Metabolomics and wood development

Catching a metabolome

Annika I. Johansson
Faculty of Forest Sciences
Department of Forest Genetics and Plant Physiology
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2013

Acta Universitatis agriculturae Sueciae
2013: 92

Cover: Björkskogen
(Artist: Karin Öst)

ISSN 1652-6880
ISBN (print version) 978-91-576-7922-2
ISBN (electronic version) 978-91-576-7923-9

Print: Arkitektkopia, Umeå 2013

Metabolomics and wood development - Catching a metabolome

Abstract

The tree industry is one of Sweden's most important economic sectors, producing pulp, paper, wooden goods and wood as a bioenergy source. Understanding the biology behind the processes of tree growth and wood development is a key to maximizing the use of wood as raw material. The term "metabolome" refers to the complete set of small-molecule metabolites to be found within a biological sample. Since metabolites can be viewed as down-stream products of gene transcription, protein expression and enzymatic activity the metabolites are believed to reflect activity in cells more accurately than the transcript and protein complements. Metabolomics aims to identify and quantify all metabolites in a biological sample and thereby link biochemical information to a phenotype. In the course of the work underlying this thesis we have developed analytical tools for metabolomics and applied these tools to reveal metabolite patterns, with a focus on various aspects of wood development.

Firstly we development of a rapid and reliable method for processing GC-TOFMS data, that enables us to move from GC-MS chromatograms via metabolite identification to biological interpretation. To address the low identification ratios characteristic of GC-MS metabolomics studies, a database containing the predicted retention indices of ~13 000 metabolite structures was developed. Since metabolomic studies only gives a snapshot of the metabolic status of a cell and it is the metabolic fluxes within the cell that are the key determinant of cell activity, an LC-MS based method for the measurement of nitrogen fluxes was evaluated using plants fed with ^{15}N.

As winter approaches trees in temperate zones halt the growth of the vascular cambium and put the meristems into a dormant state. In order to gain a better understanding of the environmental and hormonal regulation of the process of activity-dormancy transitions in aspen trees we performed metabolite profiling of the cambial tissue at different stages over the year.

Finally a metabolic roadmap of wood development in *Populus* was created by combining targeted analysis of plant hormones and amino acids in parallel with high sensitivity untargeted LC-MS and GC-MS analyses. We were able, for the first time in a plant metabolomics study, to describe changes in metabolite profiles following the route of differentiation from cell division to cell death.

Keywords: metabolomics, wood development, mass spectrometry, retention index, QSRR, nitrogen flux

Author's address: Annika Johansson, SLU, Department of Forest genetics and Plant physiology, 901 83 Umeå, Sweden
E-mail: Annika.Johansson@ slu.se

Till den största kämpen av alla.

Contents

List of Publications

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

I Jonsson, P., **Johansson, A. I**., Gullberg, J., Trygg, J., A, J., Grung, B., Marklund, S., Sjostrom, M., Antti, H. & Moritz, T. (2005). High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Analytical Chemistry*, 77, 5635-5642.

II Druart, N.*, **Johansson, A.***, Baba, K., Schrader, J., Sjodin, A., Bhalerao, R. R., Resman, L., Trygg, J., Moritz, T. & Bhalerao, R. P. 2007. Environmental and hormonal regulation of the activity-dormancy cycle in the cambial meristem involves stage-specific modulation of transcriptional and metabolic networks. *Plant Journal*, 50, 557-573.

III **Johansson, A. I**., Gullberg, J., Trygg, J., Linusson, A. & Moritz, T. A retention index prediction database to increase the rate of identification of non-annotated GC-MS mass spectra. (manuscript)

IV **Johansson, A.I.**, Aguetoni Cambui, C., Campbell, C., Hurry, V., Näsholm, T. & Moritz, T. A liquid chromatography – mass spectrometry method for determine ^{15}N- incorporation into amino acids and amines in plants. (manuscript)

V **Johansson, A.I.**, Sundberg, B. & Moritz, T. A metabolite roadmap of the wood forming tissue in *Populus.* (manuscript)

*To be considered joint first authors

Papers I-II are reproduced with the permission of the publishers.

The contribution of Annika Johansson to the papers included in this thesis was as follows:

I Planned the sample preparation, analyzed all samples, and helped with the data analysis. A.J. did part of the writing.

II Did all metabolomics analysis, including planning, data processing and analysis. A.J. did part of the writing

III Did majority of the multivariate modelling, interpretation and planning. A.J. has been responsible for majority of the writing

IV Initiated the idea combining ACQ-derivatisation and LC-MS analysis. A.J. has performed all LC-MS analyses, including data analysis. A.J. did majority of the writing.

V Planned the experiment together with B.S. and T.M. All sample preparation (excluding sampling and sectioning), MS-analysis, data processing and analysis have been performed by A.J. A.J. did majority of the writing.

Abbreviations

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

1 Introduction

Wood is becoming of increasing importance as a renewable resource because of growing demands for our society to be sustainable. Understanding the biology behind the processes of tree growth and wood development is a key to maximizing the use of tree products. The tree industry is one of Sweden's most important economic sectors, producing pulp, paper, wooden goods and wood as a bioenergy source. Between January and June 2012 the Swedish forestry sector exported products with a value of 64 billion Swedish crowns, constituting 10% of Sweden´s total export market (Statistiska Centralbyrån, http://www.scb.se). Future improvements in wood biomass production and the properties of wood fibres, through genetic engineering or by identifying elite lines using early markers for desirable wood characteristics, will be dependent upon a basic understanding of how wood development is controlled by the plant.

The term "metabolome", which was first introduced by Oliver *et al.* (1998), refers to the complete set of small-molecule metabolites (e.g. intermediates in primary metabolic pathways, hormones and secondary metabolites) to be found within a biological sample. The term was coined as a counterpart to the transcriptome (complete set of expressed genes) and the proteome (all proteins present in a sample). Since metabolites can be viewed as down-stream products of gene transcription, protein expression and enzymatic activity the metabolites are believed to reflect activity in cells more accurately than the transcript and protein complements. In 2002, metabolomics was defined as the comprehensive quantitative and qualitative analysis of the whole metabolome under a given set of conditions (Fiehn, 2002). Metabolomics analysis aims to identify and quantify all low molecular weight molecules in a biological sample and thereby link biochemical information to a phenotype (Goodacre *et al.*, 2004; Fiehn, 2002). In the post-genomic era, systems biology approaches that aim to integrate proteomic, transcriptomic, and metabolomic information

to give a holistic picture of a living organism have attracted much interest (for a review of plant systems biology see Weckwerth (2011)).

The size of the metabolomes differs between different species. It has been estimated that the *Arabidopsis thaliana* metabolome consists of approximately 5000 metabolites (Bino *et al.*, 2004), whereas the entire plant kingdom with its enormous number of secondary metabolites is estimated to contain up to 200 000 metabolites (Fiehn, 2002). Metabolomics is currently used in many different applications. Bylesjö and colleagues performed transcriptomic, proteomic and metabolomic analysis in a systems biology approach to studying lignin biosynthesis in hybrid aspen (Bylesjo *et al.*, 2009). A metabolomics study on stem cells found that the metabolic oxidation state was higher in embryonic stem cells compared to differentiated cells (Yanes *et al.*, 2010). A rapidly growing field of research that is based on metabolomics is biomarker discovery for medical applications; for example, Sreekumar and coworkers published a paper about the role of sarcosine in prostate cancer progression (Sreekumar *et al.*, 2009). Another biomarker discovery study was performed by Wang and colleagues, who found, in an LC-MS/MS based metabolomics study, that the level of trimethyl amine oxide (TMAO) has potential for use as a biomarker for cardiovascular disease (Wang *et al.*, 2011). Morreel and co-workers used a metabolomics approach to detect metabolite quantitative trait loci (mQTL) controlling the levels of metabolites in the flavonoid pathway in *Populus* (Morreel *et al.*, 2006).

Metabolomic studies can give a snapshot of the metabolic status of a cell. However it is the metabolic fluxes within the cell that are the key determinant of cell activity. In studies of metabolic fluxes, sometimes referred to as "fluxomics", changes in the metabolome are followed over time to understand, e.g., enzymatic activity (for an overview of fluxomics applications in plants see Kruger and Ratcliffe (2012)).

Although the metabolome can reflect the processes taking place in a biological system in a way that bears a closer relationship to "reality" than the transcriptome and the proteome, a combination of different "omics" approaches will give a more holistic interpretation of the biological system. This requires the joint effort of researchers from many different fields (e.g. biology, chemistry and bio-statistics), each contributing their expertise. The focus of my PhD has been the development of analytical methods and strategies to allow the measurement of at least part of the metabolome in woody tissues of *Populus*. My own expertise has developed to the point where I "know a bit about a lot". The thesis will reflect this; my aim has been to give an overview of the field of wood biology with respect to the metabolome and of the analytical challenges inherent in metabolomics.

2 The vascular cambium and wood development

Wood, or secondary xylem, is the product of xylogenesis, a developmental process in which stem cells are ultimately converted into empty shells with lignified secondary cell walls. These empty cells (or lignified cell walls) are arranged parallel to each other along the trunk of a tree, together forming the wood. Seasonal growth rings (annual rings) are the result of variations in growth rate in trees grown at temperate climate zones. In the early growing season, cells are rapidly dividing and expanding, whereas late in the growing season the rate of cell division is lower and the size of the cells are much narrower.

All plant growth and development originates from meristematic cells. Meristems are populations of small cells of equal size with “embryonic” characteristics. Meristems are self-perpetuating, meaning that not only do they produce the tissues that will form the body of the plant, but they also continuously regenerate themselves. There are three major meristems in a plant such as a tree: the shoot apical meristem (SAM), the root apical meristem (RAM) and the vascular cambium (VC). The shoot and root apical meristems are primary meristems formed during embryogenesis. Secondary meristems, such as the vascular cambium, form later in plant development and originate from the primary meristems. The VC is a lateral meristem found just beneath the bark, which produces phloem outwards and xylem inwards (as reviewed by Lachaud *et al.* (1999)) (Fig 1). It is the ultimate source of secondary xylem, and the rate of cell division in the cambium is the primary determinant of stem and root thickening (Matsumoto-Kitano *et al.*, 2008).

The molecular regulation of wood formation is poorly understood. Within the VC there are cambial initials, or stem cells, that divide to form phloem and xylem mother cells. The phloem and xylem mother cells can in turn undergo several rounds of cell division before they differentiate into the range of cell

types found in wood and bark (Schrader *et al.*, 2004; Lachaud *et al.*, 1999). Most cell divisions in the cambium are periclinal divisions, producing phloem cells outwards and xylem cell inwards. Anticlinal cell division, on the other hand, produces new stem cells (allowing the cambium to continue to surround the stem of the tree as it grows and thickens). Schrader and coworkers made an attempt to clearly identify the cambial initial cells. They hypothesized that the position of the initial cells would be determined by the site of cell division, more specifically by the site of anticlinal divisions. The authors showed that anticlinal divisions occurred only in a very narrow part of the cambium close to the phloem cells, indicating that the position of the initial cells is tightly controlled. They also found that, in the middle of the growing season, the majority of the cell divisions in the cambial zone occurred on the xylem side, explaining the difference in volume between phloem and xylem tissues (Schrader *et al.*, 2004).

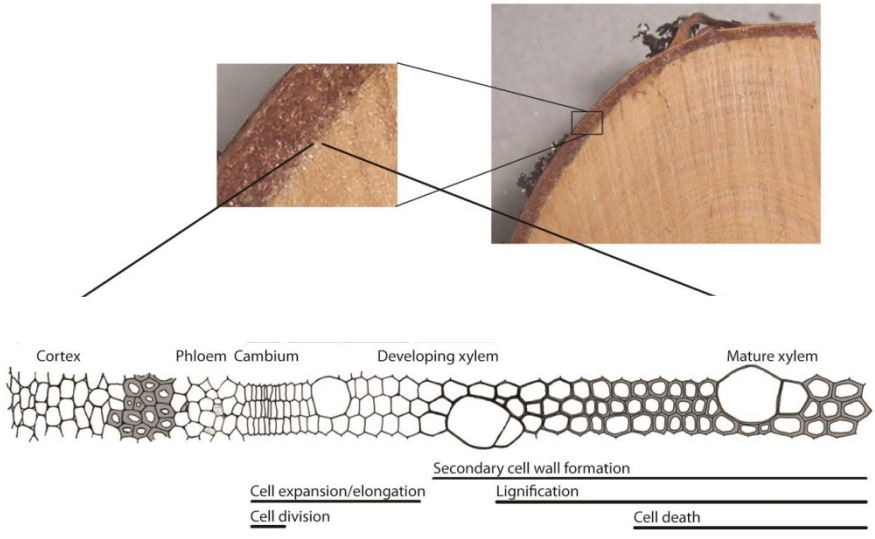

Figure 1. The developmental gradient in the wood-forming zone. Figure credit Ellinor Edvardsson (Edvardsson, 2010). Modified from (Schrader et al., 2004).

The vascular cambium consists of two different types of initial cells, the fusiform initials and the ray initials. The progeny of the fusiform initials differentiate into phloem (sieve tubes) or xylem (fibres or vessel elements). Sieve tubes and vessel elements are key components in the transportation of nutrients and water within a plant. The products of photosynthesis and a variety of other solutes are distributed to different parts of the tree within the sieve tubes in the phloem, whereas the vessel elements are responsible for water and mineral transport in the xylem. The xylem fibres give the tree structural

support. The fusiform initials are elongated cells with large vacuoles, and they have the capacity to divide longitudinally (anticlinal division), thereby ensuring the continuity of the cambium around the stem as the stem gradually increases in width (Lachaud *et al.*, 1999). The cells originating from the ray initials differentiate into ray cells, which elongate radially, stretching from the phloem into the mature xylem. The ray files are essential for translocation of nutrients from the phloem to the xylem and water from the xylem to the phloem. They also serve as storage tissue for substances such as starch, proteins and lipids (Nakaba *et al.*, 2012).

2.1 Xylogenesis - from cell division to cell death

Xylogenesis is a well-ordered process with distinct developmental stages, nicely reviewed by Mellerowicz *et al.* (2001). Following cell divisions in the cambial region, the xylem cells start to differentiate in a process that will end in programmed cell death (PCD) for all cell types except the ray cells, which will stay alive for many years. The first stage of xylogenesis takes place in the zone of cell expansion, where the cells elongate and expand in size to reach their final volume. Cell wall expansion is driven by turgor pressure and is dependent on the flexibility of the primary cell wall of the plant. The primary cell wall contains cellulose, hemicellulose and pectins (Fig 2a). Cellulose is made up of long β(1→4)-D-glucose chains. Since each glucose subunit is rotated 180^{o} with respect to both its neighboring subunits, the true repeating subunit of cellulose is the disaccharide cellobiose (Mellerowicz *et al.*, 2001). The main function of hemicelluloses such as mannans and xylans is to cross-link the cellulose microfibrils by means of hydrogen bonding to increase the rigidity of the cell wall. The hemicellulose polysaccharides contain many different sugar monomers. In addition to glucose, xylose, mannose, galactose, rhamnose, and arabinose can also be included in hemicellulose, as can acidic sugars such as glucuronic acid and galacturonic acid. Pectins are a group of highly diverse cell wall polysaccharides, but they all contain a high proportion of galacturonic acid. Pectins are present in the primary cell wall and the middle lamella but almost absent from the secondary cell wall. The degree of methylation of the pectins influences the plasticity of the primary cell wall (as reviewed in Mellerowicz *et al.* (2001)). All cell wall carbohydrates have UDP-glucose, a nucleotide sugar, as their common precursor. UDP-glucose is produced mainly from sucrose in an enzymatic reaction mediated by sucrose synthase (SuSy), which catalyzes the reaction between sucrose and UTP that forms UDP-glucose and fructose. UDP-glucose is converted to UDP-

glucuronic acid, the precursor of all nucleotide sugars that are substrates for the synthesis of hemicelluloses and pectins (Mellerowicz *et al.*, 2001). Uggla and coworkers studied the role of sugars in the formation of wood, by characterizing the levels of sucrose, fructose and glucose in 30µm sections taken from across the wood-forming zone. Sucrose levels showed a rapid decline from the cambial tissue inwards, whereas the monosaccharaides gradually increased (Uggla *et al.*, 2001).

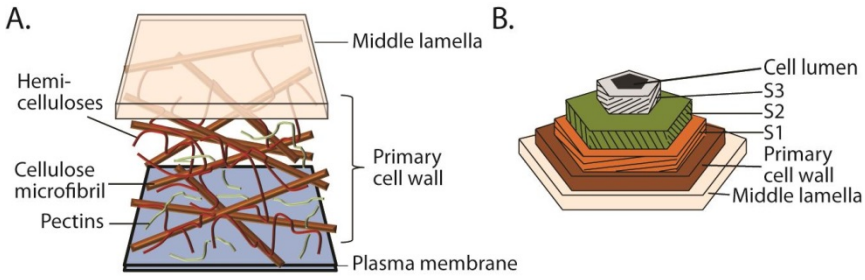

Figure 2. The carbohydrate network in the cell walls. A) A simplified model of the primary cell wall. B) The different layers of the secondary cell wall. Figure credit Emma Hörnblad (Hörnblad, 2012), modified from United States Department of Energy Genome Programs (genomics.energy.gov) and from (Côté, 1967).

When the expansion of the cell is complete, the secondary cell wall starts to form, to give the cell its rigidity and strength. As in the primary cell wall, cellulose and hemicellulose are the major polysaccharides of the secondary cell wall; however, in contrast to the primary wall, pectins are almost absent (Mellerowicz *et al.*, 2001). The secondary cell wall is composed of three different layers, S1, S2 and S3, which are distinguished by the orientation of the cellulose microfibrils, as shown in Fig 2b. The orientations of these microfibrils are fixed by the lignification process at a later stage in wood development. Lignin, which consists of phenolic polymers, gives the plant mechanical support, but also contributes to plant defence mechanisms and to the conductance of sap through lignified vascular elements (Vanholme *et al.*, 2008; Mellerowicz *et al.*, 2001). Three different hydroxycinnamyl alcohols, also known as monolignols, are used to build up the polymeric lignins: p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol. These three monolignols differ only in the degree of methoxylation of their phenyl groups. The biosynthesis of lignin starts with the conversion of phenylalanine into cinnamic acid, and subsequently to p-coumarate, via the phenylpropanoid biosynthesis pathway. The commonly used acronyms H, G and S type lignin refer to the initial letters of the p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) lignin units. Lignins from gymnosperms are composed almost exclusively

of G-units (only minor amounts of H-units are found), whereas angiosperm lignins are composed of G- and S-units (Vanholme *et al.*, 2010). Since the different monolignols interact with different cell wall components, the lignin ratio dramatically affects wood properties (Buchanan et al., 2000). At the end of the lignification process all cells, except the ray cells which stay alive for several years, undergo programmed cell death (PCD). In the process of PCD, cellular content is degraded and the vacuole collapses leaving an empty cell with the intact secondary cell walls forming a shell (Courtois-Moreau *et al.*, 2009). Courtois-Moreau and colleagues found that the programmed cell death of the wood fibres in *Populus* is complete at a distance of around 1000 µm from the cambium, whereas the PCD of vessel elements is finished at an earlier stage, around 500 µm from the cambium (Courtois-Moreau et al., 2009).

2.2 Seasonal changes in the vascular cambium

Plants in temperate zones need a strategy for surviving harsh winter conditions. Annual plants ensure the survival of their genetic material by producing and spreading seeds, but perennial plants and trees require an additional strategy. In the autumn, trees form shields to protect their apical meristems by setting buds over them, and they halt the growth of the vascular cambium, putting the meristematic tissue into dormancy. The transition from active growth in the summer to cessation of growth in the autumn and back to active growth again in spring is termed the activity-dormancy cycle, and is reviewed by Rohde and Bhalerao (2007). The meristematic cells can be in either of two different stages of dormancy: endodormancy or ecodormancy. In the endodormant stage, plant meristems will not resume growth even if the environmental conditions become favourable. In the ecodormant stage, the meristem will become active again if the outer conditions alter and become favourable. In the autumn, trees enter the ecodormant stage, and if external conditions continue to be harsh the trees will enter the endodormant stage. Trees that have entered endodormancy require a period of chilling in order to be released from this state back into ecodormancy, making it possible for the meristem to become active again when spring arrives (Fig 3.) The regulation of the activity-dormancy cycle is crucial for the survival of the tree, since failure to establish dormancy properly makes the tree vulnerable to early frosts in the autumn or cold periods in the spring. On the other hand, if dormancy is established too early in the autumn, or if the release of dormancy occurs too late in the spring, the already short growing season becomes even shorter, and this can make the tree competitively inefficient (Lachaud *et al.*, 1999).

As winter approaches, the temperature drops and the day length shortens; these two external factors are probably the triggers that start the process of growth cessation.

It is known that, in *Populus*, the most crucial factor for initiation of growth cessation in preparation for winter is the point at which the day length drops below a certain threshold (critical day length) (Wareing, 1956). It has been shown that different ecotypes of *Populus* have different critical day length. Trees from more northern latitudes require a longer critical day length than their southern "sisters" in order to initiate growth cessation (Luquez *et al.*, 2008).

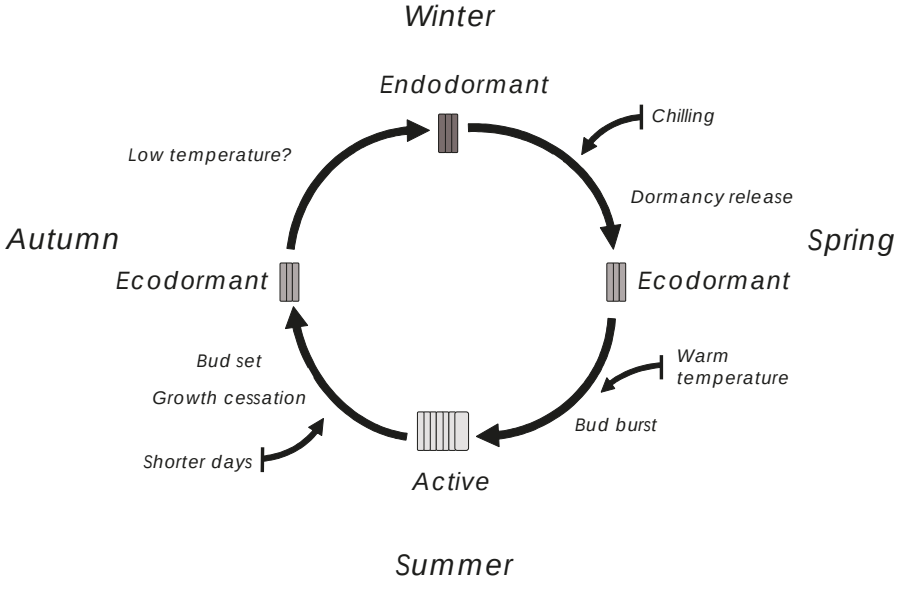

Figure 3. Activity-dormancy cycle in trees.

Simultaneously with the cessation of growth, cold hardiness develops and there is a shift in metabolism towards the accumulation of storage compounds. As cold hardiness develops, the large vacuole in each of the fusiform initials is fragmented into several smaller vacuoles and the cell wall becomes thickened. Storage compounds, including sugars, amino acids, tannins, starch and storage proteins, accumulate; the functions of these compounds are believed to include cryoprotection during the winter and as food reserves for use during reactivation of the cambium in the following spring (Lachaud *et al.*, 1999).

2.3 Plant hormones in wood development

The plant growth regulators (plant hormones) are small molecules, often present in very low amounts, that are key regulators of plant growth and development. For a review of plant hormones their action and regulation see Santner *et al.* (2009). Wood development and the induction and release of dormancy are processes that need to be tightly controlled; plant hormones are known to regulate many steps in these processes. These plant hormones include the most well-known and well-characterized classes: auxin, cytokinins (CKs), gibberellins (GAs) and abscisic acid (ABA). These plant hormones can act independently of each other, but often it is their interactions that bring about the observed physiological effects (Santner *et al.*, 2009).

Auxin was discovered as a growth factor in the 1930s and since then it has been shown to be involved in many different physiological events, such as apical dominance, shoot elongation and root initiation. The most common form of naturally occurring auxin, indole-3-acetic acid (IAA), is primarily synthesized in the buds and developing leaves and it is actively transported by a basipetal polar transport system to other parts of the plant (Buchanan *et al.*, 2000). Sundberg and colleagues have shown that in *Populus* and *Pinus silvestrus* there is an auxin gradient over the zone of wood development, with a peak in the vascular cambium (Tuominen *et al.*, 1997; Uggla *et al.*, 1996). Auxin is considered to be an important regulator of xylem development;, among its other influences, auxin appears to give the cambium positional information (Uggla *et al.*, 2001; 1998). The IAA gradient within the cambial zone is probably caused by basipetal transport of IAA rather than by de novo biosynthesis of the compound (Uggla *et al.*, 1996). There is no evidence of changes in the concentration of IAA in the vascular cambium during the induction of dormancy. However there are reports indicating that responsiveness to IAA within the cambial region is affected by short days (Baba *et al.*, 2011).

Gibberellins (GA), another class of plant hormones, involved in many plant growth processes such as, seed development, organ elongation and the control of flowering time (Santner *et al.*, 2009), may also have a role in cambial regulation. Eriksson and co-workers showed that over-expression of GA20 oxidase increases both the stem diameter and the height of transgenic *Populus* trees (Eriksson *et al.*, 2000). A more detailed study of the role of GAs in the wood-forming zone showed a gradient of active GAs in the cambial region (Israelsson *et al.*, 2005). The highest levels of active GAs (GA_4 & GA_1) were found in the cell expansion zone, whereas GA precursors and inactive GAs showed two distinct peaks, flanking the active GA peak on both sides. Based

on their results, Israelsson *et al.* (2005) suggested that the main effect of GA is on the elongation of xylem fibres rather than on radial expansion.

Cytokinins is a class of plant hormones are known to promote plant cell division, and are involved in various plant growth processes such as: germination and senescence (Santner *et al.*, 2009). Nieminen and colleagues measured the expression of the cytokinin receptor family over the cambial region and found that the peak in expression levels coincided with the expression of a marker gene for cambial cell identity. They also studied the effect of reduced cytokinin levels in the vascular cambium of transgenic *Populus* by expressing a cytokinin degrading enzyme from *Arabidopsis* in the cambial zone (Nieminen *et al.*, 2008). They showed that the number of cells in the cambial layer was reduced in the transgenic lines, indicating that the level of active cytokinins has an impact on the rate of cell division. A decrease in stem thickening in response to decreased cytokinin levels was also demonstrated in *Arabidopsis* (Matsumoto-Kitano *et al.*, 2008). This study showed that the decrease in secondary thickening observed in mutants defective in cytokinin biosynthesis could be restored, in a dose-dependent manner, upon addition of exogenous cytokinins. As in *Populus*, an increase in cell number in the cambial region as well as in the phloem and xylem regions was observed. The application of cytokinin did not, however, affect the size of the cambial cells. These results indicate that cytokinins are major hormonal regulators of the cambium, acting by controlling the rate of cell division in this tissue.

Abscisic acid (ABA) is a growth regulator which is often found to act as a growth inhibitor in response to various biotic and abiotic stresses (as reviewed in Raghavendra *et al.* (2010)). The role of ABA in wood development is unclear. However, it is thought to play a role in the regulation of dormancy, both in seeds, where ABA acts to maintain dormancy whereas GAs promote germination, and in buds, where during development of dormancy in the autumn it has been shown that ABA levels increase after a few weeks in short days (reviewed in Popko *et al.* (2010)).

3 Analytical aspects of metabolomics

There are a number of obstacles that need to be considered when performing metabolomics analysis. Firstly, in comparison with DNA and RNA, each of which has only four nucleotide bases, and the proteome, which is made up of a defined number of amino acids, there are no common building blocks in the metabolome. It consists of compounds with widely varying physicochemical properties, making chemical analysis of the metabolome very challenging. Another major analytical issue is that there is often a large variation in concentrations of metabolites, since metabolites can be present within a cell at concentrations ranging from molar down to nanomolar or even lower (Hall, 2011).

There currently exists no method that can accurately detect and measure the entire metabolome in a single analysis. Several different methods need to be combined in order to cover as large proportion of the metabolome as possible. The field has developed several different analytical approaches, and both targeted and untargeted screening methods are now included under the metabolomics umbrella; for an overview of the main metabolomics approaches, see Table 1. The methodology of choice is dependent upon the biological question being asked. If, for example, the requirement is to identify differences in the amino acid content between a wild type organism and a mutant, it is better to choose a targeted technique than an untargeted profiling method. The biological question and the experimental design are the corner-stones on which a metabolomics study should be built up, and the experimental design will be the factor with the greatest influence on the outcome. One important issue that needs to be addressed concerns the dynamics of the metabolome, which changes from second to second so that rapid sampling is necessary in order to quench metabolic activity. In the field of plant metabolomics, this is often achieved by rapid freezing in liquid nitrogen, followed by storage at -80^{o}C to minimize enzyme activity and degradation of metabolites.

Table 1. *Different metabolomics approaches, modified from (Hall, 2011) and (Goodacre et al., 2004)*

Metabolome	The complete set of small molecules, generally considered to be <1500 Da, present in a specific biological sample.
Metabolomics	The "unbiased" comprehensive quantitative and qualitative analysis of the whole metabolome under a given set of conditions.
Metabolic fingerprinting	Screening of the metabolic composition of samples, usually carried out with high-throughput techniques and involving many samples. The goal is sample comparison and classification; multivariate statistical methods such as PCA and hierarchical clustering are commonly used. Identification of metabolites is of secondary interest and therefore seldom performed.
Metabolic profiling (metabolite profiling)	The quantitative and qualitative analysis of part of the metabolome or specific groups of metabolites. In contrast to metabolic fingerprinting, metabolic profiling aims at identifying and quantifying at least some of the metabolites. Metabolic profiling can be sub-divided into two different areas. Untargeted metabolic profiling tries to detect as many metabolites as possible without focusing on any specific compounds, and is often carried out using untargeted LC-MS or GC-MS analysis. Targeted metabolic profiling methods are designed to detect 100-200 specific metabolites.
Metabonomics	Measurement of the fingerprints of biochemical perturbations caused by disease, drugs and toxins.
Targeted analysis	Methods where the extraction, separation and detection is optimized to measure a selected group of metabolites that have similar properties, such as amino acids or metabolites in a specific pathway (e.g. flavonoids). Targeted methods are usually fully quantitative.
Lipidomics	Specific characterization of lipid species.
Fluxomics	Methods to follow fluxes in the metabolome in order to understand enzymatic activity. Often carried out by means of labelling experiments using stable isotopes.

Since metabolomics studies are in most cases based on comparisons between samples, the methods used for chemical analysis need to be robust and reproducible with high precision. Chemical analysis of the metabolome can be performed using a range of different equipment configurations, often including 'hyphenated techniques' in which a chromatographic separation (e.g. GC, HPLC, UHPLC, CE) precedes the detection step (e.g. NMR or MS). NMR has the advantage of offering high throughput analysis requiring no, or little, sample preparation, and compared to mass spectrometry NMR is a highly quantitative method of detection. Despite recent improvements in instrumentation, NMR suffers from low sensitivity and large sample amounts are therefore needed compared to those required for analysis by mass

spectrometry. NMR based metabolomics will not be discussed further in this thesis, but for a review of plant metabolomics based on NMR, see Kim *et al.* (2011).

In addition to an appropriate experimental design, the key to a successful metabolomics-based study is to keep the underlying biological question constantly in mind throughout the investigation (Fig. 4).

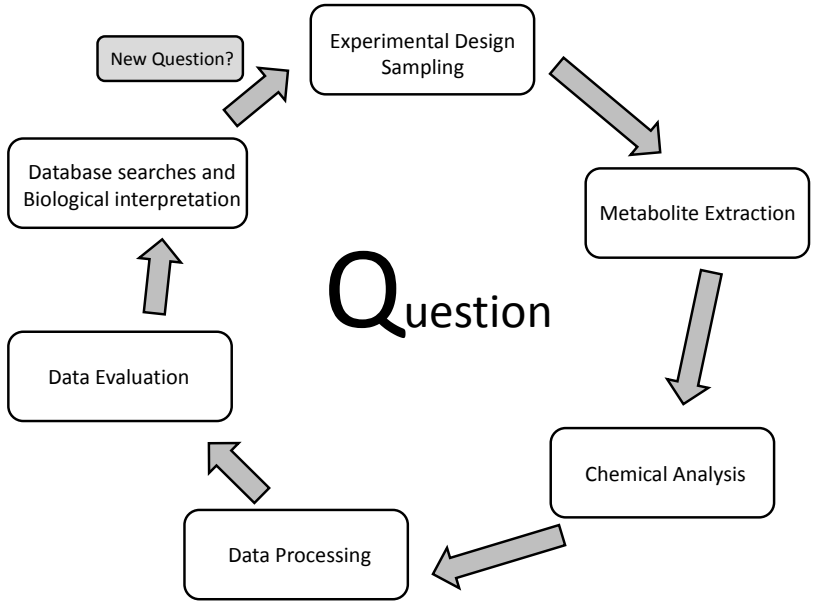

Figure 4. A schematic overview of a metabolomics study.

3.1 Metabolite extraction and purification

The first stage in analysis is sample preparation, which includes extraction, and if necessary purification, of metabolites from the tissue of interest. A reliable extraction protocol is simple and robust, with as few steps as possible so as to avoid introducing errors from, e.g., pipetting, and minimize metabolite losses. The aim of extraction is to remove salts, proteins and tissue debris. Most extraction protocols include a protein precipitation step (Yanes *et al.*, 2010; Gullberg *et al.*, 2004; Fiehn *et al.*, 2000b), since residual protein in a sample will interfere with MS analysis. However, some metabolites strongly bound to proteins will also precipitate and therefore be lost. All extraction protocols will cause some metabolite losses and the solvent used for extraction can never be optimal for all metabolites. Targeted approaches, which often focus on

metabolites with similar physicochemical properties, can be optimized with regards to extraction and purification by solid phase extraction (SPE). These methods often include the addition of stable isotope internal standards, allowing the efficiency of extraction and purification to be monitored. In untargeted approaches the situation is different, since the physicochemical properties of the compounds are very diverse and an extract can contain a large number of unknown metabolites, so stable isotope internal standards covering the entire metabolome are simply not available. The solution is to compromise, by adopting a method that can extract as many different metabolite groups as possible as well as possible. Gullberg and co-workers used a design of experiment (DOE) approach to optimize extraction and derivatization prior to analysis by GC-MS. The authors selected a number of stable isotope internal standards to cover different compound groups and determined extraction and derivatization efficiencies for the different internal standards using a range of extraction parameters (Gullberg *et al.*, 2004).
The next step is chemical analysis. The number of choices to be made within this step is huge and this has been the focus of my PhD. I will therefore try to describe the different techniques in more detail.

3.2 Chemical analysis

> "Chromatography is essential a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction" (Poole, 2003, p. 2)

In gas chromatography (GC), the mobile phase is a gas; in liquid chromatography (LC), the mobile phase is a liquid. Several different types of stationary phase are available for both GC and LC, and by changing the stationary phase it is possible to alter the selectivity of the chromatographic separation. In chromatography there is constant competition between the mobile phase and the stationary phase to "hold" the analyte. The retention time is the total time an analyte spends in the mobile phase and in the stationary phase. All analytes spend the same amount of time in the mobile phase, corresponding to the time it takes for a non-retained analyte to exit the column and reach the detector. This time is called the dead time or hold-up volume of the chromatographic system.

The sharpeness of a peak relative to its retention time is a measure of the efficiency of the system, (plate count, calculated as N). Band broadening (wider chromatographic peaks) will reduce the efficiency of the

chromatographic system, whereas the sharper the peak for a specific analyte, the better the signal-to-noise ratio and thereby increased sensitivity.

Plate Count N: Determines the efficiency of the separation (when working at isocratic conditions) Calculated by:

$$N = 5.54x\left(\frac{t_R}{w_{50\%}}\right)^2$$

Where t_R is retention time of analyte and $w_{50\%}$ is the peak width at half peak height.

Height Equivalent to Theoretical Plate (HETP): is a measure of the column's chromatographic efficiency (in micrometer) calculated from the following equation:

$$HEPT = \left({}^{L}/_{N}\right)$$

Where L is column length and N is the plate count. The efficiency of the chromatographic system increases with decreasing HEPT.

Resolution Factor, Rs: describes the difference in retention times between two peaks relative to their width. The retention factor can be calculated empirically as follows:

$$Rs = \frac{t_{R(2)} - t_{R(1)}}{0.5(w_2 + w_1)}$$

Where t_R is the retention time of peaks 1 and 2 respectively and w is their respective peak width at the peak baseline. The resolution is a critical factor when working with complex matrixes.

3.2.1 Gas Chromatography

Gas chromatography (GC) is a very widely used analytical method. It is cheap, robust, simple and fast, and when capillary columns are used it is a technique that can yield high-resolution separation. In gas chromatography, a portion of a liquid sample is injected into an inlet liner which is held at a temperature that will vaporize the analytes/metabolites. Different injection methods can be used to introduce the sample onto the column. Two common methods are split and splitless injection. In split injection, only a part of the vaporized sample (e.g. 10%) is introduced onto the column. When the entire vaporized sample is introduced onto the column (splitless mode), more of the analytes present at low abundance can be analysed, hence the splitless mode is used in most

metabolomics applications. However, in some experiments where high concentrations of e.g. sugars are present, the analysis can be greatly improved by using split injection. The vaporized analytes enter the front end of the capillary column and are pushed through the column by a continuous flow of a carrier gas (helium is commonly used). Analytes with a boiling point higher than the temperature of the GC oven will initially condense on the column's stationary phase; as the temperature in the GC oven is gradually increased the condensed analytes will vaporize again and be carried through the column by the carrier gas.

In metabolomics GC/MS applications the column is normally a fused silica capillary column with a diameter in the sub-mm range (<0.5 mm) and a length varying between 10 and 60 m. Since metabolomics studies include a large number of samples, high-throughput analytical methods are required. In our metabolomics platform we use sample run-times varying from 15 minutes to 30 minutes on our GC-TOF systems. Fast GC methods result in narrow peak widths, typically of 0.5-1.5 seconds, requiring a detection system with a high acquisition rate, e.g. a time-of-flight mass spectrometer. There are many different types of stationary phases available for capillary columns, allowing different analyte-column interactions. As an example, a methyl coated fused silica capillary column will separate analytes mainly according to their boiling points, whereas a DB-50 column with 50% phenyl groups will allow for more stationary phase-analyte interactions. In the field of plant metabolomics, the DB-5 column, where 95% of the silica is coated with methyl groups and the remaining 5% is coated with a phenyl group, is by far the most commonly used stationary phase (Schauer *et al.*, 2005). The retention time in a DB-5 type column is almost exclusively dependent on the boiling point of the (derivatized) analyte. Since volatile compounds are a requirement for analysis by GC whereas many of the metabolites within biological tissues are non-volatile, derivatization of the metabolites is usually necessary prior to GC/MS analysis. In plant metabolomics, the most commonly used derivatization protocol is methoxyamination followed by silylation (Gullberg *et al.*, 2004; Fiehn *et al.*, 2000a) (Fig. 5a). Derivatization will allow non-volatile metabolites to be analysed on GC/MS; however, the derivatization process also introduces artefacts into the chromatogram and these can cause problems during data evaluation. Experimental blanks should therefore be included in each experiment in order to distinguish putative metabolites from background peaks. Comparison of the results of different GC-based analyses is facilitated by the use of standardized retention index systems such as the Kovàts retention index. By converting the retention time of a compound into a retention index it is possible to compare the retention of a metabolite analysed on a 30 m column

with that of the same metabolite analysed on a 10 m column in a different lab, as long as the stationary phase of the GC-column is the same (Schauer *et al.*, 2005), Fig. 5b.

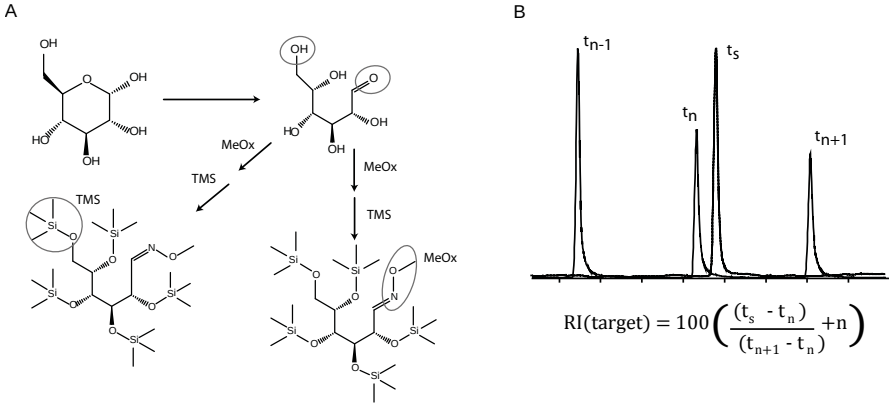

Figure 5. A) By converting the carbonyl groups into oximes (with a limited rotation along the C=N bond) before silylation, the number of tautomeric forms of reducing sugars will be reduced to two instead of the five which would be possible if only silylation were used. B) The Kovàts retention index (RI) utilizes the elution times for a series of n-alkanes to convert the retention times of analytes to retention indices. t_s *= retention time of analyte,* t_n *= retention time of closest n-alkane eluting before the analyte, n= number of carbons in that n-alkane.*

3.2.2 Liquid chromatography

In liquid chromatography, the mobile phase is commonly a combination of water and organic solvents, and depending on the choice of the stationary phase, different analytes will be retained differently on the column. Liquid chromatography does not require derivatization of the metabolites prior to analysis, thereby reducing the number of sample preparation steps necessary before analysis. This will reduce variation due to sample preparation and the number of artifacts present in the sample. Reversed phase liquid chromatography, in which separation is based on hydrophobic interactions between the analyte and the stationary phase (e.g. C18-columns), is an LC configuration used commonly in the field of metabolomics. The packing material used is crucial for the resolution of a LC column. In principal, the smaller the particle size the better the resolution. Smaller particles give a more densely packed column that improves separation but also introduces higher backpressures from the mobile phase into the system. The introduction of UPLC™ or ultra-high pressure liquid chromatography (UHPLC) in 2004 led to a reduction in particle size from 3.5 µm to sub-2 µm particles and necessitated

the introduction of pump systems that could withstand the resulting high backpressures (Swartz, 2005). The reduced particle size resulted in decreased HEPT, greater consistency in retention time and shorter analysis times, in comparison with traditional HPLC (high pressure liquid chromatography). However, as chromatographic resolution is improved, resulting in half height peak widths of less than one second, the acquisition rate of the detector must be increased so as not to lose the potential gain in resolution (Swartz, 2005). The high acquisition rate provided by time-of-flight (TOF) mass spectrometers (see chapter on mass spectrometry) is one of the reasons why these instruments have gained in popularity.

Although many different metabolites can be analyzed using reversed phase liquid chromatography, this method cannot be used for highly polar metabolites, since such metabolites are not retained on a C18 column. Non-retained metabolites will exit the column with the injection front, causing problems with, e.g., ion suppression (see chapter on mass spectrometry). A combination of different LC stationary phases is desirable in order to cover both polar and non-polar metabolites. The retention mechanism in hydrophilic interaction chromatography (HILIC) is complementary to that in reversed phase chromatography (see Hemstrom and Irgum (2006) for a review on HILIC). Nordström and co-workers have shown that by combining reversed phase chromatography and HILIC in positive and negative ionization mode, the number of metabolites from cancer cell lines that can be analyzed by LC-MS is dramatically increased. Only approximately 15% of the metabolites detected overall could be detected in both modes. (Nordström, unpublished).

Although derivatization is not essential for LC, it can sometimes improve chromatographic retention. By coupling propionyl groups to the hydroxyl groups of cytokinins, the hydrophobicity of the cytokinins was increased sufficiently for them to be analyzed on a C18 column. In addition, when the hydrophobicity of the cytokinins was increased, the sensitivity with which they could be detected in electrospray ionization mass spectrometry was increased (Nordstrom *et al.*, 2004). Derivatization prior to LC analysis can also be necessary to avoid degradation of unstable metabolites (Novák *et al.*, 2012).

3.2.3 Mass spectrometry

A mass spectrometer is essentially a sophisticated balance capable of measuring the weight of molecules. It does this by separating gas-phase ions on the basis of their mass-to-charge ratios. Separation of molecules by mass spectrometry can be divided in three stages: ion production (within the ion source), mass separation (in the mass analyzer) and mass detection (at the detector).

Ionization techniques

The analytes must be ionized and transferred into the gas-phase in the ion source before their mass-to-charge ratios can be measured in the mass analyzer. Different ionization techniques are used in conjunction with different separation techniques and for different purposes.

In GC/MS, the most commonly used ionization technique is electron (impact) ionization (EI). In EI, the gas-phase analytes are bombarded with electrons, producing positively charged radical ions. These positively charged ions are pushed out from the ion source and into the mass analyzer. EI is a hard ionization technique that causes fragmentation of the analyte (Herbert & Johnstone, 2003). The fragmentation pattern is highly repeatable and dependent on molecular structure, and the resulting mass spectra can be used for identification purposes. Libraries of mass spectra for standards of known chemical composition are available to facilitate metabolite identification. Taking the mass spectra of an analyte together with its retention index from the gas chromatography, the identity of the analyte can be determined. GC/MS and EI is in this sense a very straightforward analysis technique for use in metabolomics, although deconvolution of overlapping peaks is often difficult (see data processing chapter). Metabolites analyzed by GC/MS (using EI) can often be identified, if they are present in a mass spectral library. However, fragmentation of the analytes can be a drawback in the case of unknown analytes that are not to be found in any library. Since EI produces hardly any unfragmented molecular ions, the identification of unknown compounds becomes more difficult if no information about their elemental composition is available. One way to overcome this problem can be to reanalyze the sample using a soft ionization technique, e.g. chemical ionization. In chemical ionization the ionization chamber is filled with a reagent gas, commonly CH_4. As in EI, an electron filament is used, but in CI, because of the use of CH_4, the positive radicals formed are predominantly $CH_4^{\bullet+}$, and after one such radical collides with another CH_4 molecule, a carbonium ion, CH_5^+, is formed. When one of these carbonium ions collides with an analyte (M), there will be an exchange of protons, forming a protonated molecular ion $[M+H]^+$. In contrast to EI, CI spectra will be dominated by molecular ions and little or no fragmentation will occur (Herbert & Johnstone, 2003). A combination of the two ionization techniques can therefore be useful for de novo identifications in GC/MS based analysis.

If mass spectrometry is used in conjunction with liquid chromatography, the liquid output from the LC system must be transferred into gas phase and ionized before entrance into the mass analyzer. In LC/MS today the most commonly used ionization techniques are soft ionization methods carried out

under atmospheric pressure, which produce mainly molecular ions, in the form of $[M+H]^+$ and $[M-H]^-$. Ninety percent of all LC/MS based metabolomics applications utilize electrospray ionization (ESI) (as reviewed by Forcisi *et al.* (2013). In general, ESI is efficient for polar, thermally labile and nonvolatile compounds such as peptides, proteins and small polar metabolites. Since ionization is based upon the addition (positive) or subtraction (negative) of a proton, the molecule must have the capacity to take up or give away a proton. The ionization efficiency increases if the analyte also possess some hydrophobic part (Leito *et al.*, 2011). ESI in general is quite prone to generating multiply charged ions, especially when it comes to larger molecules such as peptides and proteins. This allows the analysis of large molecules, up to 150 000 Da, even though the mass range of a typical TOF is around 3000 m/z, since separation is based on the mass-to-charge ratio. However, as metabolites are relatively small molecules, the resulting ions are predominantly singly charged. To cover as many metabolites as possible a sample should be analyzed in both positive and negative mode.

In ESI, the liquid from the LC outlet (or the sample matrix itself, if direct injection is used) is passed through a short length of stainless steel capillary tubing. At the tip of the capillary a high positive or negative electrical potential is applied, typically 2-5 kV. The electrical potential applied will cause charge separation of the ions within the solution and an accumulation of charged ions at the liquid surface on the tip (Fig. 6). Multiply charged droplets with diameters of a few micrometers will detach from the liquid cone (known as a Taylor cone) formed at the capillary tip. As the droplets evaporate, their charge-to-volume ratio increases and reaches the point, the Rayleigh limit, at which the repulsion of equal charges within the droplet is sufficient to overcome the surface tension holding it together. At the Rayleigh limit the droplets will undergo coulombic explosions forming a spray which is emitted from the Taylor cone. After subsequent evaporation and disintegration of droplets, fully desolvated ions are formed, and these can enter the orifice of the mass analyzer. The exact process leading to the formation of a gas-phase ion is unclear, but there are two major theories to account for it: the desorption theory argues that the analyte ion is “pushed out” from the droplet by coulombic explosion whereas the charge residue mechanism postulates that the last step in gas-phase ion formation consists solely of evaporation of the remaining solvent

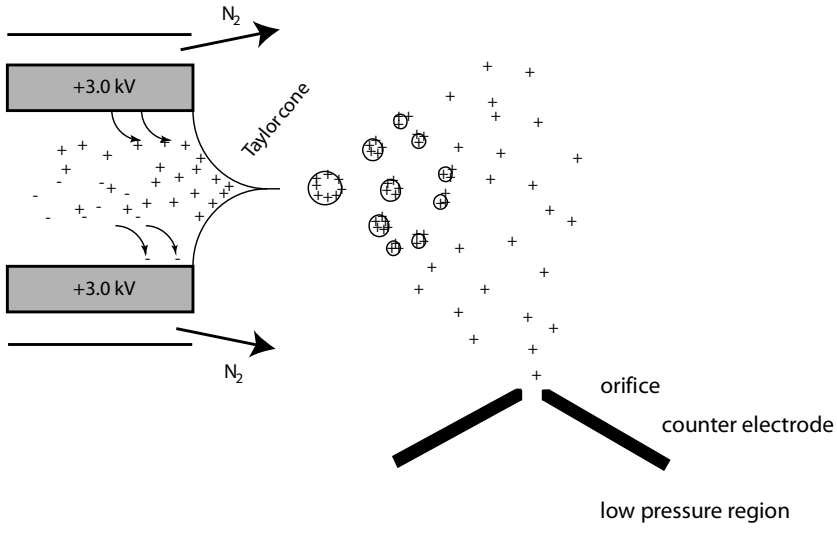

Figure 6. The principle of ESI. A liquid cone, known as a Taylor cone, will form at the capillary tip. Multiply charged droplets with diameters of a few micrometers will detach from this liquid cone. As the droplets drift towards the counter electrode, positioned close to the capillary tip, the solvent within the droplets will evaporate resulting in an increase in charge-to-volume ratio, followed by coulombic explosions.

(Herbert & Johnstone, 2003). The nebulization process is facilitated, in particular when it comes to higher flow rates, by the use of a desolvation or nebulizing gas on the outside of the capillary. To further promote evaporation, a drying gas may be used, and the temperature within the ion source is kept high (in our analyses, 100-150°C). The orientation of the needle with respect to the orifice of the mass analyzer is another important factor. The earliest ESI setups used on-axis orientation, but this led to clogging of the sampling orifice and the entry of neutral compounds into the MS causing chemical noise. More recent ESI sources have a sampling orifice positioned off-axis from the needle. Off-axis orientation decreases the amount of neutral analytes and ions of opposite charge that can enter the orifice, and hence reduces both the risk of contamination of the orifice and the phenomenon of ion suppression (Holcapek *et al.*, 2004).

Ion suppression is a major concern when working with ESI. It is a form of matrix effect that occurs mainly as a result of competition for ionization. It may affect the detection limits, reproducibility, precision and/or accuracy of an analysis. Ion suppression occurs during the ionization of analytes, hence the sensitivity and MS/MS capability of an instrument are immaterial, since the formation of ions occurs before the mass analysis stage. There are two main principles behind ion suppression: a suppressing compound can out-compete

an analyte of interest in competition for charges during the ionization process, or it can affect the viscosity and surface tension of the liquid, thereby affecting the number of coulombic explosions (Jessome & Volmer, 2006). One area in a chromatogram that is particularly likely to suffer from ion suppression is the elution front, in which a large amount of un-retained compounds elutes. One way to minimize the risk of ion suppression is therefore to avoid eluting analytes in this region of the chromatogram. Another efficient way of minimizing ion suppression effects is to use SPE to reduce the complexity of the sample matrix. Since in many cases targeted methods involve purification on SPE, they are less likely to suffer from ion suppression, but ion suppression effects can still occur. One way to reduce the error introduced by ion suppression in the case of targeted methods is to add a stable isotope form of the analyte of interest to the sample to act as an internal standard. This internal standard can be used to normalize the effects of ion suppression, but the amount of the internal standard added is crucial. Too large an amount of the standard might itself cause ion suppression of the analyte and interfere with the analysis. Other ways to reduce ion suppression include switching polarity, switching ionization method (APCI/APPI; see below), reducing the flow rate and use nanospray ionization. Nanospray ionization has proved to be less sensitive to ion suppression than other techniques and it also generally gives a higher sensitivity for detection of the analytes (Gangl *et al.*, 2001). Since the initial droplets formed by the nanospray technique are smaller and more highly charged, the number of coulombic explosions is reduced and less solvent needs to be evaporated. It is also important to bear in mind that in some cases the cause of the ion suppression cannot be detected in the mass spectra, since some of the compounds causing ion suppression do not themselves ionize.

Why is the phenomenon of ion suppression important in metabolomics analysis? Metabolomics extraction protocols aim, as previously stated, to keep the sample preparation procedure as simple as possible, so as to minimize introduction of analytical errors. The way of minimizing the risk of ion suppression effect is the opposite. ESI-LC/MS-based metabolomics screening or profiling experiments are therefore highly likely to suffer from ion suppression, and this factor needs to be taken into account when evaluating results.

In ESI, ionization of the analytes and transduction into the gas phase occur simultaneously. In atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI), the analytes are first transformed into the gas-phase and then ionized. In APCI and APPI the mobile phase is introduced into the ion source through a heated capillary (250-500^{o}C), producing a mist of small droplets. The spray thus formed is then, in the case

of APCI, allowed to flow past a corona discharge needle, initiating the ionization reactions. As in the case of CI, the analyte ion is formed through a cascade of reactions producing $[S+H]^+$ ions (S= solvent) that collide with the analyte, and finally protonated analyte ions ($[M+H]^+$) are formed through gas-phase ion-molecule reactions: $[S+H]^+ + M \rightarrow S^+ [M+H]^+$. APPI works in a similar way to APCI, but the analyte ions are produced by photons (either directly or through reactions with solvents as in APCI) from a discharge lamp instead of a corona discharge pin (Herbert & Johnstone, 2003). In comparison with ESI, APPI and APCI can ionize compounds that are more non-polar, and ideally these techniques should be run in parallel with ESI in order to cover as large proportion of the metabolome as possible. For instance, carotenoids can be successfully ionized using APCI but not with ESI (Kopec *et al.*, 2013).

Quadrupole and TOF mass analyzers

Before describing the different types of mass analyzers I should like to explain a few important terms in the field of mass spectrometry (Hart-Smith & Blanksby, 2012) modified):

Resolution: The resolution, often calculated using Full Width Half Maximum (FWHM), is the resolving power of the mass analyzer, calculated by:

$$Resolution = \frac{M}{\Delta M}$$

Where: M = mass measured at peak apex, ΔM = mass width of the peak at 50% peak height.

At a too low resolution, a mass analyzer cannot distinguish two ions with very similar mass-to-charge ratios. This will result in the average value for the two ions being output, and hence calculation of elemental composition and identification of metabolites will be more difficult.

Mass accuracy: The mass accuracy of a mass analyzer describes how well the analyzer measures mass. Mass accuracy is the difference between the experimental and theoretical mass of a given substance. It is often expressed in mDa:

$$\text{Mass accuracy (mDa)} = M_{measured} - M_{calculated}$$

or in parts per million (ppm), calculated as follows:

$$\text{Mass accuracy (ppm)} = \frac{(M_{measured} - M_{calculated})}{M_{calculated}} * 1\,000\,000$$

Sensitivity: Sensitivity refers to the smallest amount of a substance that can be analyzed and is visible in the mass spectra. Sensitivity is highly compound-dependent, and for any given compound the difference in sensitivity between different mass spectrometers can be orders of magnitude.

Duty cycle (scan time): This is the time from ion production to mass detection and calculation of mass spectra. The number of possible duty cycles or scans per second, the sampling rate, is usually reported in Hz. A good rule of thumb is that to obtain accurate quantification of a chromatographic peak, the number of data points across the peak should be at least ten (Poole, 2003). If the sampling rate of a mass spectrometer is too low, this will reduce the resolution that can be obtained in a chromatographic separation. On the other hand, reducing the scan time will allow fewer ions to reach the detector, causing a reduction in sensitivity.

Linear Dynamic Range: This is the range over which the ion signal is linearly related to analyte concentration. The dynamic range is a highly important aspect of mass spectrometry based metabolomics, especially in metabolomics studies where the aim is to quantify the "entire" metabolome.

The effective separation of mass-to-charge ratio in the mass analyzer is dependent upon the maintenance of a vacuum to allow the ions to move within the mass separator without any resistance. The vacuum within a mass analyzer is maintained by two different types of pumps: pre-vacuum or rough pumps, working in the high pressure region, and turbo pumps that work in the low pressure or vacuum region, where the actual mass separation takes place. In GC/MS, the ion source is placed within the vacuum or low pressure region, hence the ions produced can be directly transferred to the mass separator. In LC/MS, in which the ionization occurs at atmospheric pressure, the ions must travel through a gradient of pressure reduction from the orifice to the mass separator. As an example, in the Agilent QTOF instrument used in my studies the pressure is reduced from atmospheric pressure (760 Torr) at the ion source, to $3*10^{-6}$ Torr at the quadrupole and $\sim 8*10^{-8}$ Torr in the TOF flight tunnel. As the ions travel down the pressure gradient they are also focused in space with the help of a series of ion guides and octapoles/hexapoles. For a schematic overview of a QTOF see Fig. 7A.

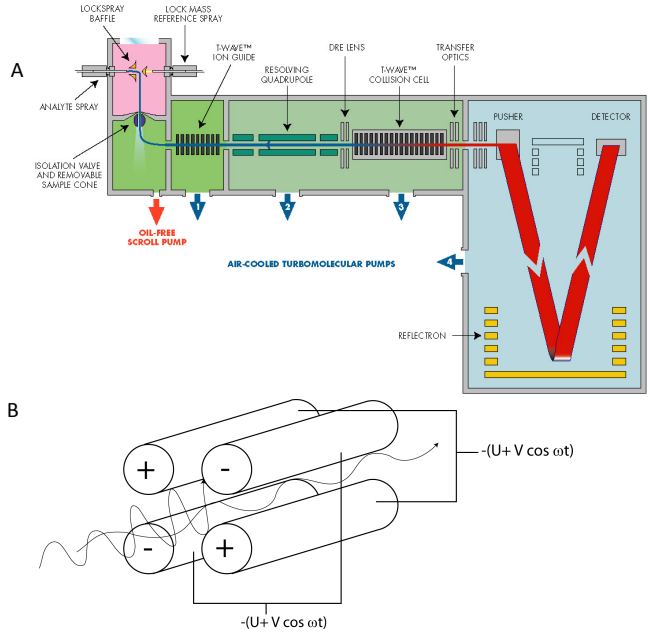

Figure 7. Schematic overview of a QTOF and the mass separation principle of a quadrupole. A) To compensate for small differences in position and kinetic energy, an electronic mirror is used to reflect the ions before they reach the TOF-detector. Ions of the same m/z ratio but with a slightly greater kinetic energy will travel further before being reflected by the electronic reflector and their travelling distance will be slightly longer, hence they will reach the detector at the same time as the corresponding ions with less kinetic energy (as reviewed by (Mamyrin, 2001). Diagram from Waters, Manchester, UK with permission. B) Ions with stable trajectories (those with the "correct" amplitude of oscillation) will pass through the quadrupole and reach the detector, whereas ions with unstable trajectories (those with incorrect amplitude) will be neutralized by striking the quadrupole rods.

Quadrupole

In a quadrupole mass analyzer, four cylindrical metal rods are placed in parallel in such a way that a cavity is formed along the central axis of the rods, as shown in Fig. 7A. As the ions enter the cavity they will start to oscillate and the trajectory of each specific m/z ratio will be determined by the voltages (U, V) and the frequency (ω) of the alternating RF potential. By varying U, V and ω, ions with different m/z ratios will develop stable ion trajectories and pass through the quadrupole (Hart-Smith & Blanksby, 2012; Herbert & Johnstone, 2003). A quadrupole mass analyzer can be run either in scan mode, in which a specified range of m/z values is scanned, allowing ions with different m/z to pass through the quadrupole, or in selected-ion mode (SIM), where the quadrupole is locked to one or a few specific m/z ratios. Quadrupoles are low

resolution instruments, typically with unit resolution, and their mass accuracy is poor. However, the linear dynamic range and the sensitivity of a quadrupole are generally good, and given their small size and relatively low cost, single quadrupole mass spectrometers are among the most commonly used types of mass spectrometer today. In metabolomics research single quadrupole MS is rarely used, due to the poor resolution and mass accuracy. However, quadrupole mass analyzers are used extensively in MS/MS instruments described later.

Time of flight

The basic function of a time-of-flight mass analyzer is to measure the time it takes for an ion of a specific mass-to-charge ratio to fly through a tunnel under low vacuum conditions. Ions fly through the tunnel at different speeds depending on their m/z ratios. The TOF mass analyzers used today are orthogonal TOFs; that is, ions enter the flight tube at the side and are pushed in an orthogonal direction by a pulsed voltage (Fig. 7A). The time it takes from the pusher pulse until the ion reaches the detector is recorded. Once all ions within the specified mass range have reached the detector, another "packet" of ions is pushed away from the pusher region, and another "start-stop-cycle" event starts.

Thanks to advances in ion mirror and detector technology, a resolution of about 40 000 (at m/z 400) can now be achieved by TOF instruments. The dynamic range of a TOF is not yet as good as that of a quadrupole, but it is constantly improving and currently varies between four and five orders of magnitude, depending mainly on the type of detector used (Hart-Smith & Blanksby, 2012). The most valuable feature of the TOF mass analyzer is its rapid acquisition. This allows for high resolution and accurate mass measurement (< 5 ppm) over narrow chromatographic peaks without losing chromatographic resolution. However, the capacity to distinguish analyte signals from noise signals decreases with increasing scan speed. The current set up on the GC/TOF instrument that I used for this research therefore has a acquisition rate of 30 Hz, even though the instrument has an acquisition limit of 500 Hz.

Since mass measurement is dependent on flight time, the speed and accuracy of the detector is of the utmost importance in TOF mass analyzers. Once the ions hit the multichannel plates (MCP) at the end of the flight tunnel the signal must be amplified and digitized. There are two different type of detectors used in contemporary TOF: TDC (time to digital converter) and ADC (analog to digital converter). ADC converters are becoming increasingly popular, since they can differentiate between one, or more than one, ion of the

same m/z hitting the detector in the same transient (start-stop cycle). A TDC detector can only register one ion bin and transient and it therefore reaches saturation faster than an ADC detector (Bristow *et al.*, 2008). A saturated detector causes problems with both quantification of analytes and mass accuracy, since the apices of the mass spectra and the chromatographic profiles cannot be accurately determined.

Tandem MS

High resolution instruments and accurate mass measurements can be used to identify the elemental composition of an analyte. Together with data from previous analysis of standard compounds and retention times from the chromatographic system, this is in some cases enough for metabolite recognition. However, the elemental composition gives no structural information about the analyte/metabolite. Tandem MS, where two mass analyzers are placed in-line, gives the researcher the opportunity to select an analyte, fragment it, and perform mass measurements on the fragment ions. MS/MS analysis can be regarded as consisting of three steps: (a) in MS(I) precursor ions are selected; (b) each precursor ion is fragmented in the collision cell; (c) the formed fragment ions are then analyzed in MS(II) (Hart-Smith & Blanksby, 2012; Herbert & Johnstone, 2003).

In many cases, the fragmentation technique used within the collision cell is collision induced dissociation (CID). In CID the selected precursor ions enter the collision cell, often a quadrupole or a hexapole, which is filled with a neutral inert collision gas, usually N_2, He or Ar. As the precursor ions travel through the collision cell they collide with the gas molecules, causing fragmentation of the precursor ions. Collision energy is also applied to the collision cell in order to increase the number of collisions and aid fragmentation. By varying the collision energy the fragmentation pattern can be controlled; low collision energies cause less fragmentation of the precursor ion (and result mainly in neutral losses of, for example, $–H_2O$ or -COOH) whereas higher collision energies can induce ion cleavage and molecular rearrangements (Sleno & Volmer, 2004).

In metabolomics profiling the combination of a quadrupole followed by a TOF (QTOF) is a powerful analytical instrument. The TOF alone will give quantitatively accurate mass measurements for protonated molecular ions, if the quadrupole and the collision cell are left “open” and inactive. In MS/MS mode, precursor ions can be selected manually or automatically in the quadrupole, fragmented using fixed or ramped collision energies in the collision cell, and the resulting product ions measured in the TOF (MS/MS spectra). By reanalyzing one or a few samples within a metabolomics profiling

study using automated MS/MS settings, the probability of identifying or classifying metabolites increases because of the structural information gained from the MS/MS experiment.

Triple quadrupoles (QqQ), in which the first and the last quadrupole work as mass filters and the middle quadrupole (or hexapole) is used as a collision cell, are extensively used in many different research fields because of the robustness, sensitivity and high dynamic range of the instrumentation. Since the two different mass analyzers can be run in fixed or scan mode, a combination of these different modes gives several different analytical choices depending on the question of interest (Fig. 8).

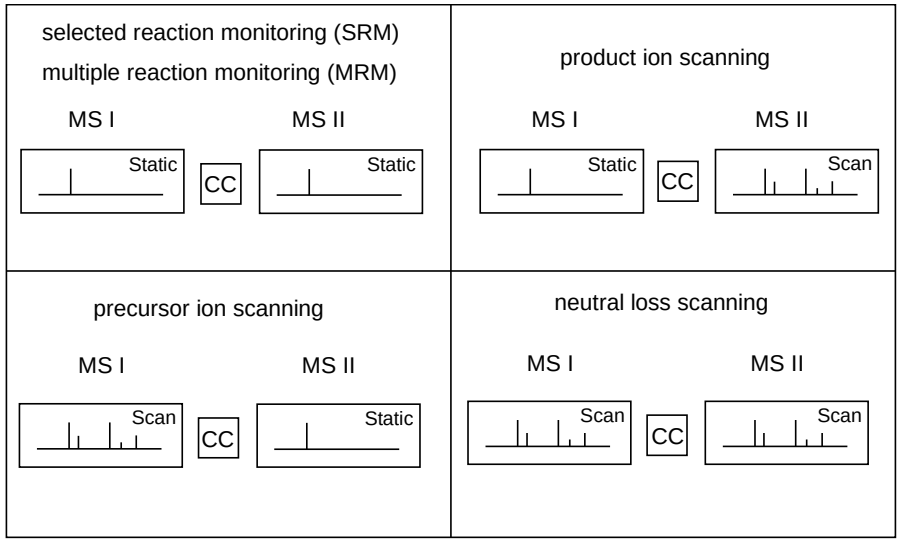

Figure 8. The different modes in which a triple quadrupole can be operated. MS I - first quadrupole, CC - collision cell (second quadrupole), MS II - third quadrupole.

In product ion scanning a precursor ion is allowed through the first quadrupole and fragmented in the collision cell, and the last quadrupole is run in scan mode, to determine the m/z ratios of the ions produced. Product ion scanning is used to acquire structural information about the precursor ions. In quantitative targeted methods the first and last quadrupole is static, letting only specific precursor and product ions through to the detector. If only one precursor-product ion combination is monitored, the term selected reaction monitoring (SRM) is used. To increase the selectivity of analysis, two different product ions are often scanned, a quantifier ion and a qualifier ion. Where several different precursor-product ion combinations are analyzed, this mode is termed multiple reaction monitoring (MRM), and is widely used in the pharmaceutical, forensic and medical industries, because of its sensitivity and

selectivity. If the last quadrupole is static, letting only product ions of a specific m/z through, while the first quadrupole is scanning (precursor ion scanning), it is possible to search for analytes that give rise to a common product ion. The final possible combination is neutral loss scanning, where precursor ions that release a specific neutral fragment upon CID fragmentation can be selectively analyzed. The two different quadrupoles are scanned in parallel with a constant mass difference, corresponding to the neutral loss, between them (Hart-Smith & Blanksby, 2012).

The resolution of a QqQ is still low, but with the selectivity gained from MRM mode in combination with a good chromatographic separation, high resolution is not needed.

3.3 Data processing and metabolite identification

The raw data output from any mass spectrometer is a series of mass spectra acquired at different time points or different scans. Each scan collected contains information about m/z values of ions and the corresponding intensities. The data files acquired can be viewed as a three dimensional cube (intensity, m/z and time). Data processing methods aim to extract the analytical information from these datasets and reduce each three dimensional cube to a two dimensional data matrix, in which the dimensions represent variables and observations. The matrix should thus consist of the peak intensity of each metabolite (variable) for every sample (observation), using mass spectra and retention time as peak identifiers.

In the case of targeted metabolite analysis, data processing methods are generally straightforward, and many steps, such as integration of quantifier ions and calibration curve calculation, are often automated using quantitative software provided by the supplier of the mass spectrometer. However, in metabolomics studies, when an unbiased data processing approach is required, the scenario becomes more complicated. There are numerous data processing methods, both commercial and freely available, including stand-alone software and scripts written for statistical packages. For a review on data processing strategies for the analysis of metabolomics data obtained from hyphenated mass spectrometry techniques see (Katajamaa & Orešič, 2007).

Often the first step in data processing is a the application of a filtering step or a data pre-treatment in an attempt to reduce the data, eliminate noise and adjust the baselines of the mass spectra and the chromatograms. After data pre-treatment, peaks or analytical features must be extracted from the raw data, and a distinction must be made between true ions hitting the detector and chemical noise. Several different strategies are available for feature extraction, and it is

at this step that the different processing methods differ the most. The simplest strategy is to bin data points in either the mass spectral direction or the chromatographic direction. Binning is an easy and efficient way to reduce the data, but it carries a high risk of missing information on the way. Jonsson and colleagues used a multivariate binning strategy in which they divided GC-MS chromatograms from overlapping samples into time windows and used multivariate statistics to search for the summed m/z channels that showed high variation within each window, and related those to differences between plants (Jonsson *et al.*, 2004). Another strategy for feature detection is peak picking (e.g. MetAlign (Lommen, 2009), XCMS (Smith *et al.*, 2006), mzmine (Katajamaa *et al.*, 2006), MarkerLynx™ (Waters, Manchester, UK)), in which all m/z channels with "peak shapes" are recognized and integrated as separate peaks, resulting in intensity values from several ions that represent the same metabolite. Peak picking is more commonly used in LC-MS, since GC-MS (EI) results in several ion fragments so that peak picking would result in a very large number of variables representing the same metabolite. In modern peak picking methods used for LC-MS (e.g. Mass Hunter Mass Feature Extraction, (Agilent Technologies Inc., Santa Clara, CA, USA)), isotopic distribution patterns and grouping of isotope clusters (from common adducts and neutral losses) are combined into mass features each representing a single compound with a calculated intensity volume. Other data processing methods use mathematical curve resolution strategies (deconvolution) (e.g. Leco ChromaTOF (St. Joseph, MI, USA), AMDIS (http://chemdata.nist.gov/mass-spc/amdis/)). The aim of deconvolution is to resolve ion traces with a common origin into one deconvoluted component. Deconvolution, which is commonly used in GC-MS metabolomics studies, exploits the fact that fragment ions originating from the same compound have the same retention time. However, deconvolution in metabolomics studies can be confounded by the complexity of a metabolomics sample, which causes a high degree of co-elution of compounds with, in some cases, overlapping isotope patterns. One way to reduce the number of overlapping peaks and hence to facilitate deconvolution (and increase the sensitivity of the separation) is to use two-dimensional chromatographic separation, for example GC*GC-MS as performed by (Oresic *et al.*, 2011). However the use of a two-dimensional separation technique increases problems arising from instrumental drift and increases the sample run time. Although deconvolution can be improved by using two-dimensional separation, the subsequent processes of alignment and sample matching are complicated. After peak picking or deconvolution, the extracted features must be aligned and combined across samples to produce a data table of the relative concentrations of all detected features using uniform variables. Alignment aims

to correct drifts in retention time that occur between injections of samples (as reviewed by Katajamaa & Orešič, (2007)).

Since identification in mass spectrometry is dependent upon data obtained from analysis of standards, mass spectral libraries play an important role in increasing the success rate for metabolite identification. In GC-EI-MS, where the fragmentation procedure is highly standardized, there are publicly available EI-mass spectral libraries containing at least 100 000 entries (NISTMS, USA). However, many of these entries are for synthetic compounds which are very unlikely to appear in metabolomics analyses. As pointed out earlier, the methods used in most GC-MS metabolomics studies include methoxyamination and silylation derivatization protocols. There is therefore a need for spectral libraries of derivatized compounds with the corresponding retention indexes. The Golm metabolome database made its GC-EI-MS mass spectra (quadrupole and TOF data) with the corresponding retention indices publicly available in 2005 (Schauer *et al.*, 2005). However there is still a need for the metabolomics community to continue to add entries to the GC/MS libraries, in order to reduce the percentage of unknown peaks. Using our current GC-MS setup we can successfully identify roughly 30% of our deconvoluted peaks, and classify an additional few percent (unpublished data). Historically, limited effort has been devoted to the development of LC-MS libraries. The lack of a standardized retention index system in LC, and the facts that different analytes require different collision energies and that the fragmentation pattern differs between different instruments (even when the same collision energy is used), are all factors making it difficult to compare LC-MS libraries between labs. However, LC-MS and LC-MS/MS libraries of analyzed standards are now being expanded rapidly (METLIN Metabolite Database, http://metlin.scripps.edu/), using data obtained with a series of different collision energies and a standardized column gradient. Although this is extremely helpful for identification purposes, most labs are, in parallel, building up their own in-house libraries on the basis of the LC-MS settings used in those labs, in order to include relevant retention-time information. One advantage of LC-MS based metabolomics analysis, using high resolution instrumentation, is that it is possible to determine the elemental composition of a metabolite. However, although high resolution instrumentation can be used to elucidate elemental composition it does not give any structural information about the metabolite, so retention-time information and MS/MS should not be discounted. The issue of identification is, and will continue to be, the major bottleneck in metabolomics studies since the number of standards commercially available for purchase is limited, meaning that costly and time-consuming de novo synthesis is often necessary. This is still a particular

problem in the field of plant metabolomics, in which the large numbers of (known and unknown) secondary metabolites mean that parallel strategies are needed to increase the proportion of metabolites that can be identified.

3.4 Chemometrics and multivariate statistics

In metabolomics analysis, the number of variables (or metabolites) is often higher than the number of observations (samples analyzed). This, together with the fact that metabolites participate in series of enzymatic reactions and are therefore often by default correlated, means that traditional univariate statistical methods which assume independent variables are of limited use. The number of significant correlations that arise by chance alone is high when univariate statistics are applied to data matrices of this type (Trygg, 2001). Another problem when the number of variables is high is that it is conceptually difficult to obtain an overview of the data. In chemometrics (Eriksson *et al.*, 2001), multivariate statistical methods such as principal component analysis (PCA, (Jackson, 1991)) or partial least squares (PLS, (Wold *et al.*, 1984)) assume that the data contain correlated variables and provide appropriate methods for data visualization, making chemometrics methods highly suitable for metabolomics analysis. The chemometric method of choice depends on the question being asked:

PCA is an unsupervised method searching for the greatest variation in a multidimensional space (X matrix). PCA reduces the multidimensionality of the space by reducing the largest variation within the dataset down to a few latent variables, i.e. principal components. The number of principal components needed to describe the data depends on the nature of the data. The first two principal components (which explain the greatest variation within the data matrix X) form a plane in the multidimensional space onto which all samples can be projected, so that they can all be visualized on a two-dimensional plot (a score plot) which can be interpreted by the human eye. The score plot can be used to visualize trends, outliers and classes within the data. The causes of outliers and trends within the data can be examined by interpreting the loadings plot. This describes the relative importance (weight) of each measured variable within the model. The variation within each sample that is not described by the new latent variables can be found in the residual.

PLS, or orthogonal-partial least squares (OPLS, (Trygg & Wold, 2002)), are used when a relationship between two blocks of data (data matrix X and known information Y) is sought. PLS and OPLS search for correlation between the data acquired and a known/measured variable. They do so by calculating latent variables that capture the Y-related variation in X. An interpretable loadings

plot describes the relative importance of each variable in the X-Y relationship. OPLS and PLS are similar methods; however, in OPLS, systematic variation within the X matrix that is orthogonal to the question asked (due to instrumental drift etc.) is removed from the data matrix and a new predictive component is calculated for the X-Y relationship with the orthogonal variation removed. The orthogonal variation that has been removed can be visualized in the orthogonal latent variables. OPLS gives predictions similar to those of PLS, but the quality of interpretation is improved since the structural noise in the model is separated from the predictable variation. PLS (or OPLS) can be used to find correlations between, for example, physicochemical properties of compounds and their retention on a column, a method known as Quantitative Structure Retention Relationship (QSRR, (Kaliszan *et al.*, 1992)). Partial least square discriminant analysis (PLS-DA or OPLS-DA) is a supervised method in which the Y matrix holds discriminate information (e.g. wild-type vs. mutant). PLS-DA or OPLS-DA is used to find differences between specified groups, e.g. differences between the metabolome of prostate cancer bone metastases and that of normal bone (Thysell *et al.*, 2010).

4 Objectives

The overall aim of this work has been to develop analytical tools for metabolomics analysis and to apply these tools in order to obtain information about metabolite profiles during wood development.

The main specific objectives were to:

Develop a processing method for GC-MS data that allows the user to go from GC-MS based metabolomics analysis to biological interpretation in a high-throughput manner. (**Paper I**)

Fine-tune metabolomics methodology to allow analysis of sub-milligram amounts of tissues in order to study the process of activation of dormancy in autumn and growth reactivation in spring in the vascular cambium. The aim was also to integrate metabolite data with gene transcript data to gain a better understanding of the complex regulation of tree dormancy. (**Paper II**)

Use multivariate statistics to predict metabolite retention behavior in GC-MS, and to use the regression model so produced to build up a database of predicted retention indexes for metabolites in order to improve identification ratios in metabolomics. (**Paper III**)

Establish a LC-MS method for measuring nitrogen fluxes in plants fed with isotopically labeled ammonium nitrate. (**Paper IV**)

Create a metabolic roadmap of wood development, by analyzing metabolites and hormones using both targeted methods and untargeted metabolite profiling approaches. (**Paper V**)

5 Results and Discussion

5.1 Development of a method for processing GC-MS data

The raw data output from a GC-MS instrument is a series of mass spectra acquired at different time points or scans. Since metabolomics studies include multiple samples, the data acquired can be viewed as a three dimensional structure (Fig. 9A). Data processing methods are designed to reduce the three dimensional data down to two dimensions (representing the variables and the observations) and to extract analytical information from individual samples. The resulting data matrix stores peak intensities for each metabolite from every sample, using mass spectra and retention time as peak identifiers. Effective data processing methods must be able to extract information from all samples and keep the variables constant in-between samples. **Paper I** describes the development of Hierarchical-Multivariate Curve Resolution (H-MCR), a semi-automatic strategy for simultaneous deconvolution of multiple samples. The strategy is illustrated in Fig. 9. The first steps in H-MCR include smoothing, background reduction and peak alignment. A hierarchical approach, in which the chromatographic region is split into time windows, is used in order to reduce the amount of data to be handled by the algorithm. The edges of the time windows are set at points where the overall signal intensity is low in order to keep peak splitting to a minimum. Prior to the Multivariate Curve Resolution step, each time window is unfolded into a 2 dimensional data matrix in which the two dimensions are chromatographic profile (scan or RT) and m/z value (Fig. 9A).

An index is calculated based on the total variation relative to the scan-to-scan variation, for each mass channel. The index will be higher for distinct peaks than for noisy regions. Mass channels with low index values are excluded from the data, with the result that data amounts as well as noise will be reduced. The data matrix is then resolved using Alternating Regression (AR). AR is an iterative method that alternates between two operations until

convergence (Karjalainen, 1989). The first step in the AR procedure is to find a starting point for the deconvolution. For this, the PURE algorithm of R. Tauler and A. de Juan (http://www.mcrals.info/) was applied to find the "purest" mass channel, which was used as the starting point (i.e. target spectrum, ST) for the iterative AR algorithm, as shown in Fig. 9B.

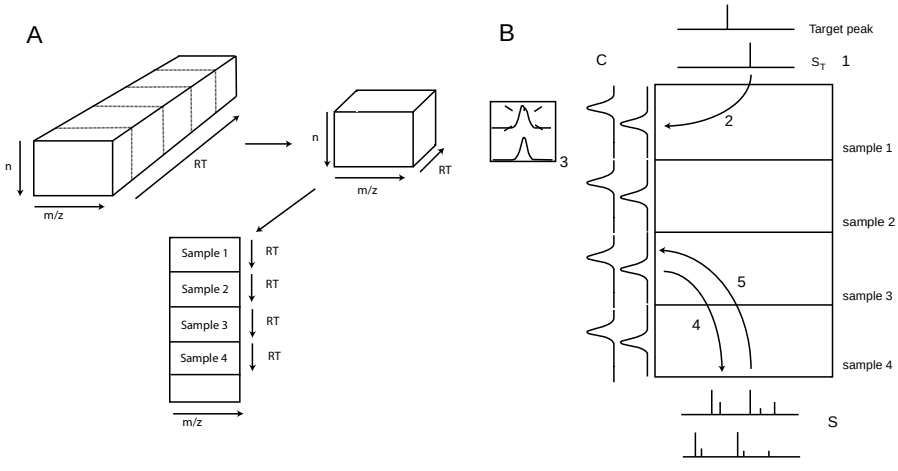

Figure 9. The principles of H-MCR

The AR algorithm starts with one target spectrum (S_T).

1. The target spectrum is multiplied by the unfolded data matrix (X), resulting in chromatographic profiles (C).
2. The chromatographic profiles (C) are then modified by applying a Gaussian (unimodal) shape filter, eliminating the possibility of multiple peaks for each mass channel and sample.
3. The cleaned-up chromatographic profiles (C) are then multiplied by the matrix (X), resulting in a full spectral profile (S).
4. The full spectral profile (S) is then multiplied by the data matrix (X), resulting in an updated chromatographic profile (C).
5. Steps 3-5 are performed until convergence is reached (with a maximum of 50 iterations).

The algorithm then attempts to add another "pure" mass channel to the target spectrum (S_T). Steps 1-5 are performed. A retention time constraint is then applied: the detectable peaks must all elute in the same order, for all samples. If there is inconsistency in the order of elution, the algorithm stops

and uses the outcomes corresponding to the last successful number of deconvoluted peaks.

In summary, for peaks to be resolved, they must fulfill one of two criteria: either their chromatographic profiles must be different, or the ratio between the samples must differ in the mass spectral dimension. The output produced by H-MCR consists of integrated areas for all components (for all samples) and the corresponding mass spectra.

The strategy, and the algorithm, which was written in Matlab, were validated using a complex standard mixture and metabolites extracted (Gullberg *et al.*, 2004) from wild type plants and four different GA mutant lines (ga1-3, gai-t6 ga1-3, rga24 ga1-3, and rga-24 gai-t6 ga 1-3). The mass spectra extracted using H-MCR were found to be very similar to mass spectra in the mass spectral library, with an average reversed match (NISTMS) of 904 for the 18 compounds whose relative concentrations were varied in the standard mixture experiment. The areas produced by H-MCR were compared with those obtained by manual integration of selected quantification masses using the software package, Leco ChromaTof, supplied with the instrument. As shown in Paper I, Fig. 4, there was good correlation between the results of manual integration and of integration performed by H-MCR. By applying multivariate statistics (PLS-DA) to the data processed by H-MCR, the four different GA mutants could be separated. Changes in the concentrations of approximately 50 metabolites were found to be responsible for the separation of the dwarf ga1-3 mutant and the semi-dwarf line rga-24 ga1-3, and the identities of some of these metabolites were determined by mass spectral matching.

H-MCR is now the workhorse of our metabolomics facility, being used to convert acquired GC-MS raw data into a data table suitable for multivariate analysis. The extracted mass spectra are used for metabolite identification, making biological interpretation possible. However the H-MCR method has its limitations and it is of the utmost importance that manual inspection of the mass spectra and the data table is carried out before any statistical evaluation is performed. A split peak, where one mass spectrum is found in two or three adjacent variables, is one of the most commonly occurring problems arising when using H-MCR. A chromatographic peak may be split due to peak noise or the positions of window edges. In these cases manual selection of quantification masses and reintegration are necessary. Another common phenomenon is the occurrence of “nonsense” peaks, which are often found at the end of the analysis, in the region where few metabolites are eluting. These peaks must be removed prior to multivariate statistical analysis. In regions of the chromatogram where multiple peaks are eluting, H-MCR will have

difficulty in resolving all overlapping peaks, resulting in confounded mass spectra and inaccurate estimates of area. Again manual curation by inspection of the raw data is needed in such cases. Nor is the integration process perfect. All m/z channels present in the mass spectra are included in the area integration. Hence 73 m/z, a fragment caused by cleavage of the trimethylsilylation derivative which is highly abundant in the mass spectra of all derivatized metabolites, is also included in the integration. This may result in skewed areas in the case of coeluting peaks, since H-MCR may have a problem finding the correct proportion of the 73 m/z-channel in each of the coeluting peaks. Careful investigation of the deconvoluted mass spectra and reintegration of raw data are therefore necessary if the correct biological conclusions are to be drawn.

5.2 Seasonal changes in the vascular cambium

Plants in temperate zones need a strategy for surviving harsh winter conditions. To protect the vascular cambium, trees halt the growth of this tissue and put the meristems into a dormant state. The transition from active growth in the summer to growth cessation in the autumn and back to active growth again in spring is termed the activity-dormancy cycle (Rohde & Bhalerao, 2007). In *Populus*, the most crucial factor regulating the initiation of growth cessation is day length (Wareing, 1956). As the day length drops beneath a certain threshold, the tree starts to prepare for the winter, by initiating cessation of growth. At the same time, cold hardiness develops, and there is a shift in metabolism towards the accumulation of storage compounds.

In the work described in **Paper II** we performed tissue-specific metabolite profiling of cambial cells from aspen in order to follow metabolic changes over a growth season. In parallel with metabolite profiling, microarray analysis was also performed to improve our understanding of the regulation of the activity-dormancy cycle in the cambial zone of trees.

Since sugars are highly abundant metabolites in woody tissues, we performed fractionation of the extracted metabolites using solid phase extraction (SPE) so as to be able to detect other metabolites that might be co-eluting with the sugars. A mixed mode cation exchange column was used and three different fractions were collected: an acidic fraction (containing sugars etc.), a non-polar fraction (containing, for example, fatty acids and organic acids) and a basic fraction (containing, e.g., amino acids). The fractions collected were analyzed separately on GC-TOFMS, and the data were processed by H-MCR and evaluated using multivariate statistics. Solid phase extraction can result in sub-fractionated samples, and in our analysis we did

indeed find an overlap between the different fractions for ~10% of the putative metabolites. However this sub-fractionation did not affect statistical evaluation. PLS-DA score plots for the different fractions are presented in Fig. 10. In the acidic and basic fractions, clear separations between samples from different time points were obtained, whereas in the organic fractions separation was less marked.

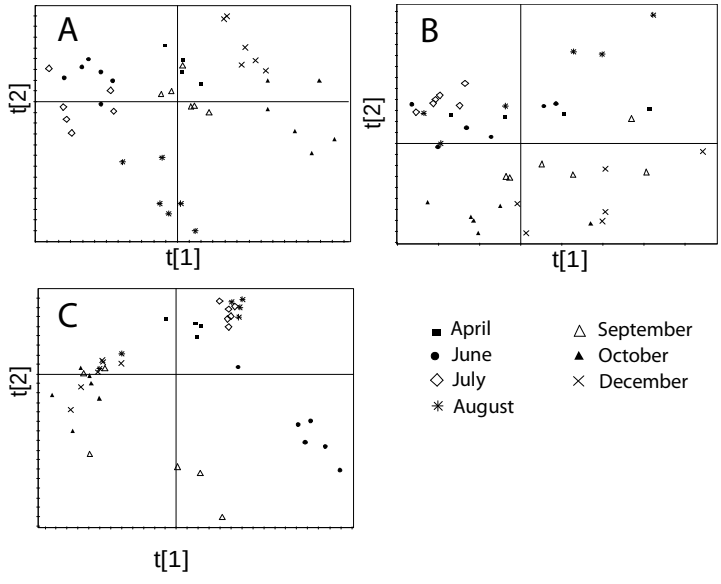

Figure 10. PLS-DA of the different fractions. A) Acidic fraction (7 components, R2X=0.65 , R2Y=0.78 , Q2Y=0.55). B) organic fraction (4 components, R2X= 0.47, R2Y= 0.43 Q2=0.11). C) basic fraction (2 components, R2X= 0.36, R2Y= 0.26, Q2=0.2).

To identify metabolic shifts occurring during spring reactivation and autumn growth cessation, pair-wise comparisons were performed between the April-June, June-August, August-September and September-December samples, for each fraction. An overview of the metabolite shifts observed is given in Paper II, Fig. 3. The accumulation of different metabolites follows temporal patterns, and is associated with specific cellular processes occurring during the successive shifts in the activity-dormancy cycle.

Acquisition of cold hardiness in the autumn involves protein synthesis and the biosynthesis of cryo-protectants. These are processes that require carbon and energy, at a time when the photosynthetic rate is declining. Gene transcript data showed induction of the genes encoding some of the key enzymes in starch breakdown from August onwards, suggesting that starch may be utilized as a carbon source in the production of sugar-based cryo-protectants such as raffinose. In our metabolite profiling experiment the levels of sucrose,

raffinose and galactinol, a precursor of raffinose biosynthesis, were found to increase from August onwards

The large vacuole in the fusiform initials is fragmented into several smaller vacuoles during the acquisition of cold hardiness; this process demands synthesis of new vacuolar membranes and hence the synthesis of various fatty acids. The gene transcript data showed induction of genes with products involved in the biosynthesis of phospholipids and in lipid desaturation in the autumn. Consistent with this finding, our metabolite profiling revealed increased levels of glyceric acid, a precursor of fatty acid metabolism, in the autumn. Many of the metabolites whose levels increased during this period could be classified as fatty acids, and in general there was a shift towards a higher proportion of fatty metabolites in the autumn. Increased levels of cinnamic acid, coumaric acid and caffeic acid, all precursors of the phenylpropanoid pathway, were found in August. These changes may be an indication that there is an increase in production of monolignols and lignification of secondary walls at this time of year.

The levels of gamma-aminobutyric acid (GABA), which is known to be induced by stress (Bouche *et al.*, 2003), increased during the same time frame as that in which genes related to cold hardiness were induced. The level of GABA also increased after low temperature treatment of Arabidopsis (Cook *et al.*, 2004). This may indicate a role for GABA in regulating the induction of cold hardiness in trees. Because the role of ABA in the regulation of cold hardiness in autumn is unclear, we also measured the levels of ABA in the cambial cells (Paper II, Table 1). ABA showed marked accumulation in September. Since the peak level of ABA content in the autumn occurred after induction of many genes related to low-temperature and cold hardiness, our findings suggest that ABA might act downstream of the short-day and low temperature signalling pathways, inducing cold hardiness at a later stage.

The greatest shift in metabolite pattern during the reactivation phase in spring was an accumulation of amino acids in June together with a rapid decline in the amount of carbohydrates from April to June. As reactivation of the cambium occurs prior to any photosynthetic activity in the spring, the dividing cells require an alternative carbon source. The rapid decline in sucrose level was accompanied by induction of sucrose synthase gene expression, suggesting that sucrose is used to produce the fructose and UDP-glucose needed for cell wall biosynthesis and other metabolic processes in the dividing cells. The accumulation of amino acids in June could be a result of storage protein breakdown during the reactivation phase, to enable synthesis of new proteins. Levels of many of the above-mentioned amino acids were also found to increase in September, coinciding with the induction of transcription of

genes encoding bark storage proteins, suggesting that storage proteins are synthesized during the transition to dormancy in the autumn.

Parallel analysis of microarray data and metabolite profiles allowed us to interpret the molecular regulation of activity-dormancy transitions in aspen in greater depth than would have been possible using either of the techniques in isolation. Many overlaps were found between the metabolite and transcript data, although in some cases the temporal trends in the two datasets were shifted relative to one another. This was particularly obvious in the case of the shift from summer to autumn. In the transcript data, the greatest difference was observed between the July and August samples, whereas the greatest shift in the metabolome occurred between August and September. Unfortunately two time-points taken in May are missing from the metabolomics data due to lack of samples. It would have been interesting to see whether an accumulation of amino acids also occurs in May, possibly indicating breakdown of storage proteins. Samples taken in June would be less informative on this point, since by then phloem transport of amino acid from photosynthesizing leaves would also contribute to the amino acid content of the cambium.

From an analytical point of view, fractionation of the extract using SPE allowed us to detect and deconvolute far more metabolites than would have been possible without fractionation. This was mostly due to the fact that the highly abundant sugars could be collected in a separate fraction. This made it possible to detect, for example, p-coumaric acid, which is not normally found in GC-TOFMS metabolite profiling projects since it co-elutes with many of the monosaccharides. However, compared to transcriptomic data, metabolite profiling is less amenable to biological interpretation, since many metabolites cannot be identified. Much effort still needs to be devoted to the identification, or at least the classification, of these unidentified metabolites.

5.3 Building a database of retention index predictions for GC-MS

Despite the structural information which can be obtained by EI-MS, and the continuous updating of mass spectral libraries, the main reason why many metabolites remain unknown is that the number of compounds in the libraries is, and will be for some time, a limiting factor. Since similar compounds generally have very similar or identical mass spectra, retention times also need to be taken into account. Schauer and colleagues showed that by converting retention times to retention indices, it is possible to compare retention characteristics from different analyses, different instruments and different labs, provided that the stationary phase of the GC-column used is the same (Schauer

et al., 2005). By specifying the structure of a compound using chemical descriptors and relating this information to retention time it is possible to calculate the elution time of a compound, in a method known as Quantitative Structure Retention Relationship (QSRR) (Kaliszan, 2007; Farkas *et al.*, 2004; Kaliszan *et al.*, 1992).

In the work reported in **Paper III** we used the QSRR strategy to create a RI database containing the predicted retention indices of the metabolites present in the KEGG metabolite compound database.

As all major GC-MS based metabolomics databases are based on methoxyamination and trimethylsilyation of the metabolites, this strategy was applied to the structures downloaded from KEGG (see Paper III Fig. 1). The database of in-silico derivatized metabolites that we created contained approximately 13 000 structures. A selection of the derivatized structures was present in our in-house GC-MS spectral database, and 310 of these unique chemical structures with known retention indices were used to calculate a QSRR model. The 310 compounds represented a variety of different chemical classes (Paper III, Fig. 2A) with retention index values ranging from 800 to 4000 (Paper III Fig. 2B). In total 175 descriptors, both "numerical descriptors" such as the numbers of carbon and nitrogen atoms and numbers of double and triple bonds, and descriptors of the physical and chemical properties of the molecule, were used for the models.

A principal components analysis of the dataset gave 6 principal components (Paper III Fig. 3), supporting our hypothesis that the dataset was indeed very diverse. One tenth of the dataset (31 compounds) were excluded from the dataset to be used later on as an external testset. A PLS model was calculated using the chemical descriptors of the remaining 279 compounds as X-variables and the RI as Y variable. The number of components in the PLS model was determined so as to maximize its predictive capacity for the external test set, i.e. to minimize the root mean square error of prediction (RMSEP). RMSEP decreased with the number of PLS components until the 9th component was added. This finding resulted in a PLS model containing eight components, describing 98% of the variation in Y (R2Y(cum): 0.98), and 96% of the variation in Y could be predicted by the model by means of cross validation (Q2(cum): 0.96). The eight component model gave a RMSEP of 159 and a root mean square error of estimation (RMSEE) of 83 (for RI), which, given the wide range of retention index values (from 800 to 4000) in our data set, indicates a good level of relative precision (83/(4000-900)= 2.7%) (Paper III, Fig. 4).

Determining the number of components in a model is a difficult task, which will have a major impact on the end result. The last 4 components in our model

had a substantial influence on its predictive capability (RMSEP decreased dramatically with the last 4 components). However as the number of components in a PLS-model increases, the risk of modelling noise also increases. On the other hand, too few components will result in poor predictive capability. In our case where the prediction is of greatest importance, the number of components was determined by the predictive capacity of the external test set (RMSEP). Given the large number of different compound classes represented in the model (24 compound classes; Paper III, Fig. 2A), the number of components added was considered to be realistic.

We predicted the retention times of the remaining in-silico derivatized compounds originating from the KEGG compound database using the QSRR model. To validate the database we had created, we searched for compounds within the RI DB that are represented in the Golm Metabolome Database mass spectral library (Schauer *et al.*, 2005) but not present in our own in-house GC-MS library and used these compounds to further validate our QSRR model. The RI predictions for these compounds, were fairly accurate (Paper III, Table 2), suggesting that the predictive DB is a useful source of RI values for compounds where a lab has no access to standards.

A closer investigation of the molecular weights of the derivatized compounds shows that the molecular weight distribution in the predictive RI DB is very similar to the distribution within the 310 compounds (Fig. 11).

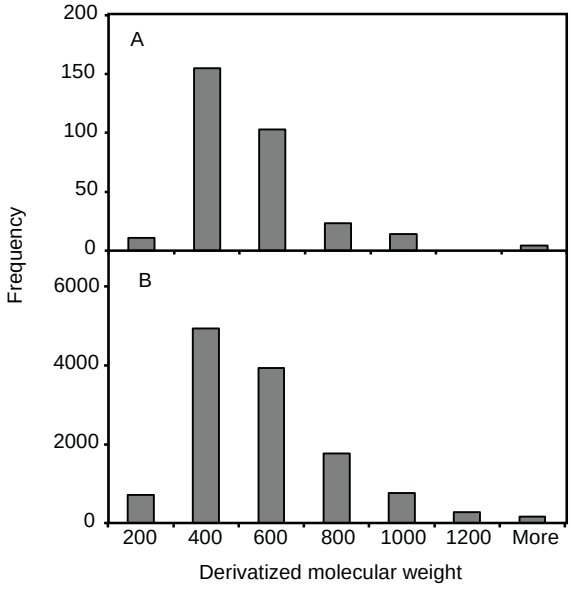

Figure 11. Histogram of the molecular weight distribution for A) the 310 compounds in the QSRR model, B) compounds in the predictive RI DB

In summary, the QSRR strategy proved to be a reliable method for predicting retention index. Given the RMSEP of the QSRR model and the predictions made for Golm Metabolome Database we suggest that an accuracy of +/- 150 in RI should be considered realistic when using the predictive RI DB. We believe that the predictive RI DB described here and in **Paper III** will be useful in determining the identity of unknown metabolites, and thus in reducing the number of non-annotated metabolites in GC-MS based metabolomics studies (Paper III, Fig. 6).

5.4 Determining ^{15}N- incorporation into amino acids and amines in plants

Amino acids are important in many different metabolic processes in plant cells. They are cornerstones of primary metabolism, they are essential as the subunits of proteins and as precursors of secondary metabolites, and they can serve as energy resources when the photosynthetic rate is limiting. In the study described in **Paper IV** we measured ^{15}N fluxes using the commercially available AccQ-Tag™ (ACQ) and standard LC-MS and LC-MS/MS setups. The incorporation of ^{15}N into different amino acids in aeroponically grown *Populus* trees fed with doubly labelled ammonium nitrate was determined.

In flux studies based on stable isotope labelling it is necessary to study isotopomer traces. In order to use widely-available MS instrumentation in stable isotope labelling experiments, adequate sensitivity and precision of detection are important factors, since the 2nd, and sometimes the 3rd and 4th, isotope traces must be quantified. In the experiments described in **Paper IV** we derivatized the amino acids with ACQ (Paper IV, Fig. 1), which results in dramatically improved sensitivity and chromatographic separation of the amino acids compared to the non-derivatized forms.

To determine the limits of detection of ^{15}N-incorporation, known amounts of the isotopically labelled amino acids were mixed with the corresponding unlabelled standards in 9 different proportions, ranging from 1% to 100% of the labelled standard. We were able to accurately determine the level of labelled nitrogen for all of the amino acids tested (Paper IV, Table 2). We concluded that the limit of detection of ^{15}N incorporation into amino acids when using the LC-TOFMS setup was approximately 5%, and that above the detection limit incorporation could be determined with high precision (~ ±0.5 percentage points). The method we developed was applied to hybrid aspen trees grown in aeroponics cultures in order to study the biosynthesis and turnover of amino acids and a few biologically important amines. Root samples were taken at five time points: just prior to addition of the ^{15}N-labelled

ammonium nitrate (0h), and 30 minutes, 1h, 3h and 7 days after addition of labelled ammonium nitrate. The absolute quantities of ^{15}N in different amino acids and amines were monitored (Paper IV, Fig. 3). As anticipated, the incorporation of ^{15}N was rapid; after 30 minutes almost 30% of the nitrogen in the glutamine pool was labelled. From the results of this *Populus* experiment we concluded that aeroponically grown plants fed with ^{15}N labelled ammonium nitrate represent a suitable system for studying nitrogen uptake and metabolism in plants using our LC-TOFMS approach. By using LC-QqQ-MRM methodology we were able to further decrease the limits of detection in our analysis. However, as the sensitivity increased we found that contaminating amino acids from other sources became an issue. For instance, glycine was always detected in our analytical blanks when they were analyzed using the triple quadrupole. An investigation into the cause of this background contamination showed that the quality of the borate buffer used was an important factor influencing background levels (Fig. 12).

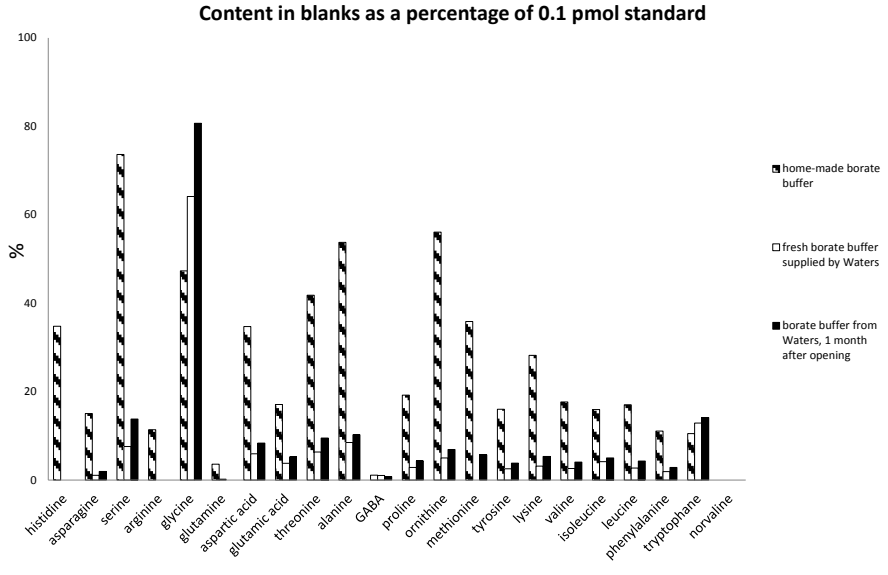

Figure 12. Amino acids in our surroundings. It is clear from our investigation that commercial borate buffer, although it still has a high level of background glycine, was superior to the borate buffer that we prepared in-house.

The amounts of different amino acids in our surroundings seem to vary; for instance glutamine, methionine and arginine were not present in large amounts

in the commercial borate buffer, whereas serine, tryptophan and particularly glycine could easily be detected. The source of the "polluting" amino acids is unclear; it may be laboratory chemicals or glassware. Interestingly, this finding suggests that it is almost impossible to grow plants in an amino acid free environment. Although we encountered problems because of background levels of certain amino acids when performing LC-QqQ-MRM analysis, for those amino acids that were not found in the blanks we achieved a decrease in detection level when moving from the TOF instrument to a triple quadrupole in dynamic MRM mode (DMRM). Another advantage of combining ACQ and triple quadrupole analysis is that it is possible to use precursor ion scanning for screening of N-containing compounds using the highly specific m/z 171 fragment. As shown in Paper IV, Fig. 5, this strategy enabled us to identify labelling in an unknown compound which was subsequently identified as ethanolamine. This strategy may be of interest for detecting incorporation of ^{15}N into compounds that are not components of the normal amino acid biosynthesis pathways.

A possible approach for future labelling studies would be to focus on low-abundance amino acids, e.g. arginine, citrulline, histidine, lysine and methionine, using the LC-QqQ-DMRM method, thereby reducing the number of transitions in DMRM. The more highly abundant amino acids could be analyzed using the TOF setup.

5.5 A metabolic roadmap of wood development

Wood, or secondary xylem, is the product of xylogenesis, a highly organized developmental process that starts with the formation of cambial derivatives and ends with empty shells consisting of lignified secondary cell walls. Fully mature xylem has undergone a series of cellular processes including cell division, cell expansion, secondary wall formation, lignification and programmed cell death (as reviewed in Mellerowicz *et al.* (2001)). Wood is becoming of increasing importance as a renewable resource because of growing demands for our society to be sustainable. Understanding the biology underlying the growth and development of wood development is key to maximizing the future use wood as raw material. In order to carry out comprehensive studies on developmental aspects of growth and development it is necessary to perform tissue-specific sampling and analysis.

In **Paper V** we describe how we analyzed tangential sections covering the region extending from the active phloem to the annual ring to produce a metabolic roadmap of wood-forming tissue in *Populus*.

To enable us to detect highly abundant primary metabolites as well as very low-abundance metabolites, including signalling compounds, we performed both untargeted metabolite profiling and targeted quantification of phytohormones and amino acids. A schematic overview of the analytical setup is shown in Paper V, Fig. 1. PCA analysis of all metabolites measured for tree 1 gave four principal components revealing the four dominant patterns of metabolites in the wood- forming zone (Fig. 13). The major pattern, represented by [t1], which accounted for 33% of the total variation in the data, is an increase from the phloem towards the outermost cambium sections followed by a rapid decrease from cambial section 3-4 inwards. The metabolites making the greatest contribution to this pattern are found in all the different analyses performed. This pattern was not unexpected, since it was assumed that metabolic activity would be high in the cambial zone and decline with distance from this zone. In contrast, metabolites detected in the positive LC-MS analysis and eluting relatively late (compounds that are likely to be non-polar metabolites) were overrepresented among those contributing the most to the second most dominant pattern [t2].

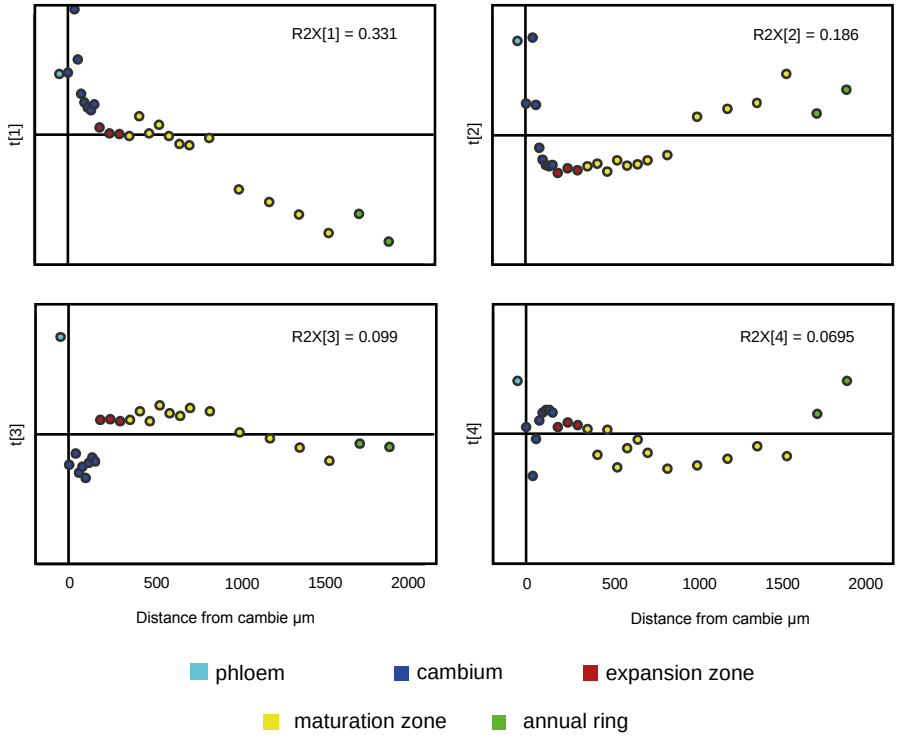

Figure 13. The major metabolic patterns of the wood forming tissue, visualized using the principal components from the PCA analysis.

Unfortunately the identities of the majority of the metabolites detected by LC-MS analysis are yet to be determined.

Plant hormones such IAA, GA and cytokinins are key regulators of plant growth and development. To explore hormone levels across the wood-forming zone, and in particular across the cambial region, the content of IAA and cytokinins per section was determined using quantitative GC- and LC-MSMS methods. A steep concentration gradient over the vascular cambium was observed for IAA (Paper V, Fig. 3A), a result similar to previously published findings Uggla *et al.*,(2001; 1998; 1996) and Tuominen *et al.* (1997). The levels of zeatin riboside (ZR) a precursor of the biologically active compound zeatin, follow the same steep gradient over the vascular cambium as those of IAA (Paper V, Fig. 3A), indicating that there is a concentration gradient of zeatin types of cytokinin over the vascular cambium. A closer investigation of the pattern of ZR and IAA in the cambial region revealed that the peak of ZR concentration is slightly shifted towards the phloem side in comparison to the IAA peak; this pattern was the same in all trees examined (Paper V, Fig. 3B). This is interesting, and a speculative explanation is that cytokinins are involved in regulating the rate of cell division whereas IAA acts as a morphogen providing the cambial region with positional information.

Concentration gradients of the sugars involved in the formation of primary and secondary cell walls are shown in PaperV, Fig. 5. As previously described, the level of sucrose declines rapidly from the cambial tissue inwards (Uggla *et al.*, 2001), whereas the monosaccharides glucose and fructose gradually increase, peaking in the zone of secondary wall formation (at around 800 µm from the cambium). The rapid decline in fructose may be an indication of fructose recycling by fructokinases. Fructose is produced in the conversion of sucrose to UDP-glucose by SuSy. The excess fructose produced by this reaction is thought to be recycled through conversion to fructose-6-phosphate, which can be further metabolized to UDP-glucose, the main precursor of all cell wall polysaccharides (Roach *et al.*, 2012). In support of this hypothesis, fructose-6-phosphate is found at its highest levels in the same region as fructose, but it is maintained at a high level further into the mature xylem than is fructose (Paper V, Fig. 5B). Xylose, which forms the sugar backbone of xylan, the dominant hemicellulose of secondary cell walls in *Populus*, showed a distinct concentration peak about 800-1400 µm from the cambium, coinciding with the formation of the secondary cell wall in this region.

As the secondary cell wall is formed the cellulose and hemicellulose network is “locked” by the process of lignification. The levels of phenylalanine, the main precursor of the phenylpropanoid pathway, decrease from the phloem inwards (Paper V, Fig. 6). As the level of phenylalanine

steadily decreases, the levels of cinnamic acid and p-coumaric acid, two monolignol precursors, increase. In the region where the levels of the monolignol precursors decrease, the levels of different di-lignols start to increase (Paper V, Fig. 6), coinciding with an increase in the levels of xylose.

Since amino acids are key primary metabolites and also in some cases (phenylalanine and tryptophan) precursors of secondary metabolites in trees, a targeted approach was used to quantify the levels of different amino acids. One of the most striking results of this quantification was that large amounts of glutamine were found in the mature xylem. The level of this amino acid in the xylem was 3-4 times higher than the level of found in the phloem tissue (Paper V, Fig. 8). This made us realize that it is important to consider xylem transport when interpreting the pattern of metabolites found in the innermost region of the maturing xylem. By using a targeted method for amino acid analysis we were also able to find trends in the patterns of different amino acids that we would not have seen using different profiling techniques. For instance methionine, which is not easily detected using GC or LC-MS, was found to have a concentration peak in the region of primary wall formation (~600-800 µm from the cambium); the reason for this increase is unknown (data not shown).

From the data presented in **Paper V** we conclude that cambial activity, cell expansion and secondary cell wall thickening are tightly coupled processes as suggested by Uggla *et al.* (2001), see Fig. 14 for a schematic overview of our main findings. In our study, both cytokinin and IAA showed distinct peaks in the cambial region. The concentration maximum for IAA was found to be towards the xylem side of this region, whereas the maximum for cytokinins was further towards the phloem. Altogether these findings suggest that IAA has a role in positional signalling in the cambium. Earlier studies have suggested that the cambial initials, in which there is a high rate of both periclinal and anticlinal cell division, are located towards the phloem side of the cambium (Nilsson *et al.*, 2008; Schrader *et al.*, 2004). Our observation of a cytokinin concentration maximum on the phloem side of the cambium suggests that cytokinins may play a key role in determining the rate of cell division, supporting the hypothesis of Nieminen *et al.* (2008). The newly formed xylem vessels and fibres elongate to reach their final dimensions in the expansion zone, and this is reflected in a dramatic increase in the levels of glucose, which is the monomeric subunit of cellulose, the main polysaccharide of the primary wall. When the fibres and vessels have attained their final dimensions in the expansion zone, the secondary cell wall is formed inside the primary wall (Mellerowicz *et al.*, 2001). We found that levels of xylose, the subunit of the

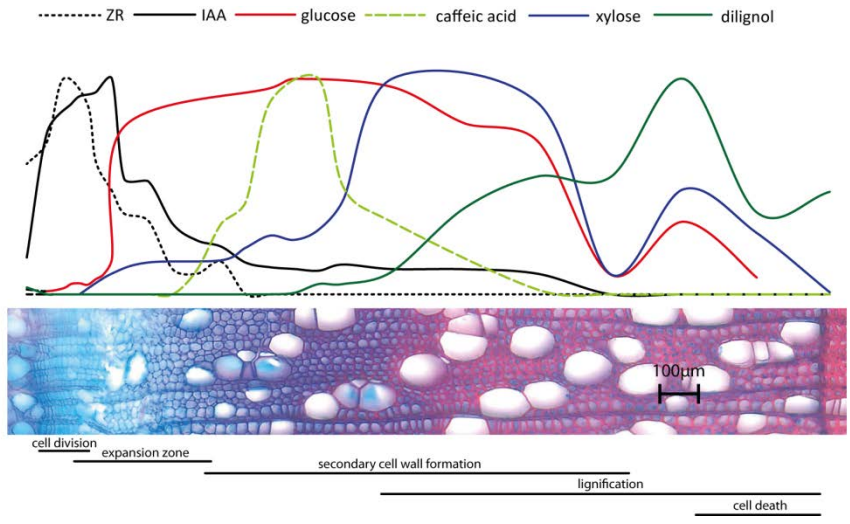

Figure 14. A schematic summary of the major metabolic patterns found in our study. Transverese section from Tree1 shown beneath.

hemicellulose xylan, increase as the glucose concentration starts to decrease, marking the transition from primary cell wall to secondary cell wall formation. Di-lignols were found to accumulate after the xylose peak, indicating that lignification occurs at a later stage of differentiation compared to formation of the secondary cell wall. However, accumulation of the monolignol precursors of lignin occurred before the accumulation of xylose, suggesting that biosynthesis of the monolignols begins during the transition from primary to secondary cell wall formation, whereas polymerization occurs at a later stage of differentiation.

Because the wood development zone is a gradient, data processing and interpretation of the data proved to be a difficult task. Matching the series of sections sampled from different trees is not straightforward, since the length of the series of sections (i. e. the width of the annual ring) varies between different trees. For instance, the cambial zone of tree 1 contained 10 sections whereas the cambial zone of tree 3 contained only 6 sections. This anatomical difference causes problems when attempting to match samples, especially in the maturation zone where no anatomical data were available and the number of sections within the zone in the individual trees varied between 64 and 87. This matching issue causes statistical problems due to the absence of true biological replicates. We therefore chose to evaluate the data based on a single tree, Tree 1. However all the metabolite patterns that we observed have been validated in all trees (Paper V, Supporting Information). The metabolite patterns are substantially different across the series of sections, with many

metabolites being absent from one part of the series and highly abundant in another part.

Analyzing plant hormones from tissue sections (with a weight of < 0.5 mg) is a difficult task; analyzing cytokinins from a fifth of a section even more so. By scaling down the SPE extraction to a minimum as described by Svacinova *et al.* (2012), and thereby increasing recovery and enabling elution of the different fractions in small (50 µL) volumes, the background due to solvents could be reduced. This approach also made it possible for the eluted fractions to be transferred into, and directly evaporated in, micro-vials, thus eliminating an evaporation and dissolution step that was necessary with larger SPE columns and larger elution volumes. The nano-flow chip cube technology used for the cytokinin analysis gave increased sensitivity in the LC-MSMS analysis. In addition, the chip is equipped with an enrichment column, and this allowed a large proportion of the sample (8/10) to be analysed, thereby further increasing the sensitivity of the method.

The results of this study highlight the need to perform multiple analyses on the same sample set in order to cover as large proportion of the metabolome as possible. By combining both targeted and untargeted methods we were able to get a more holistic view of the metabolome at different stages in xylogenesis.

6 Summary and future perspectives

In the course of the work underlying this thesis we have developed analytical tools for metabolomics analysis and applied these tools to reveal metabolite patterns, with a focus on aspects of wood development.

Since metabolite profiling results in huge amounts of raw data, processing methods are needed to extract information from these data. **Paper I** describes the development of a rapid and reliable method for processing GC-TOFMS data (HMCR). The HMCR method has become one of the corner stones in our GC-MS metabolomics platform, since using this method enables us to move from GC-MS chromatograms via metabolite identification to biological interpretation. Although the technique has its limitations, and manual reprocessing is therefore sometimes necessary, HMCR provides us with a starting point to find differences between samples. Despite the improved deconvolution and the high quality of the mass spectral information output by HMCR, identification of analytes in metabolomics studies is still challenging because of the great diversity of the metabolome. The majority of the deconvoluted peaks obtained from GC-MS analysis remain unidentified, due to the paucity of spectral libraries of analyzed standards.

In the work reported in **Paper III**, a database containing the predicted retention indices of ~13 000 methoxyaminated and trimethylsilylated metabolite structures was developed as a means of addressing the low identification ratios characteristic of GC-MS metabolomics studies. By describing the physicochemical properties of the metabolite structures with discrete information we could calculate a QSRR model able to predict a compounds' retention on a GC column. Since low identification ratios are also a problem in LC-MS based metabolomics studies, a similar approach could potentially be used to create a database for predicted retention times in liquid chromatographic separations (C18 columns). The QSRR strategy is more complex in the case of LC, since retention time is determined by interaction with the column, whereas in our GC-MS study no interaction with the

stationary phase was assumed. Changes in pH and the proportion of the organic phase in the solvent during the course of a gradient run are additional factors that will cause changes in metabolite structures during analytical separation. A LC QSRR prediction model is therefore likely to result in less accurate prediction of retention times than is possible with the GC model; nevertheless, any additional information about retention time will help in the process of identification.

Metabolomic studies can give a snapshot of the metabolic status of a cell. However it is the metabolic fluxes within the cell that are the key determinant of cell activity. In the work presented in **Paper IV**, an LC-MS based method was evaluated for the measurement of nitrogen fluxes in plants fed with ^{15}N. By derivatizing amino acids using 6-aminoquinolyl-N-succhinimidyl carbamate (ACQ), it was possible to detect the incorporation of ^{15}N into different amino acids. Although one of the major advantages of using liquid chromatography rather than gas chromatography is that LC does not necessarily require derivatization of the analytes, derivatization prior to LC-MS analysis can improve chromatographic retention and ESI efficiency and enable analysis of specific functional groups. As described in **Paper IV**, ACQ derivatization and the use of a precursor ion scanning setup on a triple quadrupole instrument allow scanning of primary and secondary amines. The method described in this paper was applied to hybrid aspen trees grown in aeroponics culture and fed with doubly labelled ammonium nitrate, and the absolute quantities of ^{15}N in different amino acids and some biologically important amines were monitored. After 30 min, incorporation of ^{15}N was already observed in many of the amino acids analyzed. We concluded that by combining ACQ derivatization of amino acids with LC-MS analysis it is possible to analyze incorporation of stable ^{15}N in labelling experiments on plants. In addition, we showed that it was possible to search for labelled amines in a non-targeted approach using precursor ion scanning. A similar strategy could be used to identify and quantify metabolites with specific functional groups in a semi-targeted approach. For instance, bromocholine will react with carboxylic acids and aldehydes, resulting in the addition of a positively charged quaternary ammonium as described by Kojima *et al.* (2009). By exploiting the specific neutral loss of 59 m/z ($N(CH_3)_3$) upon CID fragmentation, and using neutral loss scanning mode on a triple quadrupole, it would be possible to scan specifically for metabolites containing carboxylic acid or aldehydes.

Tissue-specific analysis in combination with a systems biology approach can facilitate a more holistic interpretation of the biological question being addressed. Because of the highly organized structure of the wood development

zone in trees, tangential cryo-sectioning can approximate to tissue-specific sampling (**Papers II and V**). In the study described in **Paper II** we performed metabolite profiling and transcriptomics in parallel in order to gain a better understanding of the environmental and hormonal regulation of the process of activity-dormancy transitions in aspen trees. The GC-MS metabolite profiling method was fine-tuned to enable detection of metabolite patterns during activity-dormancy transitions in isolated cambial meristem cells. The accumulation of different metabolites followed distinct temporal patterns, and was associated with specific cellular processes occurring at different stages in the activity-dormancy cycle. During the autumn transition, some of the key enzymes involved in starch breakdown were induced from August onwards, suggesting that starch may be utilized as a carbon source for the production of sugar-based cryo-protectants such as raffinose. In support of this theory, increased levels of sucrose, raffinose and galactinol, a precursor of raffinose biosynthesis, were found from August onwards in the metabolite profiling analysis. Many examples of overlap between metabolite and transcript data were found, as described in **Paper II**. In some cases the major trends in transcript data and metabolite data were temporally shifted with respect to one another. This was particularly obvious in the shift that took place between summer and autumn, when the greatest change in the transcript data was observed between the July and August samples, whereas the greatest change in the metabolome occurred between August and September.

In **Paper V** the aim was to create a metabolic roadmap of wood development. For the first time in a plant metabolomics study, by combining targeted analysis of plant hormones and amino acids in parallel with high sensitivity LC-MS and GC-MS analyses, we were able to describe changes in metabolite profiles following the route of differentiation from cell division to cell death. By using the tangential cryo-sectioning technique, which enables us to take tissues samples across a gradient of increased differentiation with increased distance from the cambium, and analyzing a wide a range of metabolites we were able to create a metabolic roadmap of the different stages of wood development from cell division in the cambium to cell death. Specific metabolite patterns were identified within the wood forming zone of *Populus*, and the metabolites included known signalling compounds as well as other type of metabolites. This study exemplifies the need to combine different analytical approaches in order to cover as large proportion of the metabolome as possible. To date, few metabolomics studies in plants have combined metabolite profiling and targeted analysis of the key growth regulators, the plant hormones. The absence of plant hormone data is a significant limitation when it comes to biological interpretation. Recent improvements in

instrumentation and purification protocols should make it possible to create a more detailed hormonal map, covering cytokinins, IAA, ABA and hopefully jasmonic acid and some of the GAs, using extracts from tangential sections of the cambial region. Such a hormonal map would give new insights into the complex regulatory mechanisms underlying wood development.

Although many improvements have been made since the start of my journey, the research field of metabolomics still contains many mountains that must be climbed. In comparison to other "omics" approaches, metabolomics still has a major limitation: the large number of so far unidentified metabolites. Although interesting metabolite profiles and patterns can be found in an untargeted fashion, biological interpretation is impossible without identification of the metabolites. It is therefore important that libraries containing both mass spectra and retention information for analyzed standards are continuously updated. Improvements in analytical approaches continue to be made, e.g. the recently developed ultra-high-performance supercritical fluid chromatographic (UHPSFC) technique in combination with QTOF MS detection has proved promising for the rapid and reproducible detection of triacylglycerols (Zhou *et al.*, 2014). The combination of UHPSFC and traveling wave ion mobility mass spectrometry is likely to be a powerful tool in metabolomics and lipidomics studies.

To conclude, I would summarize metabolomics as "the art of compromising". Choices need to be made at every stage: what is the optimal experimental design, how should the metabolites be extracted, is purification needed? Which metabolites should be analyzed and how should they be analyzed (targeted/ profiling), and how should the data be processed? The method of choice will affect the chances of extracting and detecting different metabolites, and since the number of available analytical choices is large and the number (and amount, especially in developmental biological studies) of samples often limited, compromises need to be made. The experimental design of a metabolomics study is therefore the key factor determining the biological interpretations that can be made.

References

Baba, K., Karlberg, A., Schmidt, J., Schrader, J., Hvidsten, T.R., Bako, L. & Bhalerao, R.P. (2011). Activity-dormancy transition in the cambial meristem involves stage-specific modulation of auxin response in hybrid aspen. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 108(8), 3418-3423.

Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., Trethewey, R.N., Lange, B.M., Wurtele, E.S. & Sumner, L.W. (2004). Potential of metabolomics as a functional genomics tool. *Trends In Plant Science* 9(9), 418-425.

Bouche, N., Lacombe, B. & Fromm, H. (2003). GABA signaling: a conserved and ubiquitous mechanism. *Trends in Cell Biology* 13(12), 607-610.

Bristow, T., Constantine, J., Harrison, M. & Cavoit, F. (2008). Performance optimisation of a new-generation orthogonal-acceleration quadrupole-time-of-flight mass spectrometer. *Rapid Communications In Mass Spectrometry* 22(8), 1213-1222.

Buchanan, B.B., Gruissem, W. & Jones, R.L. (2000). *Biochemistry and molecular biology of plants*. USA: American Sociaty of Plant Physiologist ISBN 0-943088-37-2.

Bylesjo, M., Nilsson, R., Srivastava, V., Gronlund, A., Johansson, A.I., Jansson, S., Karlsson, J., Moritz, T., Wingsle, G. & Trygg, J. (2009). Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *J Proteome Res* 8(1), 199-210.

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. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 101(42), 15243-15248.

Côté, W.A. (1967). *Wood ultrastructure; an atlas of electron micrographs*. Seattle, USA: University of Washington Press.

Courtois-Moreau, C.L., Pesquet, E., Sjödin, A., Muñiz, L., Bollhöner, B., Kaneda, M., Samuels, L., Jansson, S. & Tuominen, H. (2009). A unique program for cell death in xylem fibers of Populus stem. *The Plant Journal* 58(2), 260-274.

Edvardsson, E. (2010). *Integration of Arabidopsis and Poplar model systems to elucidate gene function during wood formation.* Diss. Umeå, Sweden:Swedish University of Agricultural Sciences.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. & Wold, S. (2001). *Multi and megavariate data analysis*. Umeå, Sweden: Umetrics (www.umetrics.com). ISBN 91-973730-1-X.

Eriksson, M.E., Israelsson, M., Olsson, O. & Moritz, T. (2000). Increased gibberellin biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. *Nature Biotechnology* 18(7), 784-788.

Farkas, O., Heberger, K. & Zenkevich, I.G. (2004). Quantitative structure-retention relationships XIV - Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds. *Chemometrics And Intelligent Laboratory Systems* 72(2), 173-184.

Fiehn, O. (2002). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* 48(1-2), 155-71.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. & Willmitzer, L. (2000a). Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18(11), 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. *Analytical Chemistry* 72(15), 3573-3580.

Forcisi, S., Moritz, F., Kanawati, B., Tziotis, D., Lehmann, R. & Schmitt-Kopplin, P. (2013). Liquid chromatography-mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal Of Chromatography A* 1292, 51-65.

Gangl, E.T., Annan, M., Spooner, N. & Vouros, P. (2001). Reduction of signal suppression effects in ESI-MS using a nanosplitting device. *Analytical Chemistry* 73(23), 5635-5644.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. & Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in biotechnology* 22(5), 245-252.

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. *Analytical Biochemistry* 331(2), 283.

Hall, R.D. (2011). Biology of Plant Metabolomics. In: *Annual Plant Reviews*. First edition. ed. p. 448 Blackwell Publishing Ltd.; Volume 43,). ISBN 978-1-4051-9954-4.

Hart-Smith, G. & Blanksby, S.J. (2012). Mass Analysis. In: Barner-Kowollik, C., *et al.* (Eds.) *Mass Spectrometry in Polymer Chemistry*. First edition. ed. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. ISBN 978-3-527-64184.

Hemstrom, P. & Irgum, K. (2006). Hydrophilic interaction chromatography. *Journal of Separation Science* 29(12), 1784-1821.

Herbert, C.G. & Johnstone, R.A.W. (2003). *Mass Spectromerty Basics*. United States of America: CRC Press LLC. ISBN 0-8493-1354-6.

Holcapek, M., Volna, K., Jandera, P., Kolarova, L., Lemr, K., Exner, M. & Cirkva, A. (2004). Effects of ion-pairing reagents on the electrospray signal suppression of sulphonated dyes and intermediates. *Journal Of Mass Spectrometry* 39(1), 43-50.

Hörnblad, E. (2012). *Synthesis of glucuronoxylan in higher and lower plants- Is there conservation of the enzymatic machinery?* Diss. Umeå, Sweden:Swedish University of Agricultural Sciences.

Israelsson, M., Sundberg, B. & Moritz, T. (2005). Tissue-specific localization of gibberellins and expression of gibberellin-biosynthetic and signaling genes in wood-forming tissues in aspen. *Plant Journal* 44(3), 494-504.

Jackson, J.E. (1991). *A users guide to principal components.* New York, USA: Wiley.

Jessome, L.L. & Volmer, D.A. (2006). Ion suppression: A major concern in mass spectrometry. *Lc Gc North America*, 83-89.

Jonsson, P., Gullberg, J., Nordstrom, A., Kusano, M., Kowalczyk, M., Sjostrom, M. & Moritz, T. (2004). A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry* 76(6), 1738.

Kaliszan, R. (2007). QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chemical Reviews* 107(7), 3212-3246.

Kaliszan, R., Kaliszan, A., Noctor, T.A.G., Purcell, W.P. & Wainer, I.W. (1992). Mechanism of retention of benzodiazepines in affinity, reversed-phase and adsorption High-Performance Liquid-Chromatography in view of Quantitative Structure-Retention Relationships. *Journal of Chromatography* 609(1-2), 69-81.

Karjalainen, E.J. (1989). The spectrum reconstruction problem: Use of alternating regression for unexpected spectral components in two-dimensional spectroscopies. *Chemometrics And Intelligent Laboratory Systems* 7(1–2), 31-38.

Katajamaa, M., Miettinen, J. & Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22(5), 634-636.

Katajamaa, M. & Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal Of Chromatography A* 1158(1–2), 318-328.

Kim, H.K., Choi, Y.H. & Verpoorte, R. (2011). NMR-based plant metabolomics: where do we stand, where do we go? *Trends in biotechnology* 29(6), 267-275.

Kojima, M., Kamada-Nobusada, T., Komatsu, H., Takei, K., Kuroha, T., Mizutani, M., Ashikari, M., Ueguchi-Tanaka, M., Matsuoka, M., Suzuki, K. & Sakakibara, H. (2009). Highly Sensitive and High-Throughput Analysis of Plant Hormones Using MS-Probe Modification and Liquid ChromatographyTandem Mass Spectrometry: An Application for Hormone Profiling in Oryza sativa. *Plant and Cell Physiology* 50(7), 1201-1214.

Kopec, R.E., Schweiggert, R.M., Riedl, K.M., Carle, R. & Schwartz, S.J. (2013). Comparison of high-performance liquid chromatography/tandem mass spectrometry and high-performance liquid chromatography/photo-diode array detection for the quantitation of carotenoids, retinyl esters, -tocopherol and phylloquinone in chylomicron-rich fractions of human plasma. *Rapid Communications In Mass Spectrometry* 27(12), 1393-1402.

Kruger, N.J. & Ratcliffe, R.G. (2012). Pathways and fluxes: exploring the plant metabolic network. *J Exp Bot* 63(6), 2243-6.

Lachaud, S., Catesson, A.M. & Bonnemain, J.L. (1999). Structure and functions of the vascular cambium. *Comptes Rendus De L Academie Des Sciences Serie Iii-Sciences De La Vie-Life Sciences* 322(8), 633-650.

Leito, I., Oss, M., Herodes, K. & Kruve, A. (2011). Electrospray ionization efficiency scale of organic compounds. *Abstracts Of Papers Of The American Chemical Society* 241.

Lommen, A. (2009). MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* 81(8), 3079-3086.

Luquez, V., Hall, D., Albrectsen, B.R., Karlsson, J., Ingvarsson, P. & Jansson, S. (2008). Natural phenological variation in aspen (Populus tremula): the SwAsp collection. *Tree Genetics & Genomes* 4(2), 279-292.

Mamyrin, B.A. (2001). Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry* 206(3), 251-266.

Matsumoto-Kitano, M., Kusumoto, T., Tarkowski, P., Kinoshita-Tsujimura, K., Václavíková, K., Miyawaki, K. & Kakimoto, T. (2008). Cytokinins are central regulators of cambial activity. *Proceedings of the National Academy of Sciences* 105(50), 20027-20031.

Mellerowicz, E., Baucher, M., Sundberg, B. & Boerjan, W. (2001). Unravelling cell wall formation in the woody dicot stem. *Plant Mol Biol* 47(1-2), 239-274.

Morreel, K., Goeminne, G., Storme, V., Sterck, L., Ralph, J., Coppieters, W., Breyne, P., Steenackers, M., Georges, M., Messens, E. & Boerjan, W. (2006). Genetical metabolomics of flavonoid biosynthesis in Populus: a case study. *The Plant Journal* 47(2), 224-237.

Nakaba, S., Begum, S., Yamagishi, Y., Jin, H.-O., Kubo, T. & Funada, R. (2012). Differences in the timing of cell death, differentiation and function among three different types of ray parenchyma cells in the hardwood Populus sieboldii × P. grandidentata. *Trees* 26(3), 743-750.

Nieminen, K., Immanen, J., Laxell, M., Kauppinen, L., Tarkowski, P., Dolezal, K., Tahtiharju, S., Elo, A., Decourteix, M., Ljung, K., Bhalerao, R., Keinonen, K., Albert, V.A. & Helariutta, Y. (2008). Cytokinin signaling regulates cambial development in poplar. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 105(50), 20032-20037.

Nilsson, J., Karlberg, A., Antti, H., Lopez-Vernaza, M., Mellerowicz, E., Perrot-Rechenmann, C., Sandberg, G. & Bhalerao, R.P. (2008). Dissecting the

molecular basis of the regulation of wood formation by auxin in hybrid aspen. *Plant Cell* 20(4), 843-855.
Nordstrom, A., Tarkowski, P., Tarkowska, D., Dolezal, K., Astot, C., Sandberg, G. & Moritz, T. (2004). Derivatization for LC electrospray ionization-MS: A tool for improving reversed-phase separation and ESI responses of bases, ribosides, and intact nucleotides. *Analytical Chemistry* 76(10), 2869-2877.
Novák, O., Hényková, E., Sairanen, I., Kowalczyk, M., Pospíšil, T. & Ljung, K. (2012). Tissue-specific profiling of the Arabidopsis thaliana auxin metabolome. *The Plant Journal* 72(3), 523-536.
Oliver, S.G., Winson, M.K., Kell, D.B. & Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in biotechnology* 16(9), 373-378.
Oresic, M., Tang, J., Seppanen-Laakso, T., Mattila, I., Saarni, S.E., Saarni, S.I., Lonnqvist, J., Sysi-Aho, M., Hyotylainen, T., Perala, J. & Suvisaari, J. (2011). Metabolome in schizophrenia and other psychotic disorders: a general population-based study. *Genome Med* 3(3), 19.
Poole, C.F. (2003). *The essence of chromatography*. Amsterdam, Netherlands: Elsevier Science B.V.
Popko, J., Hänsch, R., Mendel, R.R., Polle, A. & Teichmann, T. (2010). The role of abscisic acid and auxin in the response of poplar to abiotic stress. *Plant Biology* 12(2), 242-258.
Raghavendra, A.S., Gonugunta, V.K., Christmann, A. & Grill, E. (2010). ABA perception and signalling. *Trends In Plant Science* 15(7), 395-401.
Roach, M., Gerber, L., Sandquist, D., Gorzsás, A., Hedenström, M., Kumar, M., Steinhauser, M.C., Feil, R., Daniel, G., Stitt, M., Sundberg, B. & Niittylä, T. (2012). Fructokinase is required for carbon partitioning to cellulose in aspen wood. *The Plant Journal* 70(6), 967-977.
Rohde, A. & Bhalerao, R.P. (2007). Plant dormancy in the perennial context. *Trends In Plant Science* 12(5), 217-223.
Santner, A., Calderon-Villalobos, L.I.A. & Estelle, M. (2009). Plant hormones are versatile chemical regulators of plant growth. *Nat Chem Biol* 5(5), 301-307.
Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M.G., Willmitzer, L., Fernie, A.R. & Kopka, J. (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *Febs Letters* 579(6), 1332-1337.
Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M. & Sandberg, G. (2004). A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *Plant Cell* 16(9), 2278-2292.
Sleno, L. & Volmer, D.A. (2004). Ion activation methods for tandem mass spectrometry. *Journal Of Mass Spectrometry* 39(10), 1091-1112.
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. *Analytical Chemistry* 78(3), 779-787.

Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R.J., Li, Y., Nyati, M.K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G.S., Ghosh, D., Pennathur, S., Alexander, D.C., Berger, A., Shuster, J.R., Wei, J.T., Varambally, S., Beecher, C. & Chinnaiyan, A.M. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 457(7231), 910-4.

Svacinova, J., Novak, O., Plackova, L., Lenobel, R., Holik, J., Strnad, M. & Dolezal, K. (2012). A new approach for cytokinin isolation from Arabidopsis tissues using miniaturized purification: pipette tip solid-phase extraction. *Plant Methods* 8(1), 17.

Swartz, M.E. (2005). UPLC™: An Introduction and Review. *Journal of Liquid Chromatography & Related Technologies* 28(7-8), 1253-1263.

Thysell, E., Surowiec, I., Hornberg, E., Crnalic, S., Widmark, A., Johansson, A.I., Stattin, P., Bergh, A., Moritz, T., Antti, H. & Wikstro, P. (2010). Metabolomic Characterization of Human Prostate Cancer Bone Metastases Reveals Increased Levels of Cholesterol. *Plos One* 5(12).

Trygg, J. (2001). *Parsimonious multivariate models.* Diss. Umeå, Sweden:Umeå University.

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

Tuominen, H., Puech, L., Fink, S. & Sundberg, B. (1997). A Radial Concentration Gradient of Indole-3-Acetic Acid Is Related to Secondary Xylem Development in Hybrid Aspen. *Plant Physiology* 115(2), 577-585.

Uggla, C., Magel, E., Moritz, T. & Sundberg, B. (2001). Function and dynamics of auxin and carbohydrates during earlywood/latewood transition in Scots pine. *Plant Physiology* 125(4), 2029-2039.

Uggla, C., Mellerowicz, E.J. & Sundberg, B. (1998). Indole-3-Acetic Acid Controls Cambial Growth in Scots Pine by Positional Signaling. *Plant Physiology* 117(1), 113-121.

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

Vanholme, R., Demedts, B., Morreel, K., Ralph, J. & Boerjan, W. (2010). Lignin Biosynthesis and Structure. *Plant Physiology* 153(3), 895-905.

Vanholme, R., Morreel, K., Ralph, J. & Boerjan, W. (2008). Lignin engineering. *Current Opinion in Plant Biology* 11(3), 278-285.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.M., Wu, Y., Schauer, P., Smith, J.D., Allayee, H., Tang, W.H., DiDonato, J.A., Lusis, A.J. & Hazen, S.L. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472(7341), 57-63.

Wareing, P.F. (1956). PHOTOPERIODISM IN WOODY PLANTS. *Annual Review of Plant Physiology and Plant Molecular Biology* 7, 191-214.

Weckwerth, W. (2011). Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *J Proteomics* 75(1), 284-305.

Wold, S., Ruhe, A., Wold, H. & Dunn, W.J. (1984). THE COLLINEARITY PROBLEM IN LINEAR-REGRESSION - THE PARTIAL LEAST-SQUARES (PLS) APPROACH TO GENERALIZED INVERSES. *Siam Journal on Scientific and Statistical Computing* 5(3), 735-743.

Yanes, O., Clark, J., Wong, D.M., Patti, G.J., Sanchez-Ruiz, A., Benton, H.P., Trauger, S.A., Desponts, C., Ding, S. & Siuzdak, G. (2010). Metabolic oxidation regulates embryonic stem cell differentiation. *Nat Chem Biol* 6(6), 411-7.

Zhou, Q., Gao, B., Zhang, X., Xu, Y., Shi, H. & Yu, L.L. (2014). Chemical profiling of triacylglycerols and diacylglycerols in cow milk fat by ultra-performance convergence chromatography combined with a quadrupole time-of-flight mass spectrometry. *Food chemistry* 143, 199-204.

Acknowledgements

Ja, vad ska man säga. Det här åren har varit en resa såväl forskningsmässigt som privat. Och det är många personer som har gjort det här arbetet möjligt!

Tack **Thomas,** för att du alltid har ställt upp för mig. Jag upphör aldrig att förundras över vilken kunskapsbank du besitter. Och vi är många som undrar hur många timmar du egentligen har på ditt dygn… Att samarbeta med dig, för det tycker jag verkligen att vi har gjort, har varit riktigt, riktigt kul! Och jag är glad att jag fått chansen att fortsätta vara en del av metabolomics plattformen.

Tack **Johan Trygg**, min biträdande handledare, vi kanske inte har setts på regelbunden basis direkt, men jag har ändå alltid vetat att du funnits där när de multivariata dimensionerna varit för många…

Ett stort tack till **Inga-Britt & Krister**, för all hjälp på lab, såväl provupparbetning som skruvande i instrument, ni vet väl att ingenting skulle funka utan er! **Jonas**, tillsammans med dig påbörjade jag min resa och jag är så glad att du är tillbaka! Jag ser fram emot att fortsätta jobba tillsammans!

Stort tack till alla rumskamrater genom åren. **Sara**, min vapendragare, jag saknar dig, våra ibland alldeles för långa fikaraster och diskussioner om allt mellan himmel och jord. **Pernilla**, för att du är så bra på alla sätt och alltid har en förmåga att sprida glädje runt omkring dig! **Maria**, vårt kontors ständiga sekreterare, det är skönt att veta att man alltid kan fråga dig om hur det nu var man gjorde… **Linus**: Länge Leve Hälsingland! **Jonathan L**, (I miss you and our Sweden vs. Australia sport competition…) **András** for always caring about me. **Hasse**, Matlab-skriptens heliga mästare, tack för all input på dataprocesserings-biten och för att jag verkligen har fått någon att retas med. The rest of the metabolomics crew: **Ilka, Kate, Jenny L and Jenny H** for the nice time in the lab and in meetings.

The **Karin Ljungs** group, for interesting group meetings and nice fika. **Ondrej, Alès and Lenka** you are the true analytical chemists! A special thanks to **Lenka** for the StageTips used in Paper V.

Jocke för att du introducerat mig i nano-teknikens ”underbara?” värld och hjälpt mig att hitta vägen i avhandlings-byråkratin. **Anders N. och Maria N.I.** för mig kommer ni alltid att vara UPSCs-själ, tack för att ni kommit tillbaka! Tack **Janne**, för trevligt fika och lunchsällskap och för diskussioner rörande allt från Illustrator till glespanel!

Stefan L och Kjell E, dator för all hjälp med dator relaterade frågor. Tack **Inga-Lis, Ulrika, GunBritt, Nelly, Maria L, Johanna, Gertrud** för att ni håller och hållit institutionen flytande och för trevligt fika- och lunchsällskap.

Rishi & Nathalie, for good collaboration on Paper II. **Anna Linusson**, Paper III hade inte varit möjligt utan dig!

Torgny N., Catherine and Camila for introducing me to the world of hydroponics and amino acid biosynthesis, and for a nice collaboration on Paper IV. **Maggan** för trevligt sällskap och för hjälp med allehanda HPLC relaterade frågor.

Kjell O , Paper V hade aldrig varit möjlig utan ditt fantastiska arbete. Du har alltid har tid att hjälpa andra, och en förmåga att sprida glädje runt dig. Kämpa på, vi saknar dig i ”hörnan”! **Björn, Hannele och Ewa** for answering wood related questions…

Tack till fika-folket, alla nuvarande och gamla doktorander som inte finns kvar här längre, alla spexkompisar genom åren, vad kul vi har haft det!

Kemometrimaffian med **Pär** (Tack för bra samarbete på Paper I) **Anna W, Henrik och Elin * X** (hur många är ni nu egentligen?) i spetsen.

Sandra, Anna K, Maria K, & Sara med familjer för er ständiga omtanke, era upptåg och bara för att ni är helt fantastiska vänner. Nu tycker jag det är dags för den där tjejhelgen!

Karin & Micke, Roberth & Linda, Emma & Peter och alla underbara ungar. 15 år sedan TBi-98 träffades, vad mycket kul vi haft!

Hanna och David, det synd att veta att vi redan haft de bästa grannarna… Oj vad jag saknat er under hösten! I år får vi ta den där Hemavantrippen!

Släkten Bergström för att ni fått den här kustälskande ”sörlänningen” att känna sig hemma i Norrlands inland. **Micke, Jenny, Viggo, Hilda och Sally**, för all omtanke och allt bus. Och **Johanna** förstås, en fantastiskt bonus-svägerska, tack för all posthämtning, vattning och kattvaktning, fast jag vet att katter inte är ditt favoritdjur…

Hela tjocka släkten i Söderhamn, mostrar och morbröder, faster och en hel drös med underbara kusiner, tremänningar och deras familjer, för alla fantastiska somrar!! Ett extra tack till **Karin Ö** för den fina omslagsbilden. Och till Svartsundsligan med **Gun & Pelle** , och **Pelle & Bettan** i spetsen för att ni inpräntat i min själ att ålder inte spelar någon roll och att man har som roligast när man får hjälpa andra! Det finns inte många som har en morbror och moster som tar semester för att åka 40-mil enkel resa för att hjälpa sin systerdotter att bygga hus!

Gulli, jag hoppas du förstår hur mycket du har hjälp mig och Tomas. Speciellt under de senaste åren. Vårt stugbygge hade varit omöjligt utan dig! Snart får du komma och ”hälsa på” oss, på riktigt! Jag önskar att jag får vara lika pigg som dig när jag blir gammal, om inte annat hoppas jag att du fört generna vidare till sonen din…

Per, Carin, Ludwig och Felix jag önskar så ofta att vi bodde närmare varandra! Det är så mysigt de gångerna vi kan ses allihopa, vi får helt enkelt försöka se till att de gångerna kommer oftare.

Jenny, Söstra mi, jag är så glad över att vara din syster! Jag beundrar din envishet, din oerhörda omtanke för dina medmänniskor. Det finns få människor jag känner som med sådan elegans helt plötsligt bara kan vara så självklar del av en grupp. NU ska jag börja träna till Tjejvasan….

Mamma och pappa, jag hade aldrig någonsin varit där jag är idag om det inte varit för ert stöd, engagemang och er kärlek. **Mamma**, tack för att du stod på dig, och till sist fick mig att inse att det heter ”girl” och inte ”gril”, envisast vinner, eller hur? **Pappa**, vi hade en likande dust i matematik några år senare där jag gick förlorande ur striden ytterligare en gång. Trots dessa motgångar så fick ni mig hela tiden att vilja lära mig mer, att förstå att kunskap är något bland det finaste man kan äga, och att det aldrig tar slut på lärandet. Och tack

KA för att du har gett mig möjlighet att få lära mig av din verkliga expertis, jag har i alla fall lärt mig en hel drös med byggtekniska termer!

Alfred, min älskade hjälte. Tack för att du dagligen påminner mig om vad riktig livsvilja och livsglädje innebär, jag är så evinnerligt stolt över dig!

Tomas, jag har nästan inga ord kvar till dig, du är min största supporter, min stora pådrivare, min bästa vän. Det här hade aldrig varit möjligt utan dig. Du har ett hjärta av Guld och du får mig att må bra. JAG ÄLSKAR DIG!