319 222 | 258 218 | 133 162 | 148 157 | 204 154 | 290 149 |
G1_166	24.45	2335.4	Unidentified G1_166; mz: 73 999 | 147 379 | 290 330 | 133 253 | 217 220 | 129 177 | 319 92 | 74 77 | 117 59 | 284 52 |
G1_167	24.76	2367.8	Unidentified G1_167; mz: 315 999 | 316 244 | 317 128 | 301 48 |
G1_168	24.76	2368.2	Galactose-6-phosphate MEOX (TMS)
G1_169	24.89	2381.7	Glucose-6-phosphate MEOX (6TMS)
G1_170	25.08	2401.2	Glucose-6-phosphate MEOX (6TMS) [BP]
G1_171	25.15	2408.6	Unidentified G1_171; mz: 309 999 | 526 616 | 471 518 | 383 355 | 294 266 | 527 244 | 472 222 | 498 198 | 528 164 | 542 157 |
G1_172	25.28	2421.9	Unidentified G1_172; mz: 343 999 | 203 381 | 211 253 | 344 248 | 95 105 | 109 88 | 147 88 | 345 78 | 81 71 | 137 68 |
G1_173	25.35	2428.2	Unidentified G1_173; mz: 217 999 | 73 952 | 147 670 | 191 640 | 259 346 | 169 282 | 129 210 | 97 202 | 189 190 | 192 159 |
G1_174	25.54	2448.5	Unidentified G1_174; mz: 204 999 | 73 348 | 205 211 | 147 145 | 217 111 | 191 91 | 206 87 | 169 45 | 218 36 | 75 33 |
G1_175	25.69	2463.9	Unidentified G1_175; mz: 204 999 | 73 978 | 169 905 | 147 460 | 217 361 | 79 312 | 129 277 | 191 250 | 205 223 | 189 212 |
G1_176	25.78	2473.2	Unidentified G1_176; mz: 324 999 | 204 704 | 217 614 | 73 470 | 299 312 | 205 239 | 129 234 | 243 221 | 455 219 | 513 198 |
G1_177	25.84	2479.5	Myo-Inositol-2-phosphate (7TMS)
G1_178	25.89	2484.8	Unidentified G1_178; mz: 376 999 | 286 747 | 377 305 | 556 209 | 261 160 |
G1_179	25.98	2494.4	Unidentified G1_179; mz: 361 999 | 169 415 | 73 399 | 271 241 | 362 162 | 363 138 | 155 91 | 255 86 | 245 86 | 272 81 |
G1_180	26.26	2522.8	Unidentified G1_180; mz: 204 999 | 73 693 | 147 520 | 219 455 | 218 358 | 143 287 | 245 267 | 217 267 | 189 225 | 75 211 |
G1_181	26.32	2529.4	Unidentified G1_181; mz: 243 999 | 149 879 | 73 719 | 129 707 | 407 695 | 187 655 | 147 653 | 203 471 | 217 458 | 247 391 |
G1_182	26.42	2543.2	Unidentified G1_182; mz: 327 999 | 461 944 | 535 652 | 255 575 | 473 504 | 537 502 | 295 466 | 459 390 | 415 381 | 463 371 |
G1_183	26.45	2542.3	Unidentified G1_183; mz: 473 999 | 474 393 | 327 337 | 446 323 | 461 169 | 256 167 | 373 141 |
G1_184	26.56	2553.4	Unidentified G1_184; mz: 361 999 | 73 903 | 363 764 | 362 224 | 163 162 | 315 161 | 319 156 | 387 142 | 299 128 | 345 125 |
G1_185	26.57	2554.2	Unidentified G1_185; mz: 361 999 | 273 667 | 73 596 | 217 481 | 191 343 | 362 302 | 169 277 | 147 259 | 349 252 | 243 234 |
G1_186	26.69	2567.2	Unidentified G1_186; mz: 446 999 | 415 899 | 447 422 | 416 329 | 214 302 | 327 234 | 245 208 | 313 198 | 347 186 | 81 168 |
G1_187	26.78	2576.6	Salicin (?TMS)
G1_188	26.82	2581.2	Unidentified G1_188; mz: 203 999 | 217 458 | 313 413 | 131 364 | 544 268 | 218 254 | 148 252 | 242 172 | 387 168 | 109 159 |
G1_189	27.18	2613.0	Unidentified G1_189; mz: 371 999 | 372 322 | 373 102 | 459 40 | 238 11 | 342 8 | 311 5 | 385 4 |
G1_190	27.19	2613.7	Unidentified G1_190; mz: 203 999 | 147 513 | 73 263 | 109 258 | 95 237 | 83 189 | 125 129 | 137 124 | 148 118 | 57 117 |
G1_191	27.18	2618.0	Unidentified G1_191; mz: 361 999 | 204 986 | 73 681 | 243 541 | 217 512 | 331 499 | 319 489 | 129 376 | 362 349 | 169 347 |
G1_192	27.35	2634.9	Salicylic acid glucopyranoside (5TMS)
G1_193	27.37	2637.4	Hydroquinone-B-D-glucopyranoside (5TMS)
G1_194	27.53	2648.6	Unidentified G1_194; mz: 73 999 | 535 928 | 147 912 | 536 434 | 246 426 | 274 332 | 333 326 | 537 310 | 285 305 | 375 295 |
G1_195	27.77	2678.8	Unidentified G1_195; mz: 259 999 | 73 711 | 191 698 | 217 684 | 147 584 | 204 353 | 260 230 | 243 209 | 189 182 | 160 140 |
G1_196	27.99	2707.6	Sucrose (8TMS)
G1_197	28.16	2722.0	Unidentified G1_197; mz: 73 999 | 361 635 | 169 596 | 243 460 | 217 399 | 268 389 | 147 386 | 129 281 | 271 264 | 149 244 |
G1_198	28.34	2737.0	Unidentified G1_198; mz: 73 999 | 433 933 | 343 513 | 434 333 | 129 254 | 225 196 | 353 192 | 204 189 | 344 175 | 345 164 |
G1_199	28.50	2753.4	Unidentified G1_199; mz: 356 999 | 73 815 | 169 681 | 194 547 | 217 524 | 357 316 | 450 266 | 147 248 | 267 166 | 451 154 |
G1_200	28.55	2759.2	Unidentified G1_200; mz: 361 999 | 362 251 | 437 251 | 169 250 | 271 115 | 347 94 | 331 45 | 245 40 | 439 40 | 230 23 |
G1_201	28.66	2769.8	Unidentified G1_201; mz: 361 999 | 147 533 | 73 337 | 362 269 | 363 223 | 217 213 | 271 209 | 243 177 | 319 115 | 331 114 |
G1_202	28.84	2788.8	Unidentified G1_202; mz: 361 999 | 73 325 | 362 314 | 169 305 | 271 248 | 243 180 | 363 152 | 217 139 | 244 44 | 257 42 |
G1_203	28.89	2794.0	Unidentified G1_203; mz: 399 999 | 203 315 | 400 250 | 267 77 | 129 66 | 401 53 | 73 52 | 204 38 | 123 31 | 341 27 |
G1_204	28.98	2804.3	Trehalose (8TMS)
G1_205	29.21	2832.0	Unidentified G1_205; mz: 361 999 | 73 308 | 362 284 | 169 232 | 217 228 | 147 211 | 363 129 | 271 81 | 191 72 | 149 68 |
G1_206	29.49	2867.7	Unidentified G1_206; mz: 355 999 | 361 713 | 73 606 | 217 476 | 169 422 | 362 385 | 356 336 | 147 264 | 243 259 | 283 241 |
G1_207	29.77	2902.0	Unidentified G1_207; mz: 361 999 | 73 759 | 169 608 | 312 594 | 147 445 | 243 422 | 297 411 | 217 388 | 271 342 | 362 293 |
G1_208	29.97	2927.7	Unidentified G1_208; mz: 361 999 | 342 976 | 169 938 | 73 772 | 327 702 | 217 695 | 147 629 | 362 567 | 343 455 | 129 443 |
G1_209	30.04	2935.4	Unidentified G1_209; mz: 373 999 | 374 250 | 539 121 | 207 99 | 257 95 | 332 69 | 133 64 | 131 50 | 449 48 | 157 42 |
G1_210	30.10	2943.4	Unidentified G1_210; mz: 361 999 | 73 769 | 169 534 | 443 473 | 243 342 | 271 308 | 362 275 | 281 260 | 129 250 | 444 193 |

List of all 221 metabolites resolved from GC/MS chromatograms - Page 4 of 4

Peak#	RTime	RI	Identity. Where unknown: ten (if possible) most abundant masses (mz: mass relative-abundance basepeak999|)
G1_211	30.50	2992.6	Unidentified G1_211; mz: 427 999 | 203 407 | 428 287 | 147 267 | 429 111 | 97 107 | 81 102 | 129 91 | 111 72 | 83 54 |
G1_212	30.99	3053.0	Unidentified G1_212; mz: 297 999 | 217 395 | 450 339 | 73 262 | 362 250 | 243 220 | 169 184 | 225 170 | 207 164 | 299 133 |
G1_213	31.09	3065.2	Galactinol (9TMS)
G1_214	31.50	3116.1	Populin (?TMS)
G1_215	32.13	3194.1	Digalactosylglycerol (9TMS)
G1_216	32.28	3212.8	Unidentified G1_216; mz: 119 999 | 133 772 | 207 383 | 73 376 | 147 180 | 105 167 | 134 140 | 117 139 | 205 139 | 171 129 |
G1_217	33.06	3316.2	Unidentified G1_217; mz: 73 999 | 217 940 | 147 527 | 389 503 | 450 430 | 195 391 | 105 390 | 243 307 | 232 248 | 271 240 |
G1_218	33.09	3320.8	Unidentified G1_217; mz: 361 999 | 73 476 | 169 377 | 362 306 | 147 301 | 271 203 | 129 179 | 363 155 | 155 110 | 191 96 |
G1_219	33.62	3392.4	B-Sitosterol (1TMS)
G1_220	34.21	3470.9	Raffinose (11TMS)
G1_221	34.57	3518.0	Unidentified G1_221; mz: 361 999 | 362 348 | 204 210 | 73 196 | 169 170 | 243 118 | 437 105 | 257 90 | 135 80 |

Appendix B.2. Entire list of metabolites identified in LC/MS profiles. RT = retention time, MW = apparent molecular weight of molecular ion.

List of all 52 metabolites resolved from LC/MS chromatograms		
Peak#	**RT (min)**	**Identity**
L2_001	3.19	Coumaroyl glucoside; MW326
L2_002	3.24	Unidentified L2_002
L2_003	3.38	Unidentified L2_001; MW348
L2_004	3.51	Unidentified L2_004
L2_005	3.60	p-Coumaryl shikimate; MW320
L2_006	3.81	Unidentified L2_006
L2_007	4.00	Unidentified L2_007
L2_008	4.51	Unidentified L2_008
L2_009	4.66	Unidentified L2_009
L2_010	4.72	Unidentified L2_010
L2_011	5.24	Unidentified L2_011
L2_012	5.39	Unidentified L2_012
L2_013	5.37	Pinoquercetin (b); MW316
L2_014	5.55	Unidentified L2_014; MW324
L2_015	5.69	Salicin?; MW123?
L2_016	6.20	Unidentified L2_016
L2_017	6.16	Unidentified L2_017
L2_018	6.49	Unidentified L2_018
L2_019	6.61	Catechol; MW110
L2_020	6.81	Unidentified L2_020; MW402
L2_021	6.89	Unidentified L2_021; MW434
L2_022	7.09	Unidentified L2_022
L2_023	7.23	Vitexin; MW432
L2_024	7.37	Unidentified L2_024
L2_025	7.46	Unidentified L2_025
L2_026	7.55	Unidentified L2_026
L2_027	7.97	Unidentified L2_027
L2_028	7.97	Unidentified L2_028; MW518
L2_029	8.26	Unidentified L2_029; MW550
L2_030	8.58	Unidentified L2_030; MW442
L2_031	8.66	Unidentified L2_031
L2_032	8.69	Unidentified L2_032; MW404
L2_033	8.81	Unidentified L2_032; MW286
L2_034	8.83	Dihydromyricetin; MW320
L2_035	8.89	Grandidentatin; MW424
L2_036	9.19	3-ferulolquinic acid; MW368
L2_037	9.86	Unidentified L2_037; MW576
L2_038	10.36	Unidentified L2_038; MW406
L2_039	10.68	Salireposide; MW406
L2_040	10.90	Unidentified L2_040; MW132
L2_041	11.25	Salicortin; MW424
L2_042	11.31	Unidentified L2_042
L2_043	11.43	Phenyllactic acid; MW166
L2_044	11.77	Unidentified L2_044; MW454
L2_045	12.66	Unidentified L2_045; MW562
L2_046	12.76	Unidentified L2_046
L2_047	13.43	Unidentified L2_047; MW264
L2_048	13.71	Unidentified L2_048; MW584
L2_049	14.26	Unidentified L2_049; MW466
L2_050	14.99	Unidentified L2_050; MW506
L2_051	16.81	Unidentified L2_051; MW528
L2_052	20.29	Unidentified L2_052; MW666

Appendix B.3. List of all differential metabolites in the comparison between P39 and P717 hybrid poplar backgrounds, based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Average abundance of metabolites in P717 are expressed relative to P39.

Comparison between P39 and P717 wild-types

GC/MS metabolites different between backgrounds

Peak#	Identity	Average abundance P717 rel:P39
G1_002	Unidentified G1_002	1.70
G1_003	Pyruvic acid (1MEOX) (1TMS)	0.52
G1_005	Glycolic acid (2TMS)	0.68
G1_010	2-Pyrrolidinone (1TMS)	0.04
G1_012	Unidentified G1_012	1.68
G1_014	Unidentified G1_014	1.27
G1_015	Unidentified G1_015	1.70
G1_017	Unidentified G1_017	0.29
G1_018	Urea (2TMS)	0.29
G1_023	Unidentified G1_023	1.55
G1_024	Ethanolamine (3TMS)	1.56
G1_026	Phosphoric acid (3TMS)	1.81
G1_035	Glyceric acid (3TMS)	0.50
G1_036	Unidentified G1_036	1.59
G1_037	Unidentified G1_037	1.74
G1_039	Unidentified G1_039	1.71
G1_045	3-Hydroxymyristic acid (2TMS)	1.71
G1_046	Unidentified G1_044	0.11
G1_048	2-Hydroxybenzyl alcohol (2TMS)	1.57
G1_049	Unidentified G1_049	3.70
G1_050	Unidentified G1_050; Phenolic	1.42
G1_053	Malic acid (3TMS)	0.75
G1_054	Unidentified G1_054	0.40
G1_055	L-Asparagine (2TMS)	4.35
G1_059	Pyroglutamic acid (2TMS)	0.66
G1_060	4-Aminobutyric acid (3TMS)	0.65
G1_062	L-Norvaline (3TMS)	3.64
G1_063	Unidentified G1_063	0.63
G1_067	Unidentified G1_067; Organic acid	0.42
G1_075	Unidentified G1_075; Amino acid	3.14
G1_076	Unidentified G1_076; junk	3.00
G1_077	Unidentified G1_076; junk	3.17
G1_079	4-Hydroxybenzoic acid (2TMS)	1.39
G1_089	D-Ribonic acid lactone (3TMS)	0.27
G1_092	Unidentified G1_092	1.68
G1_096	Unidentified G1_096	0.69
G1_104	Unidentified G1_104; Carbohydrate	0.52
G1_105	Unidentified G1_105	0.65
G1_108	Unidentified G1_108	0.61
G1_109	Ribonic acid (5TMS)	0.59
G1_113	Unidentified G1_113	4.39
G1_119	Unidentified G1_119	9.08
G1_120	Unidentified G1_120	0.41
G1_123	Quinic acid (5TMS)	0.13
G1_124	Fructose MEOX (5TMS)	0.29
G1_125	Sorbose MEOX (5TMS) [BP]	0.26
G1_126	Fructose MEOX (5TMS) [BP]	0.21
G1_132	Galactitol (6TMS)	2.27
G1_133	Unidentified G1_133; Organic acid	0.67
G1_135	Unidentified G1_135	0.36
G1_137	Unidentified G1_135; Phenolic	1.55
G1_138	Unidentified G1_138	1.75
G1_140	Glutamine (4TMS) rep?	0.60
G1_143	Palmitic acid (1TMS)	1.61
G1_145	Galactaric acid (6TMS)	1.62
G1_147	Unidentified G1_147	0.46
G1_149	Unidentified G1_149; Organic acid	1.95
G1_151	Inositol (6TMS)	0.37
G1_152	3-Deoxy-arabino-hexaric acid (5TMS)	1.94
G1_156	Unidentified G1_156; Phenolic	1.39
G1_160	Unidentified G1_160	0.26
G1_162	Unidentified G1_162	0.70
G1_167	Unidentified G1_167	0.75
G1_169	Glucose-6-phosphate MEOX (6TMS)	0.64
G1_171	Unidentified G1_171; Phenolic	0.25
G1_173	Unidentified G1_173; Carbohydrate	0.44
G1_176	Unidentified G1_176	42.86
G1_178	Unidentified G1_178	0.36
G1_182	Unidentified G1_182	0.26
G1_183	Unidentified G1_183	1.63
G1_187	Salicin (?TMS)	1.89
G1_196	Sucrose (8TMS)	0.70
G1_198	Unidentified G1_198	4.81
G1_199	Unidentified G1_199	0.31
G1_204	Trehalose (8TMS)	0.26
G1_206	Unidentified G1_206; Phenolic	1.92
G1_208	Unidentified G1_208	0.79
G1_210	Unidentified G1_210	0.38
G1_215	Digalactosylglycerol (9TMS)	0.62

LC/MS metabolites different between backgrounds

Peak#	Identity	Average abundance P717 rel:P39
L2_002	Unidentified L2_002	0.31
L2_005	p-Coumaryl shikimate; MW320	0.53
L2_007	Unidentified L2_007	0.72
L2_008	Unidentified L2_008	0.73
L2_012	Unidentified L2_012	0.65
L2_013	Pinoquercetin; MW316	0.51
L2_016	Unidentified L2_016	0.67
L2_017	Unidentified L2_017	0.11
L2_018	Unidentified L2_018	2.55
L2_019	Catechol; MW110	1.81
L2_020	Unidentified L2_020; MW402	2.81
L2_021	Unidentified L2_021; MW434	0.44
L2_023	Vitexin; MW432	0.50
L2_024	Unidentified L2_024	0.55
L2_025	Unidentified L2_025	0.39
L2_027	Unidentified L2_027	14.66
L2_028	Unidentified L2_028; MW518	0.47
L2_029	Unidentified L2_029; MW550	0.47
L2_030	Unidentified L2_030; MW442	1.28
L2_031	Unidentified L2_031	0.43
L2_032	Unidentified L2_032; MW404	0.47
L2_033	Unidentified L2_032; MW286	1.79
L2_036	3-ferulolquinic acid; MW368	30.40
L2_040	Unidentified L2_040; MW132	17.68
L2_041	Salicortin; MW424	1.64
L2_042	Unidentified L2_042	0.19
L2_043	Phenyllactic acid; MW166	21.85
L2_044	Unidentified L2_044; MW454	2.93
L2_045	Unidentified L2_045; MW562	2.28
L2_049	Unidentified L2_049; MW466	7.06
L2_050	Unidentified L2_050; MW506	10.77

Appendix B.4. List of all "collective" differential metabolites in the comparisons between P39 C3′H-RNAi and wild-type background and P717 C3′H-RNAi and wild-type background, based on Select Lines dataset. Significance of differences determined by Student's t-test (α = 0.01). Average abundance of metabolites in modified lines are expressed relative to respective wild-type background.

Genetic Construct: C3'H-RNAi

GC/MS metabolites different between Mod and WT

Peak#	Identity	Avg abundance rel:WT P39	P717
G1_002	Unidentified G1_002		0.84
G1_003	Pyruvic acid (1MEOX) (1TMS)		0.56
G1_005	Glycolic acid (2TMS)		0.44
G1_006	L-Alanine (2TMS)	2.08	
G1_011	Unidentified G1_011	0.63	
G1_014	Unidentified G1_014		0.83
G1_017	Unidentified G1_017		0.27
G1_018	Urea (2TMS)		0.59
G1_030	Glycine (3TMS)		0.51
G1_031	Succinic acid (2TMS)	0.27	0.34
G1_033	Unidentified G1_033		0.65
G1_034	Unidentified G1_034		0.69
G1_035	Glyceric acid (3TMS)		0.41
G1_036	Unidentified G1_036		0.70
G1_038	Fumaric acid (2TMS)		0.72
G1_041	L-Serine (3TMS)	2.48	
G1_042	Cytosine (2TMS)		0.76
G1_044	Unidentified G1_044	3.40	4.59
G1_047	Unidentified G1_047	0.47	0.64
G1_050	Unidentified G1_050; Phenolic		0.73
G1_053	Malic acid (3TMS)	0.54	0.60
G1_054	Unidentified G1_054	0.24	0.30
G1_059	Pyroglutamic acid (2TMS)	0.61	0.67
G1_060	4-Aminobutyric acid (3TMS)	1.44	0.72
G1_063	Unidentified G1_063	0.49	0.56
G1_065	Threonic acid (4TMS)		0.62
G1_067	Unidentified G1_067; Organic acid		0.31
G1_071	Glutamine (4TMS) split peak 1		0.36
G1_072	Glutamine (4TMS) split peak 2		0.34
G1_074	Unidentified G1_074; Organic acid	0.47	0.55
G1_079	4-Hydroxybenzoic acid (2TMS)	0.57	
G1_081	Unidentified G1_081		0.36
G1_088	2,4,5-Trihydroxypentanoic acid (4TMS)		0.44
G1_096	Unidentified G1_096	0.39	0.59
G1_097	Xylitol (5TMS)		0.63
G1_102	Unidentified G1_101; Sugar alcohol	0.26	0.35
G1_103	Unidentified G1_103		0.46
G1_104	Unidentified G1_104; Carbohydrate		0.52
G1_105	Unidentified G1_105		0.48
G1_106	L-Glycerol-3-phosphate (4TMS)	0.64	0.67
G1_108	Unidentified G1_108		0.40
G1_109	Ribonic acid (5TMS)	0.41	0.40
G1_110	Unidentified G1_108; Organic acid	0.54	0.46
G1_111	Unidentified G1_111		0.61
G1_112	Shikimic acid (4TMS)		8.55
G1_114	Unidentified G1_114; Organic acid	0.46	0.49
G1_119	Unidentified G1_119		3.01
G1_120	Unidentified G1_120	1.54	
G1_123	Quinic acid (5TMS)		0.23
G1_124	Fructose MEOX (5TMS)		0.29
G1_125	Sorbose MEOX (5TMS) [BP]		0.44
G1_126	Fructose MEOX (5TMS) [BP]		0.24
G1_128	Glucose MEOX (5TMS)		0.22
G1_130	Glucose MEOX (5TMS) [BP]		0.18
G1_132	Galactitol (6TMS)	0.72	0.57
G1_133	Unidentified G1_133; Organic acid	0.50	0.31
G1_135	Unidentified G1_135		0.32
G1_137	Unidentified G1_135; Phenolic		0.76
G1_139	Unidentified G1_139; Carbohydrate		0.26
G1_140	Glutamine (4TMS) rep?		0.51
G1_141	Gluconic acid (6TMS)	0.06	0.24
G1_142	Galactonic acid (6TMS)	0.08	
G1_143	Palmitic acid (1TMS)		0.76
G1_144	Glucaric acid (6TMS)	0.15	0.17
G1_145	Galactaric acid (6TMS)	0.20	0.18
G1_147	Unidentified G1_147	0.37	0.33
G1_148	Unidentified G1_148; Organic acid	0.17	0.21
G1_149	Unidentified G1_149; Organic acid	0.15	0.15
G1_150	Unidentified G1_150; Organic acid	0.37	0.46
G1_151	Inositol (6TMS)		0.73
G1_152	3-Deoxy-arabino-hexaric acid (5TMS)	0.35	0.38
G1_154	Unidentified G1_154	0.28	
G1_156	Unidentified G1_156; Phenolic		0.64
G1_158	Octadecadienoic acid (1TMS)	0.59	
G1_160	Unidentified G1_160		0.22
G1_162	Unidentified G1_162		0.59
G1_167	Unidentified G1_167	0.65	0.52
G1_168	Galactose-6-phosphate MEOX (TMS)		0.54
G1_169	Glucose-6-phosphate MEOX (6TMS)	0.61	0.52
G1_170	Glucose-6-phosphate MEOX (6TMS) 2nd Pk	0.52	0.54
G1_172	Unidentified G1_172		0.76
G1_173	Unidentified G1_173; Carbohydrate		0.70
G1_174	Unidentified G1_174	0.21	0.23
G1_175	Unidentified G1_175; Carbohydrate	0.19	0.22
G1_178	Unidentified G1_178		0.20
G1_181	Unidentified G1_181		0.34
G1_183	Unidentified G1_183	0.14	0.16
G1_186	Unidentified G1_186		0.13
G1_187	Salicin (?TMS)		0.38
G1_189	Unidentified G1_189		0.70
G1_190	Unidentified G1_190; Fatty acid		0.72
G1_194	Unidentified G1_194		0.67
G1_196	Sucrose (8TMS)		0.62
G1_197	Unidentified G1_197; Glycoside	10.50	4.42
G1_198	Unidentified G1_198	0.08	
G1_203	Unidentified G1_203; Phenolic		0.72
G1_204	Trehalose (8TMS)		0.22
G1_205	Unidentified G1_205; Carbohydrate	26.14	3.15
G1_208	Unidentified G1_208		0.35
G1_209	Unidentified G1_209	0.56	0.74
G1_210	Unidentified G1_210		0.34
G1_212	Unidentified G1_212	0.39	0.20
G1_213	Galactinol (9TMS)		0.38
G1_219	B-Sitosterol (1TMS)		0.65

LC/MS metabolites different between Mod and WT

Peak#	Identity	Avg abundance rel:WT P39	P717
L2_001	Coumaroyl glucoside; MW326	24.48	12.40
L2_002	Unidentified L2_002		2.10
L2_003	Unidentified L2_001; MW348	0.68	0.32
L2_004	Unidentified L2_004	0.41	0.28
L2_005	p-Coumaryl shikimate; MW320	14.63	11.39
L2_006	Unidentified L2_006	0.25	0.18
L2_007	Unidentified L2_007	0.36	0.30
L2_008	Unidentified L2_008	0.61	0.57
L2_009	Unidentified L2_009	0.25	0.36
L2_010	Unidentified L2_010		1.44
L2_011	Unidentified L2_011	3.06	
L2_013	Pinoquercetin; MW316		0.55
L2_014	Unidentified L2_014; MW324	0.35	0.40
L2_015	Salicin?; MW123?		0.63
L2_016	Unidentified L2_016		0.32
L2_017	Unidentified L2_017		0.59
L2_018	Unidentified L2_018	0.50	0.31
L2_019	Catechol; MW110		0.64
L2_020	Unidentified L2_020; MW402		0.65
L2_021	Unidentified L2_021; MW434		0.61
L2_023	Vitexin; MW432		0.57
L2_024	Unidentified L2_024	0.51	0.24
L2_025	Unidentified L2_025		0.58
L2_026	Unidentified L2_026		0.25
L2_029	Unidentified L2_029; MW550	0.67	0.57
L2_030	Unidentified L2_030; MW442	0.55	0.66
L2_031	Unidentified L2_031		0.61
L2_032	Unidentified L2_032; MW404		0.51
L2_034	Dihydromyricetin; MW320	1.70	
L2_035	Grandidentatin L2_034; MW424	20.45	69.00
L2_036	3-ferulolquinic acid; MW368	0.23	0.54
L2_037	Unidentified L2_037; MW576		0.52
L2_038	Unidentified L2_038; MW406		0.58
L2_040	Unidentified L2_040; MW132	0.70	
L2_041	Salicortin; MW424	4.32	1.73
L2_043	Phenyllactic acid; MW166	0.65	0.47
L2_046	Unidentified L2_046		0.59
L2_047	Unidentified L2_047; MW264	0.37	0.51
L2_048	Unidentified L2_048; MW584	0.03	0.03
L2_049	Unidentified L2_049; MW466		0.58

APPENDIX C

Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and NIR-based prediction modeling

A version of this appendix was accepted for publication in The Plant Journal, Jan 14th 2009. Robinson, A.R., and Mansfield, S.D. 2009. Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and NIR-based prediction modeling. doi: 10.1111/j.1365-313X.2009.03808.x

Summary

Determination of the physico-chemical attributes of plant cell walls, such as lignin content and composition, is of paramount importance in germplasm screening, and for evaluating the results of plant breeding and genetic engineering. There are escalating needs for analyses to be robust, reproducible, accurate, and efficient. We have recently modified an established protocol for lignin monomer discrimination, thioacidolysis, with the goal of increasing sample throughput while maintaining accuracy and reducing equipment load and reagent consumption.

Numerous methodological changes related to volume scaling, processing vessel selection, and sample handling were addressed. The revised protocol permitted rapid processing of *ca.* 50+ samples per person per day. A direct comparison between methods using hybrid poplar (P. alba x tremula) wood samples, resulted in *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) lignin monomer quantities that were equivalent to those derived from the original protocol. The revised methodology was then applied to quickly generate phenotypic trait data from 267 hybrid poplar trees (including wild-type, and eight C4H::F5H transgenic lines), for the development of an NIR-based model for predicting lignin monomer proportion across a broad phenotypic range of S:G. The resulting PLSR model performed well under full cross-validation, giving strong, linear relationships between actual and predicted monomer proportions, and very high predictive accuracy for the predominant G and S monomers. This research brings considerable refinement to the thioacidolysis procedure, and establishes a method for rapidly and accurately quantifying cell wall lignin composition that could effectively be employed in routine phenotypic screening platforms.

Introduction

The lignin heteropolymer is an integral cell wall constituent that significantly influences the physical properties of plants, via its involvement in architectural support, water transport and defence. Lignins comprise the second most abundant polymer class in the biosphere, and their combinatorial biosynthesis renders these polymers among the more complex biomacromolecules synthesized by plants. This intricate macromolecule is assembled via the combinatorial free radical coupling of monolignol precursors derived from three *p*-hydroxycinnamyl alcohols varying in their degree of methoxylation (Ralph et al. 2004), resulting in varying proportions of *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) lignin sub-units according to plant species, tissue type and response to environmental stress (Campbell and Sederoff, 1996). The effective use of plants for a range of natural and industrial purposes is largely dependent on the extent to which the plant cell wall is lignified. This is particularly true when considering biomass conversion for bioenergy, forage digestibility in ruminants, or processes such as pulping and papermaking from woody materials. The effective removal of lignin is expensive and often the limiting factor to many applications. Lignin is therefore a key plant breeding or genetic engineering target to provide improvements in cell wall conversion. Consequently, analytical techniques that permit the precise determination of the abundance and chemical attributes of lignin have become important tools in the analysis of plant cell walls. Advances in lignin analysis, in terms of both the efficiency and accuracy of determination, are necessary now due to the demand for medium- to high-throughput phenotypic screens to identify valuable plant germplasm for specific applications (*e.g.* bioenergy crops), especially in the case when the individuals originate from large breeding studies, mutant populations or mapping families.

One important aspect of lignin composition that can affect the utility of the plant cell wall is the proportional content resulting from the different monomeric units. For example, in chemical pulping, lignin monomer ratio has been shown to significantly impact delignification efficiency (both pulp yield and residual lignin content) and pulp bleachability (Huntley et al., 2003; Stewart et al., 2006). An established technique for determining monomeric composition is the thioacidolysis reaction. By this approach, the uncondensed arylglycerol-β-aryl ether linked *p*-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) monomers of lignin may be cleaved from the polymer and the abundances

determined (Rolando et al., 1992). Although proven and valuable, the established protocol for this analysis is time, reagent, and equipment intensive, especially where the processing of large sample batches is concerned, and is therefore not suitable for high-throughput screens.

In contrast, several spectroscopic techniques have been employed to analyse/screen plant cell walls, including near infrared reflectance (NIR) (Blakeney and Flinn, 2005; Huang et al., 2007; Via et al., 2007) and Fourier-Transform Infrared (FT-IR) spectroscopies (Carpita and McCann, 2002; Chen et al., 1998). NIR uses infrared overtones and combination vibrations, while FT-IR employs mid-infrared regions based primarily on functional and fundamental vibrations. NIR is a promising technique for the rapid determination of physical and chemical wood properties, based on calibration and estimation rather than direct measurement. Although, the NIR signatures are often overlapping and make direct structural assignment difficult, there are several good examples of NIR-based modeling in the determination of lignin composition, in terms of both monomer ratio and total lignin content (Alves et al., 2006; Bailleres et al., 2002; Huang et al., 2007; Jones et al., 2006; Li et al., 2007; Maranan and Laborie, 2008; Poke and Raymond, 2006; Sykes et al., 2005; Takayama et al., 1997; Yamada et al., 2006; Yeh et al., 2004). However, this trait prediction process still first requires the measurement of NIR spectra and compositional trait(s) (e.g. the determination of H, G and S monomer composition of lignin by wet chemical techniques, such as thioacidolysis or nitrobenzene oxidation) in a 'calibration' population that covers the phenotypic range of interest, in order for a predictive model to be generated. Furthermore, distinct prediction models are typically required for each plant species, or at least for each genus of interest. However, once an accurate model is established, the easily obtained NIR reflectance data is all that will be required to allow the estimation of compositional traits in samples with undetermined characteristics.

The research herein describes a modified thioacidolysis protocol that enables the medium- to high-throughput screening of plant cell wall lignin monomer composition. We describe the details of a series of simple revisions to the original thioacidolysis protocol that reduce its resource demand while permitting a marked increase in sample throughput. Furthermore, we utilise the described wet-chemical procedure to rapidly

develop an accurate NIR-based prediction model for lignin monomer composition in hybrid poplar.

Materials and methods

Plant material and sample preparation

All trees were propagated from apical explants by sterile tissue culture techniques (three weeks growth cycle), and then re-potted into a soil medium. Trees were then grown in a temperature-controlled greenhouse under a fixed 16 h photoperiod with supplemental lighting (radiant flux density of 300 W m^{-2}). Daily watering with fertigated water was achieved by flood table irrigation. The trees were harvested twelve weeks after transfer to the greenhouse, and individual wood samples collected from each stem, 10 cm above the base of the root collar. The samples were then stripped of bark and pith and oven-dried (50 °C) for two days. Once dry, the wood was ground to a flour in a Wiley mill (40 mesh), and extracted for 12 h with hot acetone in a Soxhlet apparatus to remove extractives.

Thioacidolysis procedure

Wood samples involved in the comparison between thioacidolysis procedures were initially processed as per Rolando et al. (1992) (Appendix D.6), using 10 mg of ground, extract-free, oven-dried wood flour as the substrate, and a ratio of 1 ml of reaction mixture to 1 mg of sample. The same samples, as well as those used in NIR-based predictive modeling, were also processed according to the following revised protocol. For each sample, 10 mg of ground, extract-free, oven-dried wood flour was weighed into a glass 5 ml Wheaton vial with teflon-lined screw-cap. One ml of freshly made reaction mixture (10% boron trifluoride etherate and 2.5% ethanethiol, in recently distilled dioxane (v/v)) was added to each vial and blanked with nitrogen gas prior to sealing. Vials were then placed together in a dry heating block (100 °C) for 4 h, with periodic (hourly) manual agitation. The reaction was halted by placing the reactions at -20 °C for 5 min. Internal standard (5 mg/ml tetracosane in methylene chloride, 0.2 ml) was then added to each vial, and enough 0.4 M sodium bicarbonate to bring reaction pH to between 3 and 4 (~0.3 ml, as determined by pH indicator paper). To extract the reaction products from the aqueous mixture, 2 ml of water and 1 ml of methylene chloride were added to each vial, which was then recapped, vortexed, and allowed to

settle, phase-separating the upper (aqueous) and lower (organic, and containing lignin breakdown products) phases. An aliquot (1.5 ml) of the organic phase was removed by autopipette, and simultaneously cleared of residual water and filtered by passing through a Pasteur pipette packed with a small tissue paper plug and an inch (~50 mg) of granular anhydrous sodium sulphate, and transferred directly into a 2 ml polypropylene microfuge tube. Samples were then collectively evaporated to dryness in a Vacufuge (Eppendorf) (approximately 1.5 h at 45 °C), and resuspended in 1 ml of methylene chloride.

Samples prepared by both conventional and optimised methods were derivatised by combining 20 µl of resuspended sample with 20 µl of pyridine and 100 µl of N,O-bis(trimethylsilyl) acetamide (Sigma). After incubation for at least 2 h at 25 °C, 1 µl of this reaction product was analysed by gas chromatography (GC).

Gas chromatography

Gas chromatography was conducted on a Hewlett Packard 5890 series II instrument, fitted with an autosampler, splitless injector, flame ionising detector (FID), and 30 m RTX5ms 0.25 mm ID capillary column. One microlitre injections were separated using helium as a carrier gas at 1 ml/min. Inlet and detector temperatures were set to 250 °C, while the oven profile consisted of: initial temperature 130 °C, hold 3 min, ramp temperature 3 °C/min for 40 min to give a final temperature of 250 °C, hold 5 min, cool. Peak identification was consistent with that in Rolando et al. (1992). The yield and relative concentration of hydroxyphenyl, guaiacyl, and syringyl lignin was determined from the response factor for each, as the ratio of relative concentration to the relative peak area of the internal standard.

NIR Spectroscopy

The light reflectance of wood samples across the near infrared spectrum was measured with an Analytical Spectral Devices Inc., QualitySpec Pro near infrared (NIR) spectrophotometer, equipped with a round 1.5 cm diameter sample window (Muglite). The scanning range was from 350–2500 nm, with a 2 nm interval, interpolated to 1 nm.

Statistical analysis

Prediction modeling was conducted using the Partial Least Squares Regression (PLSR) package provided in ‘The Unscrambler 9.1’ software (Camo Technologies, Woodbridge, New Jersey), employing full cross-validation as a modeling option. Prior to PLSR, NIR

reflectance data were transformed into the Savitzky-Golay first derivative, with the averaging/smoothing process spanning 25 wavelengths either side of each data point, and an order of 2 for the polynomial approximation process.

Results

Thioacidolysis reaction

An initial experiment comparing the performance of the original (Lapierre et al. 1985; Rolando et al., 1992) and our revised thioacidolysis method was conducted to determine the relative efficacy of *p*-hydroxyphenyl, guaiacyl and syringyl lignin monomer subunit quantification, in woody plant tissue. In order to estimate the monomer proportions and total monomers yield (as determined by both methods) ground, extract-free wood samples from single trees of three lines of hybrid poplar P717 (*P. alba x tremula*) were analysed, with five technical replicates conducted on each sample. These trees represent two transgenic lines, both of which exhibit markedly altered wood lignin chemistry, and the wild-type control. The transgenic lines included an overexpressing C4H::F5H line, which exhibited a considerable increase in the S:G ratio of lignin, without significant effect on *p*-hydroxyphenyl monomers or the relative lignin content of the wood, as previously determined by thioacidolysis (Franke et al., 2000; Huntley et al., 2003; Meyer et al., 1996; Meyer et al., 1998), and an RNAi-suppressed line targeting *p*–coumaroyl-CoA 3-hydroxylase (C3′H), with substantially reduced levels of cell wall lignification and elevated levels of *p*-hydroxyphenyl units (Coleman et al., 2008).

The intention was to detect, measure and combine the contributions from all of the three major lignin monomer subunits liberated from the cell wall lignin, where H monomers represent 4-hydroxyphenyl, G monomers represent 4-hydroxy-3-methoxyphenyl, and S monomers represent 4-hydroxy-3,5-dimethoxyphenyl subunits. The monomers were detected as both the *erythro*- and *threo*-isomers of (H, G or S)–CHR-CHR-CH_2-R, and (H,G or S)-CHR-CHR-CHR_2, as per Rolando et al. (1992). In practice, clean, discrete, measurements of each form was only possible for the G- and S-derived monomers, as the H-derived monomers were at trace-levels in most samples, and the peaks frequently did not resolve well from one another in gas chromatography. Approximate elution times under the conditions described herein were: 22.2–23.5 min

for H-derived monomers, 23.8–24.3 min for G-derived monomers, and 25.8–26.4 min for S-derived monomers (Figure 1).

Lignin subunit ratios and total monomer yield differed between the lines, with the wild-type exhibiting an S:G ratio of ~70:30 as is typical of many angiosperm deciduous species, the over-expressing C4H::F5H transgenic line exhibiting an elevated S:G ratio (~84:15), and the C3H:RNAi transgenic line exhibiting a decrease in total monomer yield of ~40% (compared to wild-type), with a concurrent increase (~10%) in H monomers largely at the expense of G monomers (Table 1).

It is apparent from Figure 2 and Table 1 that the averaged composition of lignin monomers, liberated as a result of thioacidolysis, was very similar when comparing the original and revised methods, with differences generally falling within a single standard deviation. Similarly, thioacidolysis yield (as measured by the sum of the *threo*- and *erythro*-isomers of H, G, and S-monomers) was very similar when. Furthermore, the level of error does not appear to favour any monomer over another.

Traditional thioacidolysis techniques permit, on average, the preparation of ~6-10 samples per day prior to monomer quantification by GC analysis. In contrast, the modified method permits upwards of 50 individual samples to be prepared in a single day, which can then be run on the GC for analysis the subsequent day. This substantially increases preparation number, and offers a medium- to high-throughput analytical protocol for accurately determining plant cell wall lignin monomer subunit composition. Furthermore, the reduced volumes of reaction solvents (5- to 10-fold reduction in dioxane, BF_3 and ethane thiol) and purifying steps (up to 20-fold reduction in methylene chloride rinse volume *etc*.) offer added savings in preparation costs.

Prediction of lignin monomer proportion via NIR spectroscopy

Following confirmation that the modified thioacidolysis protocol was suitable to quantify lignin monomer composition of plant cell wall moieties, 267 independent thioacidolysis reactions were run to determine the monomer composition of wild-type and transgenic wood samples originating from individual trees exhibiting a board spectrum of lignin monomer compositions. These same samples were then NIR-scanned and the relationship between H, S and G lignin proportions and the NIR reflectance data (both raw and derivative transformed) modelled by PLSR analysis, with complete cross-validation. Substantial variation in monomer composition was achieved by including

trees from eight C4H::F5H transformed lines originating from independent transformation events, and that displayed a change in S:G ratio when compared with the wild-type control trees. These were the same lines previously analysed by Huntley *et al.* (2003). The sample and phenotypic structures of the dataset are described in Table 2.

A predictive model including the first eight (of 15 calculated) principal components was deemed most suitable, as this model accounted for 95.896% of variance in the response variables, while the addition of subsequent components resulted in only marginal improvement (97.669% with all 15 components). The resultant NIR prediction model proved to be highly accurate under cross-validation as indicated in the plots of measured versus predicted monomer proportion (Figure 2), and in the corresponding table of regression descriptors (Table 3). As anticipated (due to the low level and quantification difficultly for H-monomers), the accuracy of prediction was notably less for H proportion than for G and S.

Discussion

Revised thioacidolysis procedure

In this post-genomic era, there is a pressing need to develop high-throughput phenotyping tools to work in parallel with, and complement, the rapidly advancing functional genomics toolbox (*i.e.* transcriptomics, proteomics and metabolomics). Furthermore, advanced breeding programs, mapping populations and mutant populations, as well as our traditional plant resources, require techniques that permit the accurate rapid characterization of plant cell walls, with minimal inputs (time, reagents and equipment). Thus, the principal aim of this research was to modify an accepted protocol (*i.e.* thioacidolysis) to improve throughput, while concurrently maintaining accuracy. The resultant protocol could then serve as a standalone platform for plant cell wall monomer compositional analysis, or as the wet chemical procedure required to establish predictive models based on other analytical tools, such as spectroscopy. Because independent models are typically required on a species by species basis, due to the substantial variation in chemical and ultrastructural architecture of plants, extensive calibrations sets continue to be needed and the greater the number of samples, the higher the statistical accuracy.

Revisions to the procedure were attempted to increase the efficiency of the traditional thioacidolysis reaction employed to liberate lignin monomers for quantification by gas chromatograph. Several iterations and modifications were attempted (data not shown) but, attesting to the development by the procedures originators, in the end the established chemical reaction was retained, and most revisions were methodological rather than empirical. Essentially, the reaction volume was scaled down with as few changes to chemical ratios as possible. Volume scaling was central to achieving higher efficiency, as this allowed simultaneous processing of many samples via the use of alternative equipment. Simply reducing the reaction volume to small-scale 1 ml levels facilitated large-scale incubation of several reactions together in a temperature-controlled heating dry-block. Furthermore, the reduction in reaction volumes permitted the aqueous quenching and the subsequent organic solvent extraction procedures to be conducted in a single 5 ml glass vial, without the need for a separatory funnel. Additionally, the final volume of extract in organic solvent was less than 2 ml, which allowed for the rapid evaporation of a large number of samples together in a Vacufuge (Eppendorf). Ultimately, these simple modifications now permit 50 or more samples to be processed per person, per day, without the need for specialized equipment and with substantially lower chemical input.

Equivalence of results from original and revised thioacidolysis protocols

Although the procedural revisions involved mainly scaling and equipment choice, there were some changes that potentially could have affected the analysis results. Comparison between the performance of the original and revised protocols, as applied to samples with highly deviating lignin monomer composition, was intended to assess any possible effects.

One major concern during scaling was the decision to supply the same amount of sample (as the traditional method) to the modified reaction in order to maintain weighing accuracy and good sample representation and homogeneity. This meant that, in the revised protocol, the ratio between reaction volume and sample weight was decreased ten-fold compared to the original method. In their discussion of the analysis of lignocellulose residues (*i.e.* extractive-free wood flour) using thioacidolysis, Rolando *et al.* (1992) suggest that because the reagent is quickly consumed through reaction with glycosidic units in the sample, it is advisable to provide surplus reagent in order to

ensure that this does not become a limiting factor. They suggest a minimum of 0.5 ml of reagent per mg of tissue – a ratio that was reduced considerably in the revised protocol. Another concern was the decision to reduce both the relative volume and number of organic solvent extraction cycles to isolate the lignin breakdown products from the aqueous mixture, post-reaction. This was done primarily to reduce sample handling time, and to keep the final extract volume within the 2 ml capacity needed for ease of sample handling (single glass vial) and evaporation by Vacufuge (Eppendorf). However, this carried with it the risk of decreasing product yield or skewing monomer representation.

The results indicate that the revisions did marginally reduce yield (~5%), but had a limited influence on monomer quantification (Table 1). Yield in thioacidolysis is ascribed to the β–O–4-ether content and the extent of its cleavage; only β-ether units that are either free-phenolic or are further 4–O–β-etherified will release the thioacidolysis monomers. However, it has been shown that thioacidolysis does not completely cleave all existing β-ethers (Ralph and Grabber, 1996), and as such the minor reductions in observed yield in the revised protocol may be a function of using lower reagent loadings. Furthermore, in poplar, some lignin units are gamma-*p*-hydroxybenzoylated; these units do not completely release the (unacylated) monomer (Grabber et al. 1996), and therefore as a consequence of reagent scaling may account for the under-quantification. The yield is therefore always an underestimate of β-ether content, but if linkage analysis is the objective, then the modified protocol will deliver a further slight underestimation compared to the traditional method. However, monomer proportion, rather than yield, is the primary concern of the thioacidolysis method. As long as the yield is high enough to facilitate accurate measurement of relative abundance, this analysis has achieved its goal. The shifts observed in monomer proportion were small enough (and frequently statistically insignificant) that the measurements generated by the original and revised protocols could be considered ‘substantially equivalent’. This study therefore indicates that for wood of varied lignin composition, from 12 week-old hybrid poplar, as little as 0.1 ml of reaction mixture per mg of tissue is sufficient to maintain a good yield, and accurate quantification. From a purely pragmatic point of view, especially for use as a large population screening tool, the slight decrease in yield and minimal shifts in monomer proportions observed under

the revised protocol are suggested to be acceptable, given the considerable resource- and time-related benefits it provides.

Predictive modeling of lignin monomer proportion

According to visual and numerical indicators, the performance of the NIR-based predictive model for lignin monomer proportion was excellent, and comparable to the results achieved in other species and applications (Alves et al., 2006; Bailleres et al., 2002; Maranan and Laborie, 2008; Yamada et al., 2006). Notably, for G and S monomers prediction in cross-validation exhibited minimal error across the range of phenotypic extremes. Because the correlation coefficients and slopes of the regression lines of these monomers' predicted value against their measured value are approaching one, a high level of confidence could be placed in further predictions made by this model for additional, similar wood specimens. The comparatively low accuracy of prediction for H monomers is likely related to the ordinarily low abundance of this monomer in poplar in combination with the error possibly introduced when measuring near the baseline of GC traces, and the limited variation observed in H monomer abundance across the samples modeled (the C4H::F5H construct had little effect on relative H monomer content).

The cross-validation regression plots (Figure 2) supports the relationship existing between G and S monomer proportion in angiosperm wood of C4H::F5H modified trees, as has previously been shown (Huntley et al. 2003) by thioacidolysis (Table 2). This is most apparent in the plots of regression coefficients for NIR wavelengths against the prediction model, in which the strength and importance of individual NIR wavelengths for G and S monomer prediction are complementary (Figures 1E and 1F). It is probable that the strength of this relationship contributed substantially to the accuracy of the predictive model, as PLSR involving multiple response variables takes implicit advantage of such patterns.

Concluding remarks

It is apparent from these findings that the revised thioacidolysis protocol described is a useful alternative to the classic procedure. This new protocol promotes a substantial increase in sample processing efficiency while markedly reducing equipment load and

reagent consumption, and achieves this with little effect on the determination of lignin composition. The increased performance under this protocol makes thioacidolysis an effective tool for rapidly generating the large sets of phenotypic data required for NIR-based predictive modeling of lignin compositional traits. In applying such modeling to the wood of hybrid poplar, it was confirmed that data generated via the revised thioacidolysis protocol was of sufficient quality to allow accurate modeling of the relationships existing between NIR reflectance spectra and the proportion of the primary monomeric constituents in lignin (as measured by thioacidolysis), across a broad phenotypic range.

Acknowledgements

Funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada Grants held by SDM. AR would also like to recognise the Top Achiever Doctoral Scholarship (Bright Future Scholarships, Tertiary Education Commission, Wellington, N.Z.) held between 2002 and 2005.

Figures

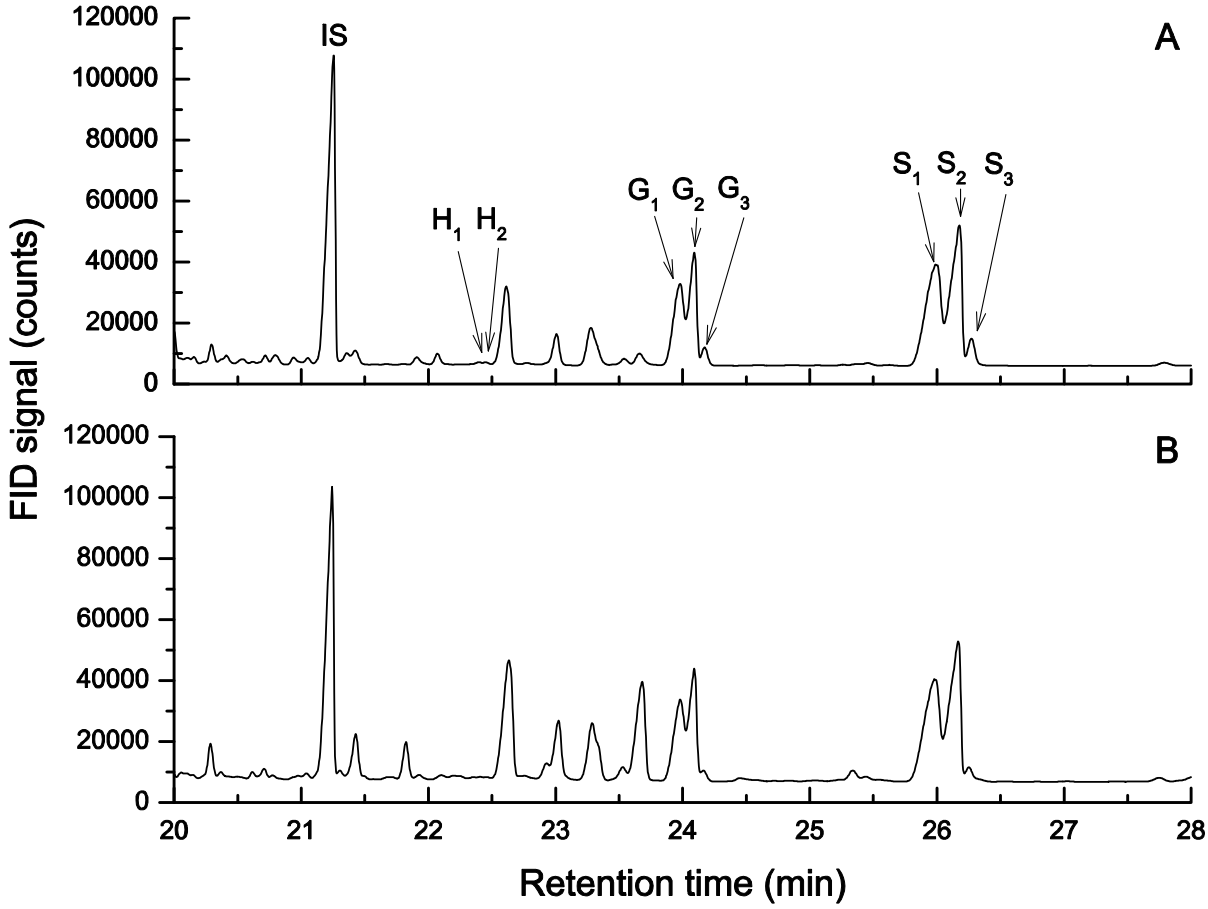
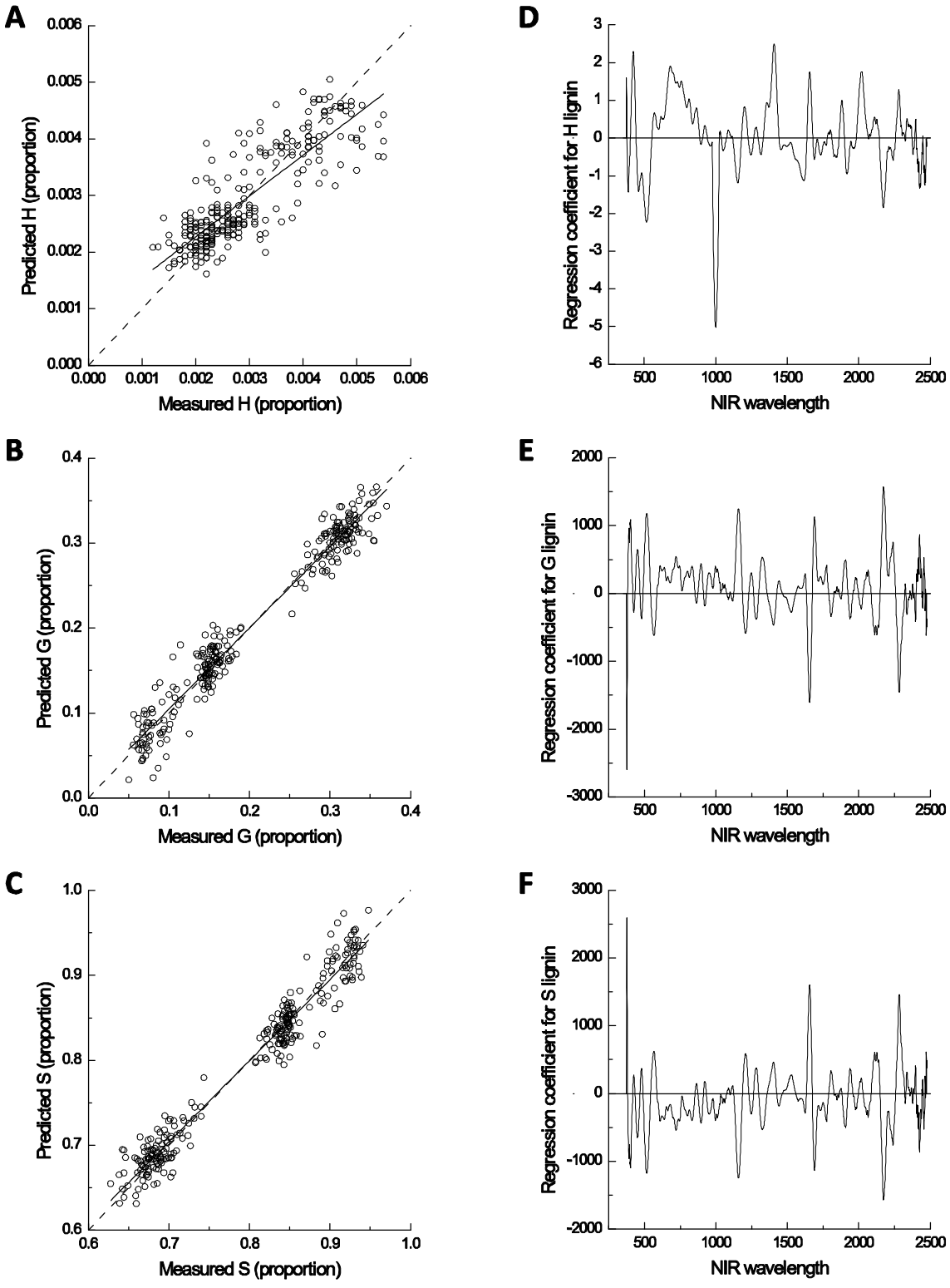


Figure 1. Typical GC FID traces for wild-type P717 poplar thioacidolysis samples prepared by A) original, large-scale method, and B) revised, small scale method. H_1: *p*-hydroxyphenyl-CHR-CHR-CH_2R erythro. H_2: *p*-hydroxyphenyl-CHR-CHR-CH_2R threo. G_1 = Guaiacyl-CHR-CHR-CH_2R erythro. G_2 = Guaiacyl-CHR-CHR-CH_2R threo. G_3 = Guaiacyl-CH_2-CHR-CHR_2. S_1 = Syringly-CHR-CHR-CH_2R erythro. S_2 = Syringyl-CHR-CHR-CH_2R threo. S_3 = Syringyl-CH_2-CHR-CHR_2.

(Overleaf)

Figure 2. NIR-based PLSR prediction modeling of lignin monomer proportion in P717 poplar exhibiting variation in S:G ratio due to transgene activity. Results presented are derived from a model using the first eight principal components, under cross-validation. Figures A-C are plots of measured *vs.* predicted values for lignin monomer proportion for *p*-hydroxyphenyl, guaiacyl and syringyl monomers, respectively. Solid line represents regression lines of best fit, while dashed line represents optimal 1:1 relationship between measured and predicted values. Figures D-F are the corresponding plots of regression coefficients for the first derivative NIR wavelengths with the PLSR-derived prediction model, for *p*-hydroxyphenyl, guaiacyl and syringyl monomers, respectively (also based on the first eight principal components).


A
Predicted H (proportion)
Measured H (proportion)
0.000
0.001
0.002
0.003
0.004
0.005
0.006
B
Predicted G (proportion)
Measured G (proportion)
0.0
0.1
0.2
0.3
0.4
C
Predicted S (proportion)
Measured S (proportion)
0.6
0.7
0.8
0.9
1.0
D
Regression coefficient for H lignin
NIR wavelength
-6
-5
-4
-3
-2
-1
0
1
2
3
500
1000
1500
2000
2500
E
Regression coefficient for G lignin
NIR wavelength
-3000
-2000
-1000
0
1000
2000
F
Regression coefficient for S lignin
NIR wavelength
-2000
-1000
0
1000
2000
3000

Tables

Line/ treatment	Thio lignin yield (µMol/g Klason lignin)		Thio lignin composition (proportion)					
			Mean			StdDev		
	Mean	Std dev	H	G	S	H	G	S
P7-WT-R	1798.30	19.32	0.0037	0.3128	0.6835	0.0002	0.0034	0.0033
P7-WT-O	1941.93	89.14	0.0031	0.3161	0.6808	0.0002	0.0023	0.0024
P7-CF-R	2013.52	108.50	0.0039	0.1490	0.8471	0.0006	0.0015	0.0013
P7-CF-O	2129.42	326.68	0.0043	0.1527	0.8430	0.0009	0.0007	0.0004
P7-CR-R	1574.62	54.67	0.1212	0.2355	0.6433	0.0038	0.0049	0.0023
P7-CR-O	1649.18	254.84	0.1053	0.2413	0.6534	0.0035	0.0018	0.0033

Table 1. Lignin yield and proportion measured by original and revised thioacidolysis reaction procedures. Line/treatment descriptors: P7 = P717 hybrid poplar, WT = wild type, CF = C4H::F5H, CR = C3H:RNAi, and suffixes 'O' and 'R' indicate the use of 'original' or 'revised' protocol. 'H', 'G' and 'S' = *p*-hydroxyphenyl, guaiacyl and syringyl lignin liberated by thioacidolysis. Mean and standard deviation values are based on five technical reaction replicates conducted with homogenised, ground and extracted wood samples. 'StdDev' = standard deviation.

Tree line	Tree count	Thio lignin yield (µMol/g Klason lignin)		Proportion of lignin monomers by class					
				Hydroxyphenyl		Guaiacyl		Syringyl	
		Mean	Std dev	Mean	StdDev	Mean	StdDev	Mean	StdDev
P7-WT	104	2321.59	280.12	0.0028	0.0010	0.3158	0.0209	0.6814	0.0206
P7-CF-A	9	2519.57	100.95	0.0028	0.0006	0.2753	0.0197	0.7219	0.0196
P7-CF-B	82	2504.30	275.10	0.0033	0.0012	0.1556	0.0116	0.8411	0.0110
P7-CF-C	7	2750.77	152.40	0.0025	0.0006	0.1473	0.0143	0.8502	0.0146
P7-CF-D	5	2634.48	76.71	0.0023	0.0006	0.1360	0.0381	0.8617	0.0379
P7-CF-E	5	2771.64	147.65	0.0026	0.0003	0.1103	0.0116	0.8872	0.0117
P7-CF-F	12	2790.52	150.8	0.0028	0.0005	0.0883	0.0125	0.9089	0.0126
P7-CF-G	6	2773.49	116.35	0.0032	0.0006	0.0854	0.0069	0.9114	0.0069
P7-CF-H	37	2707.53	242.47	0.0028	0.0009	0.0803	0.0390	0.9169	0.0393
Total	267								

Table 2. Lignin properties and sample counts of tree lines employed in NIR-based modeling of lignin monomer proportions by PLSR. Tree line descriptors: P7 = P717 hybrid poplar, WT = wild type, CF = C4H::F5H, and an alphabetic suffix denotes identity of transgenic lines derived from single transformation events. Lines have been ordered according to increasing S:G ratio. Lignin monomer proportions were determined by the revised thioacidolysis procedure, described herein. 'StdDev' = standard deviation.

	Model calibration			Model validation		
Descriptor	H	G	S	H	G	S
Slope	0.731	0.956	0.968	0.721	0.956	0.956
Offset	0.008	0.090	0.255	0.008	0.090	0.346
Correlation	0.855	0.979	0.984	0.843	0.979	0.979
R-square	0.731	0.957	0.968	0.710	0.957	0.958
RMSEC/RMSEP	0.005	0.202	0.176	0.005	0.202	0.201

Table 3. Performance indicators of NIR-based PLSR model for prediction of lignin monomer proportion, as determined by a comparison between actual measured and model-predicted values. Indicators are provided for both model calibration and model cross-validation scenarios. Corresponding graphical representation of cross-validation scenario is provided in Figures 1A-C. 'Descriptor' explanations are as follows: 'Slope' = slope of the regression line between measured and predicted values (ideally = +/-1), 'Offset' = Y-intercept of the regression line, 'Correlations' = correlation coefficient between measured and predicted values, 'R-square' = coefficient of determination (a measure of the degree of fit of the regression), RMSEC/RMSEP = root mean square error of calibration/prediction (the average modeling error).

References

Alves, A., Schwanninger, M., Pereira, H., and Rodrigues, J. (2006). Calibration of NIR to assess lignin composition (H/G ratio) in maritime pine wood using analytical pyrolysis as the reference method. *Holzforschung* **60,** 29-31.

Bailleres, H., Davrieus, F., and Pichavant, F.H. (2002). Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. *Ann. For. Sci.* **59,** 479-490.

Blakeney, A.B. and Flinn, P.C. (2005). Determination of non-starch polysaccharides in cereal grains with near-infared reflectance spectroscopy. *Mol. Nutr. Food Res.* **49,** 546-550.

Campbell, M.M. and Sederoff, R.R. (1996). Variation in lignin content and composition - Mechanism of control and implications for the genetic improvement of plants. *Plant Physiol.* **110,** 3-13.

Carpita, N.C. and McCann, M.C. (2002). The functions of cell wall polysaccharides in composition and architecture revealed through mutations. *Plant Soil* **247,** 71-80.

Chen, L.M., Carpita, N.C., Reiter, W.D., Wilson, R.H., Jeffries, C., and McCann, M.C. (1998). A rapid method to screen for cell-wall mutants using discriminant analysis of Fourier transform infrared spectra. *Plant J.* **16,** 385-392.

Coleman, H.D., Park, J.Y., Nair, R., Chapple, C., and Mansfield, S.D. (2008). RNAi-mediated suppression of p-coumaroyl-CoA 3'-hydroxylase in hybrid poplar impacts lignin deposition and soluble secondary metabolism. *Proc. Natl. Acad. Sci., USA* **105,** 4501-4506.

Franke, R., McMichael, C.M., Meyer, K., Shirley, A.M., Cusumano, J.C., and Chapple, C. (2000). Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *Plant J.* **22,** 223-234.

Grabber, J.H., Quideau, S. and Ralph, J. (1996). *p*-Coumaroylated syringyl units in maize lignin; implications for β-ether cleavage by thioacidolysis. *Phytochem.* **43**, 1189-1194.

Huang, A.M., Jiang, Z.H., and Li, G.Y. (2007). Determination of holocellulose and lignin content in Chinese fir by near infrared spectroscopy. *Spectrosc. Spectr. Anal.* **27,** 1328-1331.

Huntley, S.K., Ellis, D., Gilbert, M., Chapple, C., and Mansfield, S.D. (2003). Significant increases in pulping efficiency in C4H-F5H-transformed poplars: Improved chemical savings and reduced environmental toxins. *J. Agric. Food Chem.* **51,** 6178-6183.

Jones, P.D., Schimleck, L.R., Peter, G.F., Daniels, R.F., and Clark, A. (2006). Nondestructive estimation of wood chemical composition of sections of radial wood strips by diffuse reflectance near infrared spectroscopy. *Wood Sci. Technol.* **40,** 709-720.

Lapierre, C., Monties, B., and Rolando, C. (1985). Thioacidolysis of lignin: comparison with acidolysis. *J. Wood Chem. Technol*. **5**, 277-292.

Li, G.Y., Huang, A.M., Wang, G., Qin, D.C., and Jiang, Z.H. (2007). Rapid determination of mason lignin content in bamboo by NIR. *Spectrosc. Spectr. Anal.* **27,** 1977-1980.

Maranan, M.C. and Laborie, M.P.G. (2008). Rapid prediction of the chemical traits of hybrid poplar with near infrared spectroscopy. *J. Biobased Mater. Bioenergy* **2,** 57-63.

Meyer, K., Cusumano, J.C., Somerville, C., and Chapple, C.C.S. (1996). Ferulate-5-hydroxylase from Arabidopsis thaliana defines a new family of cytochrome P450-dependent monooxygenases. *Proc. Natl. Acad. Sci., USA* **93,** 6869-6874.

Meyer, K., Shirley, A.M., Cusumano, J.C., Bell-Lelong, D.A., and Chapple, C. (1998). Lignin monomer composition is determined by the expression of a cytochrome P450-dependent monooxygenase in Arabidopsis. *Proc. Natl. Acad. Sci., USA* **95,** 6619-6623.

Poke, F.S. and Raymond, C.A. (2006). Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in *Eucalyptus globulus*. *J. Wood Chem. Technol.* **26,** 187-199.

Ralph, J. and Grabber, J.H. (1996). Dimeric β-ether thioacidolysis products resulting from incomplete ether cleavage. Holzforschung. **50,** 425-428.

Ralph, J. Lundquist, K., Brunow, G., Lu, F., Kim, H., Schatz, P. F., Marita, J. M., Hatfield, R. D., Ralph, S. A., Christensen, J. H. and Boerjan, W. (2004). Lignins: natural polymers from oxidative coupling of 4-hydroxyphenylpropanoids. *Phytochem. Rev.* **3**, 29-60.

Rolando, C., Monties, B., and LaPierre, C. (1992). Thioacidolysis. in Lin, S. and Dence, C. (Eds), Methods in Lignin Chemistry. Springer-Verlag, Berlin, pp. 334-349.

Stewart, J.J., Kadla, J.F., and Mansfield, S.D. (2006). The influence of lignin chemistry and ultrastructure on the pulping efficiency of clonal aspen (Populus tremuloides Michx.). *Holzforschung* **60,** 111-122.

Sykes, R., Li, B.L., Hodge, G., Goldfarb, B., Kadla, J., and Chang, H.M. (2005). Prediction of loblolly pine wood properties using transmittance near-infrared spectroscopy. *Can. J. For. Res.* **35,** 2423-2431.

Takayama, M., Johjima, T., Yamanaka, T., Wariishi, H., and Tanaka, H. (1997). Fourier transform Raman assignment of guaiacyl and syringyl marker bands for lignin determination. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.* **53,** 1621-1628.

Via, B.K., So, C.L., Groom, L.H., Shupe, T.F., Stine, M., and Wikaira, J. (2007). Within tree variation of lignin, extractives, and microfibril angle coupled with the theoretical and near infrared modeling of microfibril angle. *IAWA J.* **28,** 189-209.

Yamada, T., Yeh, T.F., Chang, H.M., Li, L.G., Kadla, J.F., and Chiang, V.L. (2006). Rapid analysis of transgenic trees using transmittance near-infrared spectroscopy (NIR). *Holzforschung* **60,** 24-28.

Yeh, T.F., Chang, H.M., and Kadla, J.F. (2004). Rapid prediction of solid wood lignin content using transmittance near-infrared spectroscopy. *J. Agric. Food Chem.* **52,** 1435-1439.

APPENDIX D

Miscellaneous protocols

Appendix D.1. Klason procedure for the determination of total lignin content

Extractive-free, oven-dried wood equivalent to 0.2 g and 0.15 g of oven-dried pulp was treated with 3 mL of cold (4 ºC), 72% H_2SO_4 (Fisher Scientific) at 20°C. The mixture was initially macerated continuously for 2 minutes, and then stirred every 10 minutes for 2 hours. Acid hydrolysis was stopped with the addition of cold deionized water, whereby the wood mixture was diluted with 112 mL of deionized water and the pulp with 112.5 mL of deionized water, to achieve a final acid concentration of 4% (w/w) H_2SO_4. The mixture was then transferred to a serum bottle, which was sealed with a septa cap and autoclaved (Castle Thermatic 60) at 121°C for 1 h.

Klason lignin was determined gravimetrically with the hydrolysates filtered through pre-weighed, medium coarseness, sintered-glass crucibles. The filtrate was added back to the septa bottles and re-filtered to ensure recovery of all solids, and the filtrate retained. The solids were then washed with 100 mL of 40°C deionized water and oven-dried at 105°C for 12 hours. The oven-dried crucibles with acid-insoluble lignin were then weighed.

The Klason lignin filtrate was further analysed using TAPPI Useful Method UM-250 to determine the portion of acid soluble lignin. 30 µL of wood hydrolysate samples were diluted with 970 µL of 4% H_2SO_4 in a test-tube and mixed thoroughly, such that spectophotometer (Milton Roy Spectronic 1001 Plus) absorbance readings were between 0.2 and 0.7 absorbance units (AU) at 205 nm. The solution was transferred to a quartz cuvette, 4% H_2SO_4 was used to calibrate the spectrophotometer and then absorbance values for each sample was taken. An expression of Beer's Law is used to calculate the percent of acid-soluble lignin as follows:

Acid-soluble lignin % = B·V·100/1000·W

Where:

B = Absorbance · volume of diluted filtrate / 110 · volume of original filtrate

V = Total volume of filtrate

W = Oven-dry weight of wood

Appendix D.2. Carbohydrate analysis of woody material

The Klason lignin filtrate was filtered through 0.45 mm HV filters (Millipore, MA, USA) prior to injection of a 20 µL sample volume. The HPLC system (Dionex DX-600, Dionex, CA, USA) was equipped with an ion exchange PA1, 4 × 250 mm, (Dionex) column, an ED50 Electrochemical detector (Dionex), and an AS 50 autosampler (Dionex). The column was equilibrated with 250 mM NaOH (BDH), eluted with deionized water at a flow rate of 1.0 mL/min., and post-column wash of 200 mM NaOH at a flow rate of 0.5 mL/min. Fucose (Sigma) (5 mg/mL) was used as an internal standard. Cell wall carbohydrates analysed in this manner included arabinose, galactose, glucose, mannose and xylose.

Appendix D.3. Preparation of Woody Plant Medium (WPM) (McCown and Lloyd, 1981)

WPM Medium (pH 5.6)

all units either mL or g; stock concentrations are as below

To volume dH_20	0.5 L	1 L	5 L	10 L
WPM A	10	20	100	200
WPM B	10	20	100	200
WPM C	2.5	5	25	50
WPM D	2.5	5	25	50
WPM E	2.5	5	25	50
LM FeEDTA	5	10	50	100
WPM Vitamins	0.5	1	5	10
MS Glycine	0.5	1	5	10
Ca gluconate	0.325	0.65	3.25	6.5
sucrose	10	20	100	200

pH 5.6, NaOH

500ml to each 1 ml autocalve bottle

to each 500 mL bottle add:

Agar	1.5
Phytagel	0.55

Steam autoclave for 20 min at 121°C at 15 psi,
plus time for heating, cooling, and pressure adjustment

WPM A (50X)	1 L
NH4NO3	20.0 g
Ca(NO3)2*4H2O	27.8 g

WPM B (50X)	1 L
K2SO4	49.5 g

WPM C (200X)	1 L
CaCl2*2H2O	19.2 g

WPM D (200X)	1 L
KH2PO4	34 g
H3BO3	1.24 g
Na2MoO4*2H2O	50 mg

WPM E (200X)	1 L
MgSO4*7H2O	74 g
MnSO4*4H2O	4.46 g
ZnSO4*7H2O	1.72 g
CuSO4*5H2O	50 mg

WPM Vitamins (1000X)	1 L
myo-inositol	100 g
thiamine-HCl	100 mg
nicotinic acid	500 mg
pyridoxine-HCl	500 mg

MS Glycine (1000X)	1 L
glycine	2 g

LM FeEDTA (100X)	1 L
FeSO4*7H2O	2.78 g
Na2EDTA*2H2O	3.74 g

Appendix D.4. RNA extraction (Kolosova et al 2004), and cDNA synthesis protocols

Treenomix - Total RNA Isolation and Quality Control Protocol – Version 1.0 March 30th, 2004

1. Total RNA isolation chemical and reagent inventory

Chemical Name	Formula Weight	Company	Catalogue
Tris (Hydroxymethyl) aminoethane (500g)	121.14	VWR	9210
Lauryl sulfate lithium salt (50g)	272.3	Sigma	L-9781
Lithium chloride (500g)	42.39	Sigma	L-9650
Disodium salt EDTA (500g)	372.24	VWR	EX0539-1
Deoxycholate acid, Sodium Salt (100g)	414.5	Fisher	BP349-100
Tergitol (100g)		Sigma	NP-40
Aurintricarboxylic acid	422.35	Sigma	A1895-5G
Dithiothreitol (10g)	154.2	Sigma	D-9779
Thiourea (100g)	76.12	Sigma	T-8656
2% PVPP (100g)		Sigma	P-6755
Sodium chloride (5Kg)	58.44	VWR	7760
Sodium acetate trihydrate (500g)	136.08	VWR	7610
CTAB (500g)	364.5	Sigma	H-5882
Isoamyl alcohol (500mL)	88.15	Sigma	I-9392
Agarose (500g)		Invitrogen	15510-027
Boric acid (1Kg)	61.83	VWR	2710
Glacial acetic (2.5L)		Fisher	A38-212
Ethidium Bromide (10mg/mL) (10mL)		BioRad	161-0433
Diethyl pyrocarbonate (100mL)	162.1	Sigma	D-5758
dATP (100mM)		Invitrogen	55082
dCTP (100mM)		Invitrogen	55083
dGTP (100mM)		Invitrogen	55084
dTTP (100mM)		Invitrogen	55085
Oligo $(dT)_{18}$ (500 ng/µL)		Qiagen	N/A
RNaseOUT (5,000U @ 40U/µL)		Invitrogen	10777019
M-MLV RT (40,000U @ 200U/µL)		Invitrogen	28025-013
5x First Strand Buffer (1mL)		Invitrogen	Y00146
0.1 M DTT (500µL)		Invitrogen	Y00147
5x T4 DNA polymerase buffer (1mL)		Invitrogen	Y02284
T4 DNA polymerase (50U @ 5U/µL)		Invitrogen	18005-017
1 Kb Ladder (250µg)		Invitrogen	15615-016
100 bp Ladder (50µg)		Invitrogen	15628-019
Microspin S-300 columns (50)		Amersham	27-5130-01
Alpha dGTP (^{32}P) 250 µCi (10µCi/µL) 3000 µCi/mmol)		Amersham	AA0006
100% ethanol			N/A
70 % ethanol			N/A
100% isopropanol			N/A
Chloroform (4 L)		Fisher	C298-41
Liquid Nitrogen			N/A
Sodium Dodecyl Sulfate (500 g)	288.38	VWR	7910

2. Total RNA isolation from spruce and poplar tissues

The following methods for total RNA isolation and quality control analysis were developed in the Treenomix laboratory by Natalia Kolosova, Dr. Barbara Miller, Dawn Cooper, Sharon Jancsik, Hesther Yueh and Dr. Steven Ralph. The total RNA isolation protocol is based on Wang et al. (2000) Isolation and purification of functional total RNA from woody branches and needles of Sitka and White Spruce. Biotechniques 28(2):292 and Chang et al. (1993) A Simple and Efficient Method for Isolating RNA from Pine Trees. Plant Molecular Biology Reporter 11(2):113.

a. Preparation of solutions for total RNA isolation

All glassware, stir bars, metal spatulas, etc. for preparing solutions, or for use during total RNA isolation should be baked for at least 4 hours at 180°C prior to use. Likewise, mortars and pestles should be treated in the same manner. This will inactivate any RNases that are present.

0.1% DEPC-treated H_2O – 1 L	Volume Added
DEPC	1 mL
Milli-Q H_2O	999 mL

Handle DEPC in fume hood. Mix for 1 hour at room temperature (closed bottle) and then autoclave. Store at room temperature.

Stock extraction buffer – 500 mL	Volume Added	Final Concentration
1 M Tris-HCl (pH 8.5)	200 mL	0.4 M
Lauryl sulfate lithium salt	7.5 g	55 mM
Lithium chloride (10 M)	15 mL	0.3 M
Disodium salt EDTA (0.5 M)	10 mL	10 mM
Sodium deoxycholate, Sodium Salt	5 g	24 mM
Tergitol NP-40	5 mL (solid-microwave to melt)	
DEPC-treated H_2O	to 500 mL	

Do not autoclave and store at 4°C.

Working extraction buffer

	200 mL	100 mL	150 mL	Final Conc.
Aurintricarboxylic acid	0.0844 g	0.0422 g	0.0633 g	1 mM
Dithiothreitol	0.3084 g	0.1542 g	0.2313 g	10 mM
Thiourea	0.0763 g	0.03815 g	0.057225 g	5 mM
PVPP*	4 g	2 g	3 g	2 %

*PVPP is insoluble. Addition of PVPP will cause the extraction buffer to become opaque. Also note that PVP and PVPP are NOT the same chemical. Do not autoclave working extraction buffer. Prepare and use the same day.

TE buffer pH 8.0 – 1 L	Volume Added	Final Concentration
1 M Tris (pH 8.0)	10 mL	10 mM
500 mM Disodium salt EDTA (pH 8.0)	2 mL	1 mM

Adjust to 1 L using DEPC-treated H_2O
Autoclave, aliquot and store at -20°C.

5 M sodium chloride
sodium chloride 146.1 g
Milli-Q H_2O to 500 mL

Add stock DEPC to 0.1%. Handle DEPC in fume hood. Mix for 1 hour at room temperature (closed bottle) and then autoclave. Aliquot and store at -20°C.

3.3 M sodium acetate pH 6.1

Sodium acetate trihydrate	224.5 g
Milli-Q H2O	to 500 mL

pH to 6.1 with HCl. Add stock DEPC to 0.1%. Handle DEPC in fume hood. Mix for 1 hour at room temperature (closed bottle) and then autoclave. Aliquot and store at -20°C.

10% CTAB

CTAB	50 g
DEPC-treated H_2O	to 500 mL

Do not autoclave. Heat to 65°C to dissolve prior to use. Store at room temperature.

10 M LiCl

Lithium chloride	211.9 g
DEPC-treated H_2O	to 500 mL

Filter sterilize, do not autoclave. Store at 4°C.

500 mM Disodium salt EDTA (pH 8.0)

Disodium salt EDTA	186.12 g
Mill-Q H_2O	to 1 L

pH to 8.0 with 5 M NaOH. Add stock DEPC to 0.1%. Handle DEPC in fume hood. Mix for 1 h at room temperature (closed bottle) and then autoclave. Aliquot and store at -20°C.

Chloroform:isoamyl alcohol (IAA) (24:1) - 1L

Chloroform	960 mL
Isoamyl alcohol	40 mL

Mix and store at room temperature.

100% ethanol
Store at -20°C.

Chilled isopropanol
Store at -20°C.

b. Isolation of total RNA

This protocol has been used to successfully isolate high quality total RNA from a variety of tissues from spruce (e.g. xylem, phloem, needles, roots, bark, cones) and poplar (e.g. xylem, phloem, leaves, stems, bark, flowers, roots). These RNA samples have been used to construct full-length cDNA libraries and for RNA expression analysis using cDNA microarrays and RT-PCR. This protocol is designed for tissue samples of wet weights between 0.5 g and 1.5 g. For samples of this size, 15 mL of working extraction buffer per sample is suitable. For tissues exceeding 1.5 g, the quantity of working extraction buffer must be scaled accordingly (15 mLs per 1.5 g) and the tissue must be divided after grinding into multiple units for processing. This protocol may not be suitable for tissue weights less than 0.5 g.

Protocol: **(Note: store all fractions at 4°C until RNA isolation has been verified)**		
	1.	Pre-chill mortar and pestle at –20°C.
	2.	Weigh & add tissue to chilled mortar, add sufficient liquid N_2 to freeze and grind tissue to powder using mortar. Do not allow powder to thaw during grinding.
	3.	Add 7.5 mL of working extraction buffer and mix well. Note that tissue and buffer will freeze solid. Continue to add more liquid N_2 and grind until tissue is again reduced to a powder.
**	4.	Transfer powder to 50 mL Falcon tube, add remaining 7.5 mL working extraction buffer, and vortex vigorously until consistent solution is achieved. Snap freeze in liquid N_2 for 10 seconds.
	5.	Place samples in waterbath at 37°C just until thawed, then invert 2 or 3 times to ensure a consistent solution.
	6.	Spin in Sorvall (RTH-750 rotor) at 4,000 rpm or Beckman (JS-5.3 rotor) at 3,000x g for 20 minutes at 4°C.
Steps 7-10 not needed for 'easy' tissues	7.	Filter supernatant through a folded kimwipe (placed in a funnel) into a 50 mL Falcon tube (keep tube buried in ice). This step is only required if there are large visible particles in the supernatant.
	8.	Add 1/30 volume of 3.3 M NaAcetate pH 6.1 and 1/10 volume 100% ethanol. If step 7 is omitted, transfer supernatant directly from step 6 into a new 50 mL Falcon tube and then add NaAcetate and ethanol. *Step 8 is required for tissues with a high polysaccharide content, which must be determined for each tissue. If after centrifugation (step 11), the polysaccharide pellet is very small then step 8 may be omitted. Examples of issues with high polysaccharide content include poplar leaves and some xylem samples, whereas spruce leader tissue is generally low in polysaccharides and does not require steps 7-10.*
	9.	Incubate on ice for 10 minutes.
	10.	Spin in Sorvall (RTH-750 rotor) at 4,000 rpm or Beckman (JS-5.3 rotor) at 3,000x g for 30 minutes at 4°C.
	11.	Transfer supernatant to new 50 mL Falcon tube and add 1/10 volume 3.3 M NaAcetate pH 6.1 and 100% volume chilled isopropanol. Invert several times to mix thoroughly. If solution is not clear, filter through kimwipe as in step 7.
**	12.	Store at –80°C for at least 30 minutes.
	13.	Thaw samples at room temperature and spin in Sorvall (RTH-750 rotor) at 4,000 rpm or Beckman (JS-5.3 rotor) at 3,000x g for 40 minutes at 4°C.
	14.	Pour off supernatant and resuspend pellet in 2 mL TE and 2 mL 5 M NaCl with periodic vortexing, until pellet is completely resuspended (minimum of 30 minutes on ice). Transfer to 15 mL Falcon tube.
	15.	Add 1 mL of 10% CTAB and mix by vortexing at room temperature.
	16.	Incubate 5 minutes in 65°C water bath.
	17.	Add an equal volume (5 mL) of chloroform:IAA and mix thoroughly with vortexing (20-30 seconds). Separate phases by centrifugation in Sorvall (RTH-750 rotor) at 4,000 rpm or Beckman (JS-5.3 rotor) at 3,000x g for 20 min at 4°C. Transfer aqueous phase (top) to a new Falcon tube and repeat extraction by adding another 5 mL of chloroform:IAA.
	18.	Add ¼ volume 10 M LiCl to aqueous phase and mix by vortexing.
	19.	Incubate overnight in –20°C freezer. Samples may be left at –20°C for a few days if

		necessary.
	20.	Centrifuge in Sorvall (RTH-750 rotor) at 4,000 rpm or Beckman (JS-5.3 rotor) at 3,000x g for 30 minutes at 4^{o}C.
	21.	Pour off supernatant and carefully remove all remaining liquid with pipette. Dissolve pellet in 900 µL TE buffer on ice, vortexing occasionally (note that resuspension may take up to 1 hour). Transfer sample into a 2 mL microfuge tube.
**	22.	Add 900 µL of chilled isopropanol (1 volume) and 100 µL of 3.3 M NaAcetate pH 6.1 (1/9 volume) to each tube. Mix by inversion and precipitate for 30 minutes at –80^{o}C.
	23.	Spin in Hettich microfuge at 14,000 rpm for 30 minutes at 4^{o}C to pellet RNA. Remove supernatant and wash pellet with 1 mL of 70% ethanol. Repeat spin for 10 minutes. Remove supernatant by pipette without disturbing pellet. Repeat spin for 1 minute. Remove any remaining liquid with pipette. Dry pellet for 5-10 minutes at room temperature and resuspend in 500 µL autoclaved DEPC-treated H_2O on ice for 20 - 30 minutes, vortexing occasionally. After QC analysis, divide into aliquots of 125 µL per 1.5 mL tube.

** At these steps samples may be stored overnight or up to a few days at –80oC.

The quality and quantity of total RNA samples is determined by first measuring UV absorbance, followed by gel electrophoresis analysis on a standard agarose gel.

c. UV absorbance quantification of RNA samples

a. While samples are resuspending in 500 µL of DEPC-H20, turn spectrophotometer on (located in GDC Rm 3342). The spectrophotometer requires ~15 minutes to warm up before it is operational. An error message will appear regarding one failed diagnostic/calibration test; simply press enter to continue. Record your name and period of use on the diagnostic test report that is generated each time the spectrophotometer is turned on (place on lower shelf directly above spectrophotometer).

b. Vortex RNA samples briefly to ensure pellet is thoroughly resuspended. Use tabletop microcentrifuge to collect liquid at the bottom of the microfuge tube. Prepare a dilution of each RNA sample in a 1.5 mL microfuge tube as follows:

Samples. 99 µL of TE Buffer (pH 8.0) and 1 µL of RNA sample (vortex briefly and centrifuge briefly to collect liquid at the bottom of the microfuge tube). Keep on ice.

Reference. 100 µL of TE Buffer (pH 8.0).

c. Once the instrument calibration is complete, initiate set-up by selecting “nucleic” and then “enter”. Then choose “scan” and press “enter”. The scan range for “nucleic” acids has a default setting of 200-350 nm and set points of 260 nm and 280 nm. This is appropriate for RNA scans and does not need to be adjusted. The “background on” option may be selected if additional set points are desired. For RNA analysis, the correction “factor” option must be set to 40 µg/mL in order to correctly determine the concentration of RNA from your scan. Press “enter” twice more to proceed to the next screen.

d. The instrument will prompt you to then press “run” on reference once the cuvette containing your reference sample has been placed in the holder (note that the measurement windows must be oriented from left to right). After your reference measurement is complete remove the cuvette from the instrument, rinse with distilled water, carefully tap the cuvette dry, add your first sample and place the cuvette back into the holder.

e. The instrument will then prompt you to press “run” on sample. If a chromatogram is required be sure to select “print” prior to measuring your next sample. The spectrophotometer does not save any record of your results. Additional samples may then be quantified after the cuvette is cleaned with distilled water and dried between samples.

f. Quality RNA samples will typically have absorbance measurements at 230 nm (polysaccharide peak) of less than 0.1 and a 260 nm/280 nm ratio of 2.0-2.4. A lower ratio is indicative of either protein and/or DNA contamination.
   i. The stock RNA concentration may be calculated by multiplying the “concentration” taken from the chromatogram by your dilution factor.
   ii The total yield is then obtained by multiplying your stock RNA concentration by the total volume of your stock RNA. The yield varies by tissue but typically ranges from 50-500 µg total RNA/g of tissue.

d. Gel electrophoresis analysis of RNA quality

Total RNA samples are resolved on standard agarose gels to determine if any RNA degradation has occurred. We have not found it necessary to use denaturing formaldehyde gels for RNA analysis as these are labour intensive to produce and dangerous to handle.

a. First clean gel combs and trays with either RnaseZap or a 1% Sodium Dodecyl Sulphate (SDS) solution to remove any RNases that may be present and lead to RNA degradation. Use only gel trays and combs that are reserved for RNA analysis and select sizes appropriate for the number of samples being processed.

b. All gels are prepared using a 1% agarose matrix (1 g of agarose per 100 mL buffer) in either 0.5X TBE or 1X TAE buffer (see below for buffer composition). Note that the buffer used to prepare the gel must be the same as the running buffer in the gel box. For small gels, use 10 mL of liquefied agarose and for large gels use 20 mL. Allow at least 30 minutes at room temperature for the gel to solidify.

c. Transfer gel into gel box and cover with appropriate buffer. Load 300 ng (6 µL) of 1 Kb ladder (Invitrogen) and 2 µg of sample (sample volume between 5-10 µL, adjusted with DEPC-H2O – add 1 µL of 10x Loading Dye per sample). Run at 100 volts until the bromophenol blue (purple) dye has migrated ¾ of the length of the gel.

d. Transfer gel to ethidium bromide (carcinogen – wear nitrile gloves) RNA staining solution and leave in the dark for 20-30 minutes.

e. Transfer gel to RNA destain container (DEPC-H20) for an additional 20-30 minutes.

f. Photograph gel using the UVP gel imaging system.
   i. Adjust image using coarse and fine focuses under white light, with aperture set to 14 and then switch to UV light (302 nm) and adjust integration time until foreground and background intensity are optimal. If signal intensity is unsatisfactory, repeat staining and/or destaining steps.

   ii. Dispose of all ethidium bromide contaminated materials (i.e. gel, gloves, paper towels etc.) in the ethidium bromide solid waster container adjacent to the gel electrophoresis station.

   iii. Quality RNA samples will have prominent ribosomal RNA bands with no sign of degradation (bright smear trailing below prominent ribosomal band). The number of ribosomal bands will vary (between 2 and 10) depending on the tree species and tissue source. There should be no evidence of genomic DNA present (high molecular weight bands greater than 10 Kb).

g. Solutions used in gel electrophoresis procedure

<u>5X TBE (1 L)</u>
54 g Tris base
27.5 g boric acid
20 mL 0.5 M pH8.0 EDTA
Bring to 1 L with distilled H2O.

Do not autoclave.
Store at room temperature.

<u>50X TAE (1 L)</u>
242 g Tris base
57.1 g glacial acetic
100 mL 0.5 M pH8.0 EDTA
Bring to 1 L with distilled H2O.

Do not autoclave.
Store at room temperature.

Ethidium bromide staining solution
100 mL DEPC-H2O
10 µL of Ethidium Bromide solution (Biorad-10mg/mL)

Store in dark at room temperature.

Destaining Solution (0.1% DEPC-H2O)

DEPC	1 mL
Milli-Q H2O	999 mL

Handle DEPC in fume hood. Mix for 1 hour at room temperature (closed bottle) and then autoclave. Store at room temperature.

3. cDNA synthesis from total RNA

DNA digest (if required): DNase I DIGEST (AMBION TURBO)

1. add 5 uL of TURBO 10X buffer to 10 ug of RNA in 44 uL (bring up with DEPC water)
2. add 1 uL of TURBO DNase (total volume now 50 uL)
3. incubate at 37°C for 30 minutes
4. add 5 uL of Inactivation solution (vortex this before using); vortex samples at setting 5 briefly to mix; let stand for 2 minutes at room temperature
5. spin at 10 000 g for 2 minutes at 4°C
6. transfer the supernatant to a fresh 1.5 mL microtube
7. freeze in -80°C, or use immediately; quantify using GeneQuant

First-strand cDNA synthesis: SUPERSCRIPT II RT (INVITROGEN)

Cat #: 18064-014 (10,000 units)

1. add 1 ug of RNA to a PCR tube, bring to 10 uL with DEPC water
    a. for triplicates, pool 0.33ug of RNA from each triplicate to give 1 ug
2. add 1 uL of oligo(dT_{18} 500 ug/mL) to mixture
3. add 1 uL dNTP (10 mM stock)
4. total volume now 12 uL
5. heat to 65°C for 5 minutes using in thermal cycler
6. add 4 uL of 5X First-Strand Buffer
7. add 2 uL of 0.1 M DTT
8. incubate 42°C for 2 min in thermal cycler
9. add 1 uL (200 units) of SuperScript II RT (mix by pipetting)
10. incubate @ 42°C for 50 minutes, then kill @ 70°C for 15 minutes, then 4°C hold
11. place cDNA in -20°C until needed

Appendix D.5. Poplar transformation

The protocol for transforming the two poplar hybrids *P. grandidentata* × *alba* (P39) and *P. tremula* × *alba* (P717), with the two constructs C4H::F5H and C3′H-RNAi, was as follows. Wild-type lines were transformed with *Agrobacterium tumefaciens* harbouring an appropriate binary plasmid, by a standard leaf disk inoculation technique. For C3′H-RNAi, this involved the *Agrobacterium tumefaciens* EHA105 strain harbouring the pCC600 binary. For C4H::F5H, this involved the *Agrobacterium tumefaciens* EHA105 strain harbouring the PCC153 (pGA482+C4H::F5H) binary. Plasmid-harbouring bacteria were incubated overnight in liquid woody plant medium with 2% sucrose (WPM) and 100 μM acetosyringone. Leaf disks were cut and cocultured with EHA105 for 1 h, blotted dry, plated onto WPM solidified with 3% (wt/vol) agar and 1.1% (wt/vol) phytagel, and supplemented with 0.1 μM each α-naphthalene acetic acid (NAA), 6-benzylaminopurine (BA), and thiadiazuron (TDZ) (WPM+NAA/BA/TDZ). After 3 days, the discs were transferred to WPM+NAA/BA/TDZ containing carbenicillin disodium (500 mg/L) and cefotaxime sodium salt (250 mg/L). After 3 additional days of selective growth, the discs were transferred to WPM+NAA/BA/TDZ containing carbenicillin, cefotaxime, and kanamycin (25 mg/L). After two consecutive 5-week periods on this medium, shoot tips were isolated to WPM with no antibiotics. For C3′H-RNAi, plants were confirmed as transgenic by PCR using 35S promoter (5′-gcagctgacgcgtacacaacaag-3′) and poplar CYP98 (C3′H-3)-specific primer oligonucleotides (5′-caattggggtaccgcagtgatca-3′), generating a 1.508 Kb amplicon. PCR amplification was achieved under the following conditions: hot start at 94°C for 3 min, 35 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 2 min, followed by 10 min at 72°C. For C4H::F5H, plants were confirmed as transgenic by PCT using F5H-specific primer oligonucleotides “cc14” (5′-GGTCCGGTCGGTCTCTTG-3′) and “cs278” (5′-TATCTCACCCGGAATTGCCT-3′), generating a ~1 Kb amplicon. PCR amplification was achieved under the following conditions: hot start at 95°C for 5 min, 35 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min, followed by 10 min at 72°C.

Appendix D.6. Traditional method for thioacidolysis (Ronaldo et al 1992)

Rolando, C., Monties, B., and LaPierre, C. (1992). Thioacidolysis. in Lin, S. and Dence, C. (Eds), Methods in Lignin Chemistry. Springer-Verlag, Berlin, pp. 334-349.

In some cases, the lignin monomer composition of woody material was determined following the traditional thioacidolysis procedure of Ronaldo et al (1992). Briefly, the solvolysis step began with extract-free wood ground through a 40-mesh screen using a Wiley mill and extracted for 12 hours with acetone in a Soxhlet apparatus. A 10 mg sample was weighed into a 25 mL Kimax test tube fitted with a Teflon coated cap. A 10 mL aliquot of freshly prepared reaction mixture (1:4:40 ratio of boron trifluoride diethyl etherate (Sigma), ethanethiol (Sigma), and dioxane (Fisher Scientific), respectively) was added to the wood. Nitrogen gas was used to evacuate the air, and the cap was closed tightly. The reaction was allowed to proceed in a 100°C silicon oil (Aldrich) bath for exactly 4 hours and was mixed by shaking every 30 min. Afterwards, the mixture was removed to an ice bath for 5 min to stop the reaction. 3 mL of deionised water and 2 mL of internal standard composed of 0.25 mg/mL tetracosane (Aldrich) in dichloromethane (Fisher Scientific), were then added to a separation funnel followed by the addition of the reaction mixture. The tubes were rinsed twice with 10 mL of deionised water and the contents emptied into the separation funnel. Approximately 4 mL of 0.4 M sodium bicarbonate in water was added to attain a pH of 3-4. Next, 30 mL of dichloromethane was added and the mixture in the separation vial was shaken well, left to separate for 5 minutes and then the lower fraction collected into an Erlenmeyer flask. The addition of 30 mL of dichloromethane was repeated twice more and the lower fractions pooled. The mixture was dried over excess Na_2SO_4 (Fisher Scientific), and filtered through Watman #4 filter paper into a 250 mL round bottom flask. The Erlenmeyer flask was rinsed twice with 15mL dichloromethane and the contents of the filter paper rinsed once with 30 mL dichloromethane.

The total pooled dichloromethane mixture was evaporated at ~40°C under reduced pressure (using a Büchi RE III (Switzerland) rotovap) until approximately 3 mL remained. 4 mL of methanol (Fisher Scientific) was added to the 3 mL mixture and evaporated (as described) to dryness. The residue in the round bottom flask was re-

suspended in 2 mL dichloromethane, transferred to a 5 mL vial fitted with a Teflon cap, wrapped with aluminium foil and stored at 4°C.